

Article

ASD Recognition Through Weighted Integration of Landmark-Based Handcrafted and Pixel-Based Deep Learning Features

Asahi Sekine ^{1,*}, Abu Saleh Musa Miah ^{1,*}, Koki Hirooka ¹, Najmul Hassan ¹, Md. Al Mehedi Hasan ², Yuichi Okuyama ¹, Yoichi Tomioka ¹ and Jungpil Shin ^{1,*}

¹ School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan

² Research Institute for Electronic Science (RIES), Hokkaido University, Sapporo 001-0020, Japan

* Correspondence: musa@u-aizu.ac.jp (A.S.M.M.); jpshin@u-aizu.ac.jp (J.S.)

Abstract

Autism Spectrum Disorder (ASD) is a neurological condition that affects communication and social interaction skills, with individuals experiencing a range of challenges that often require specialized care. Automated systems for recognizing ASD face significant challenges due to the complexity of identifying distinguishing features from facial images. This study proposes an incremental advancement in ASD recognition by introducing a dual-stream model that combines handcrafted facial-landmark features with deep learning-based pixel-level features. The model processes images through two distinct streams to capture complementary aspects of facial information. In the first stream, facial landmarks are extracted using MediaPipe (v0.10.21), with a focus on 137 symmetric landmarks. The face's position is adjusted using in-plane rotation based on eye-corner angles, and geometric features along with 52 blendshape features are processed through Dense layers. In the second stream, RGB image features are extracted using pre-trained CNNs (e.g., ResNet50V2, DenseNet121, InceptionV3) enhanced with Squeeze-and-Excitation (SE) blocks, followed by feature refinement through Global Average Pooling (GAP) and DenseNet layers. The outputs from both streams are fused using weighted concatenation through a softmax gate, followed by further feature refinement for classification. This hybrid approach significantly improves the ability to distinguish between ASD and non-ASD faces, demonstrating the benefits of combining geometric and pixel-based features. The model achieved an accuracy of 96.43% on the Kaggle dataset and 97.83% on the YTUIA dataset. Statistical hypothesis testing further confirms that the proposed approach provides a statistically meaningful advantage over strong baselines, particularly in terms of classification correctness and robustness across datasets. While these results are promising, they show incremental improvements over existing methods, and future work will focus on optimizing performance to exceed current benchmarks.



Academic Editor: Hersh Sagreiya
Sagreiya

Received: 25 December 2025

Revised: 4 February 2026

Accepted: 7 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Autism Spectrum Disorder (ASD); image recognition; Convolutional Neural Network (CNN); Squeeze-and-Excitation (SE)

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that manifests through challenges in social interaction, communication, and behavior. Individuals with ASD may exhibit repetitive behaviors, a strong preference for routines, sensory sensitivities, and difficulties in understanding social cues. These traits, which often become evident in early childhood, can create significant challenges in daily functioning and are associated

with a wide range of severity [1]. Traditionally, ASD diagnosis has relied on subjective methods such as clinical interviews, behavioral observations, and psychological assessments, often requiring expert judgment to assess the child's development [2]. While effective, these traditional methods are susceptible to biases related to gender, ethnicity, and race, potentially leading to inconsistent diagnoses. Moreover, they can be time-consuming and heavily reliant on human expertise, highlighting the need for more objective and efficient diagnostic tools. Early detection is crucial for implementing timely interventions, which can significantly improve outcomes for individuals with ASD. Deep learning (DL) has emerged as a promising alternative, offering a more objective, automated method for diagnosing ASD by identifying complex patterns in large datasets. DL models can extract features that are not easily discernible by humans, providing a more accurate and scalable approach to ASD recognition [3]. Applications of DL to neuroimaging data, such as fMRI [4] and EEG [5], have revealed neurological differences in individuals with ASD. Despite their high accuracy, these methods tend to be costly and require specialized personnel. In contrast, facial image analysis, which can be conducted using widely available cameras, provides an accessible, cost-effective tool for early ASD detection. Facial datasets, when processed with deep learning algorithms, offer a valuable resource for identifying ASD-related features, such as subtle facial expressions or behavioral patterns [6,7]. Although there is growing interest in multimodal approaches for Autism Spectrum Disorder (ASD) recognition, such as integrating EEG, video, facial images, and landmark data, many of these studies still face challenges in achieving high performance due to the lack of effective and complementary feature sets. Notably, existing work has often treated pixel-based and landmark-based features separately, neglecting the potential advantages of combining them for more robust ASD recognition. To the best of our knowledge, no prior research has explicitly fused facial image-based pixel features with facial landmark-based features for ASD classification. In this work, we propose an incremental improvement in this area by introducing a hybrid deep learning model that combines both pixel-based features and facial landmark-based features, thereby enhancing the accuracy and robustness of ASD recognition. The contributions of our proposed model are as follows:

1. **Dual-Stream Model:** We propose a dual-stream model that integrates facial landmark-based handcrafted features with pixel-based deep learning features for classifying ASD and non-ASD faces. This approach effectively combines landmark-based features with RGB image-based features, capturing complementary information from both feature types. By leveraging both geometric and pixel-based data, the model significantly enhances its ability to distinguish between ASD and non-ASD faces, thus improving classification accuracy and robustness.
2. **Two-Stream Description:** In the first stream, geometric and blendshape features are extracted using Mediapipe, which are then processed through dense layers for further feature refinement. The second stream, on the other hand, processes RGB images by normalizing them using a preprocessing function corresponding to the selected backbone (e.g., ResNet50V2, DenseNet121, InceptionV3). Features are then extracted using a Convolutional Neural Network (CNN) enhanced with Squeeze-and-Excitation blocks and DenseNet layers to capture high-level image features.
3. **Feature Concatenation:** After processing the features from both streams, the outputs are concatenated, merged, and refined. A final softmax layer is applied to classify the ASD and non-ASD faces, ensuring that both feature sets contribute meaningfully to the final decision.
4. **Experimental Results:** On the Kaggle dataset, the model achieves an accuracy of 96.43%, a precision of 97.10%, a recall of 95.71%, and an F1 score of 96.40%. For the

YTUIA dataset, the model performs even better, reaching an accuracy of 97.83%, a precision and recall of 97.78%, and an F1 score of 97.78%.

5. **Contribution and Advantages:** The contribution of our approach lies in the fusion of both facial landmark and RGB pixel features, which significantly improves the classification of ASD. This hybrid model offers enhanced performance, increased robustness, and better integration of complementary features, making it a promising solution for ASD recognition.

2. Related Work

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by social communication difficulties and restricted, repetitive behaviors. Early diagnosis of ASD, similar to Parkinson's Disease (PD) and Alzheimer's Disease (AD), is crucial for effective intervention. In recent years, the use of facial image datasets has gained attention for ASD detection. Deep learning models, particularly Convolutional Neural Networks (CNNs), have shown promise in analyzing facial images to detect subtle morphological features associated with ASD. For instance, these models can identify facial asymmetry, which is often present in individuals with ASD [8]. Multimodal approaches, combining facial features with EEG or speech data, have also been explored to enhance diagnostic accuracy for ASD, PD, and AD [8]. These advances demonstrate the potential of deep learning in the non-invasive diagnosis of neurodevelopmental and neurodegenerative disorders. Studies like those by [9,10] have employed facial asymmetry as a key feature for ASD detection, achieving high accuracy using deep learning algorithms. These studies used 3D blend shapes to analyze facial asymmetry, quantifying differences in key facial regions such as the eyes, eyebrows, and mouth. However, one limitation of existing studies is their reliance on single-domain datasets, which may not generalize well to new datasets or populations. Domain adaptation techniques have been proposed to address this challenge. These techniques allow models to transfer knowledge learned from one dataset to another, enabling the model to perform well on data from diverse sources. Research by [11] demonstrated the effectiveness of domain adaptation in improving model robustness when applied to various medical imaging tasks, including ASD detection. Another approach that has gained popularity is active learning. This method focuses on selecting the most informative data for labeling, reducing the need for large labeled datasets. Active learning has been successfully applied in many medical image classification tasks, including ASD detection, where it helps improve model performance while minimizing the annotation effort. Studies like [12,13] have shown that combining active learning with deep learning models can enhance the accuracy and efficiency of ASD diagnostic systems. Recent work has emphasized combining active learning and domain adaptation in ASD detection. Active learning can help address data variability across different clinical settings by selecting samples that improve model performance [14]. Domain adaptation, on the other hand, allows the model to generalize better to diverse datasets, ensuring its robustness in real-world applications.

Several studies have developed DL-based methods [15–17] for ASD identification using facial images, with most works relying on the Kaggle ASD dataset. Taher M. Ghazal et al. [18] proposed a modified AlexNet-based model, termed ASDDTLA, which reported a comparatively lower accuracy of 87.7%. Furthermore, Alam et al. [15] conducted a systematic ablation study to optimize network architectures, optimizers, and hyperparameters. Using the Xception model with an optimal parameter configuration, they achieved a maximum accuracy of 95%. More recently, Narinder Kaur et al. [19] and M. Ikermane et al. [20] reported accuracies of 70% and 98%, respectively. Across these studies, CNN models pretrained on the ImageNet dataset were predominantly used to extract

discriminative facial features from the Kaggle ASD dataset, underscoring the effectiveness of transfer learning for ASD facial image classification.

3. Dataset Description

In this study, we propose a hybrid ASD recognition framework that combines deep features from facial images with handcrafted features derived from facial landmarks detected using the MediaPipe FaceMesh model. We quantify facial asymmetry by calculating the differences between symmetric landmark pairs across key facial regions. The fusion of deep and geometric features is aimed at improving classification performance while enhancing interpretability and robustness. Table 1 shows the demographic description of the datasets used in the study. We base our approach on a dataset that includes facial images of individuals with autism and normal controls. The dataset enables us to explore the relationship between facial asymmetry and ASD, which can improve the diagnostic capabilities of the model.

Table 1. Demographic summary of Datasets (D1: Kaggle ASD Facial Image Dataset; D2: YTUIA YouTube-based Facial Dataset).

Attribute	Description (Dataset 1 (Kaggle) 152)	Description (YTUIA-YouTube Dataset)
Source	Kaggle (compiled by Gerry Piosenka)	YouTube + SSBD + therapists/institutions
Data type	2D RGB facial images	Video frames (MTCNN-based face extraction)
Age range	2–14 (mainly 2–8)	NC individuals: 1–11, ASD individuals: 3–11
Male/Female ratio	approx. 3:1	NC: 117 male, 56 female; ASD: 93 male, 30 female
ASD/NC ratio	1:1 across all splits	1:1 in training and test sets
Train/Test/Val split	86.38%/10.22%/3.41%	Train: 1068, Test: 100
Demographics	No SES, ethnicity, clinical history, or ASD severity	No SES, ethnicity, clinical history, or ASD severity
Preprocessing	N/A	Detection, alignment, cropping, resizing
Notes	Images collected from online sources	100 YouTube videos curated; 75 from SSBD

3.1. Dataset 1 (Kaggle)

The first dataset was collected from Kaggle and contains facial images of children with autism. It consists of 2D RGB images of children aged between 2 and 14 years, with most samples belonging to the 2–8 year age group. The dataset shows an overall male-to-female ratio of approximately 3:1; however, the distribution between the ASD group and the normal control (NC) group is balanced at nearly 1:1. The data are divided into training, testing, and validation sets, accounting for 86.38%, 10.22%, and 3.41% of the samples, respectively. Each split preserves an equal proportion of ASD and NC images. The dataset was compiled by Gerry Piosenka using publicly available online sources and does not include additional demographic or clinical information such as ethnicity, socioeconomic background, medical history, or ASD severity.

3.2. YTUIA-YouTube Dataset

The second dataset, named YTUIA, was collected at Universiti Islam Antarabangsa and is based on YouTube videos. It includes videos derived from the Self-Stimulatory Behaviours Dataset (SSBD), a widely used resource for autism research. Although only 50 relevant videos were directly available on YouTube, additional videos were identified with the help of therapists and specialized institutions, resulting in a total of 100 YouTube

videos used for frame extraction. For the normal control group, videos showing kindergarten activities within the same age range were selected from YouTube. In the initial stage, faces were detected in each video frame using the MTCNN algorithm. This was followed by a preprocessing process that involved face alignment, cropping, and resizing. The normal control group consisted of 173 unique children, including 117 males and 56 females, aged between 1 and 11 years. The ASD group contained 123 individuals, with 93 males and 30 females, whose ages ranged from 3 to 11 years. To ensure balanced learning, the dataset was divided into a training set of 1068 samples and a test set of 100 samples, maintaining a 1:1 ratio between ASD and normal control participants.

4. Proposed Methodology

The proposed workflow architecture is shown in Figure 1. The proposed methodology for face-based Autism Spectrum Disorder (ASD) recognition uses a two-phase approach: training and testing. During the training phase, input images from both ASD and non-ASD subjects are preprocessed by resizing and normalizing to standardize the data. Feature extraction is performed through two distinct streams: the first stream uses facial landmarks extracted via Mediapipe, capturing 478 points initially and selecting 137 symmetric landmarks, from which geometric and blendshape features are derived. The face position is adjusted using an in-plane rotation based on angles from the outer eye corners (landmarks 33 and 263). These features are processed through Dense layers with dropout for regularization, concatenated, and refined with additional Dense layers to produce the final output for Stream-1. The second stream processes the RGB images, resizing and normalizing them before passing through a Convolutional Neural Network (CNN) augmented with Squeeze-and-Excitation (SE) blocks to extract high-level image features. Global Average Pooling (GAP) reduces dimensionality, followed by DenseNet and a final Dense layer to extract features for Stream-2. The outputs from both streams are then concatenated and merged through a softmax gate with weighted concatenation, dynamically adjusting the contribution of each stream based on its relevance for classification. A final Dense layer with dropout refines the features before passing them through a softmax layer to generate the probabilistic classification score. In the testing phase, input images are preprocessed, features extracted from both streams, and modality-gated fusion applied to output the prediction. Additionally, the model is evaluated by averaging predicted probabilities or using majority voting with validation-optimized thresholds. The ensemble approach, incorporating multiple backbones, further ensures robust classification. This hybrid methodology, combining geometric landmark-based features and pixel-based deep learning features, enhances the model's ability to distinguish between ASD and non-ASD faces, improving both accuracy and robustness.

4.1. Preprocessing

In this study, we employed preprocessing for two streams, RGB-based facial images and facial landmarks, which were extracted using Mediapipe. The details of the preprocessing are provided below.

4.1.1. RGB Based Preprocessing (Resizing and Normalization)

Given an input image $I \in \mathbb{R}^{H \times W \times C}$, where H is the height of the image, W is the width of the image, and $C = 3$ for RGB images. The first step is resizing the image to a standard size of 224×224 . Each input image is resized to $I' \in \mathbb{R}^{224 \times 224 \times 3}$, so that the image dimensions are consistent for the network. This resizing ensures that all input images have the same shape, making them compatible for model processing. The image is then normalized using the preprocessing function corresponding to the chosen backbone (e.g.,

ResNet50V2, DenseNet121, InceptionV3). This step standardizes the pixel values of the image I' , ensuring that they are in a range suitable for the pretrained model's weights. The normalized image I'' is now ready for feature extraction:

$$I'' = \text{Preprocess}_{\text{backbone}}(I')$$

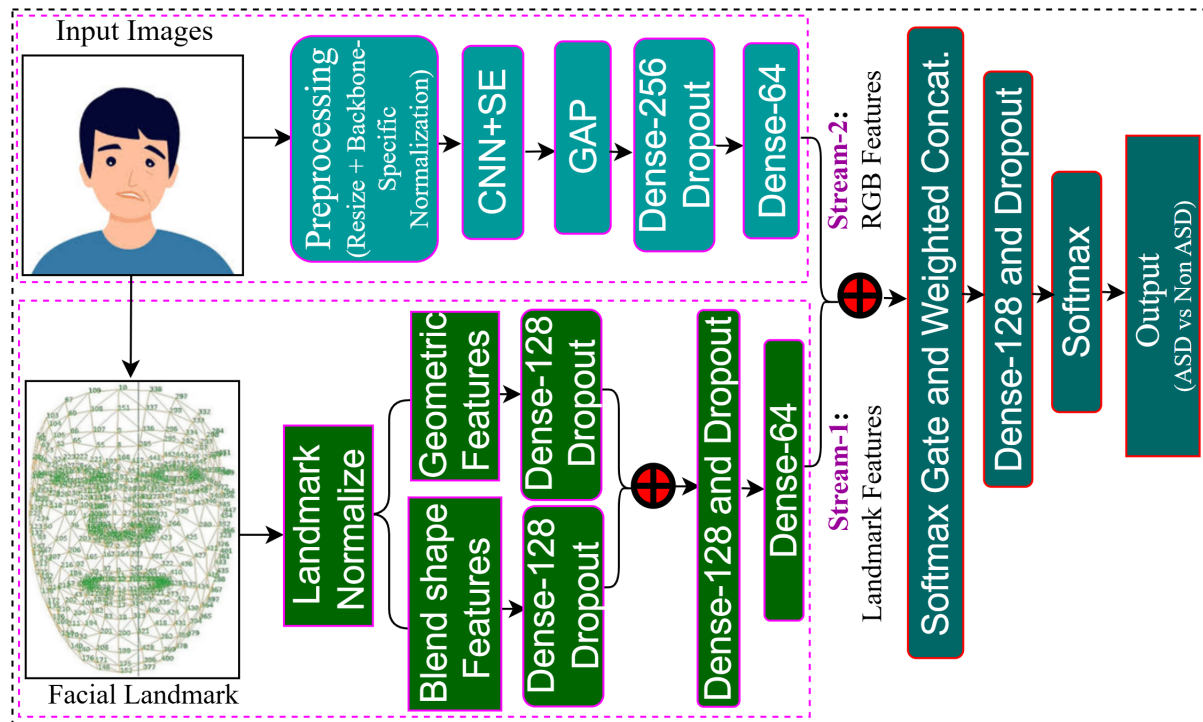


Figure 1. Overview of the proposed two-stream architecture. In Stream-1, face alignment is performed using in-plane rotation based on outer eye corner landmarks (33 and 263) to ensure consistent landmark positioning, followed by extraction and concatenation of geometric and blendshape features. In Stream-2, input images are resized and normalized using backbone-specific preprocessing (ResNet50V2, DenseNet121, and InceptionV3), then processed through a CNN with SE blocks and GAP to obtain deep features. Features from both streams are fused and classified using a softmax layer.

4.1.2. Facial Landmark Based Processing and Selected 137 Symmetric Landmarks

We applied Mediapipe's FaceMesh to extract facial landmarks from the input image. A total of 478 landmarks are initially detected, which includes key facial points such as those around the eyes, eyebrows, nose, and lips. These landmarks provide detailed information about the geometry of the face.

Out of the 478 landmarks, we selected 137 points from symmetric facial areas, such as the eyes with eyebrows, the nose, and the lips. These landmarks were chosen for their relevance in distinguishing subtle features between ASD and non-ASD faces.

In addition to the facial landmarks, we extract handcrafted features from the detected landmarks and expression-related blendshapes using Mediapipe Face Landmarker v2. These blendshapes model facial expressions and contribute further discriminative information to the feature set.

The facial landmarks are represented as

$$\mathbf{lms} = \{(x_i, y_i, z_i)\}_{i=1}^{478}$$

where (x_i, y_i, z_i) represent the 3D coordinates of the i -th landmark. After extracting the landmarks, we use the 137 symmetric points to compute geometric features (e.g., distances

and angles between points) and blend shape features, which are then used for classification. These features, alongside RGB-based features from the second stream, are merged to form the final feature set for classification.

4.2. Stream-1: Facial Landmark-Based Features

In Stream-1, first, we fixed the face position by applying an in-plane rotation using the eye corner angles calculated from the outer eye corners (landmarks 33 and 263). This ensures that all facial landmarks are in a consistent position for feature extraction, regardless of initial face orientation. The in-plane rotation angle θ is derived from the eye-line and used to rotate the face. The face alignment is achieved by applying trigonometric functions $\sin(\theta)$ and $\cos(\theta)$ to correct the face position. This alignment step is crucial for ensuring that the extracted features are consistent across various face orientations.

4.2.1. Geometric Features

The face is aligned based on the eye positions by rotating the face with an in-plane rotation angle θ , calculated as

$$\theta = \text{atan2}(y_{263} - y_{33}, x_{263} - x_{33}), \quad (1)$$

where (x_{33}, y_{33}) and (x_{263}, y_{263}) are the coordinates of the outer canthi of the eyes. After rotation, the face is aligned, and we append $[\sin(\theta), \cos(\theta)]$ to the feature set to capture the corrected face orientation.

Next, we compute geometric features by calculating the 3D and 2D distances between key facial points. The 3D distance between two points a and b is given by

$$D_{3D}(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}, \quad (2)$$

while the 2D distance is calculated as

$$D_{2D}(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}. \quad (3)$$

These distances help capture facial features such as face width and height.

To align facial symmetry, the landmarks are scaled by the inter-outer-canthi distance d , which is calculated as

$$d = \sqrt{(x_{33} - x_{263})^2 + (y_{33} - y_{263})^2}, \quad (4)$$

After this scaling, the landmarks are rotated to align the eye-line horizontally, and the face is centered by subtracting the mid x -coordinate between the two outer canthi.

For each symmetric pair of landmarks (a, b) , we compute the differences in the x - and y -coordinates, as well as their absolute values:

$$[\Delta x, \Delta y, |\Delta x|, |\Delta y|], \quad (5)$$

and when depth information is available, we extend the calculation to the z -coordinates:

$$[\Delta z, |\Delta z|]. \quad (6)$$

For each facial part (eyes, brows, lips, nose, contour), we compute several statistics, including the center, variance, range, and the principal direction (via the leading eigenvector of the 2D covariance matrix). These are used to generate an 8-dimensional feature vector for each part. For the eyes and brows, left-right differences are used, while for the lips, nose, and contour, single-part summaries are computed.

To capture symmetry errors, we reflect the x -coordinates of a subset of points (i.e., $x \rightarrow -x$) and calculate the nearest-neighbor distance to the original set. The mean, standard deviation, and 90th percentile of these distances are aggregated to form additional features:

$$\text{Mirror-Symmetry Error} = \text{mean, std, 90th percentile.} \quad (7)$$

4.2.2. Blendshape Features

In addition to the geometric features, we extract blendshape features that represent facial expressions, such as a smile or a frown, which are indicative of emotional states or behavior. These features are computed from the detected landmarks and capture the dynamics of facial movements. If the blendshape model asset is unavailable, the blendshape features are replaced with a zero vector of the target dimensionality to ensure consistency across all images. The blendshape features are normalized using a reference distance d , which is calculated between landmarks 33 and 263, as described in Equation (4). The x - and y -coordinates of the landmarks are normalized by this reference distance to standardize facial position and scale. Additionally, the face is rotated by an angle θ , computed in Equation (1), to align the face consistently across different images. Finally, the landmarks are translated so that the midpoint between landmarks 33 and 263 is centered at zero in the x -coordinate, ensuring consistent positioning of facial landmarks. For each pair of symmetric blendshape indices, to quantify expression differences between the left and right sides of the face for each pair of symmetric blendshape indices, we derive from the 52 scores (range 0–1) produced by MediaPipe Face Landmarker v2 the following features: for each left–right pair, (diff, To quantify expression differences between the left and right sides of the face for each pair of symmetric blendshape indices, we derive from the 52 scores (range 0–1) produced by MediaPipe Face Landmarker v2 the following features:

$$\text{diff} = s_L - s_R, \quad |\text{diff}| = |s_L - s_R|, \quad \text{sum} = s_L + s_R, \quad \text{max} = \max(s_L, s_R),$$

for each individual key

$$v, \quad v^2, \quad \mathbf{1}[v > 0.5],$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, and group aggregations (mean, max, variance) are computed for the eyes, brows, mouth, jaw, and nose.

4.2.3. Concatenated the Geometric and Blendshape Features

After extracting both geometric and blendshape features, we concatenate these features from the landmark branch and then fuse them with the image branch via a learned modality gate:

$$F_{\text{LM}} = \text{Concatenate}(F_{\text{geom}}, F_{\text{blend}}), \quad F_{\text{fused}} = \text{Gate}(F_{\text{CNN}}, F_{\text{LM}}) \quad (8)$$

where “Gate” refers to a soft attention mechanism that generates modality weights and combines the two modalities into a weighted sum.

After processing the facial landmarks and blendshape features, the resulting features are passed through Dense layers to produce the final feature set for Stream-1. The concatenation and fusion of these features allow the model to leverage both geometric and expression-based information for improved classification.

4.3. Stream-2: RGB-Based Feature Extraction with CNN Backbones

In Stream-2, we process the RGB image I , which is resized and normalized to a standard size and range, as described in the RGB-based preprocessing section. The image I'' after normalization is passed through a Convolutional Neural Network (CNN) backbone,

such as ResNet50V2, DenseNet121, or InceptionV3, all of which are pretrained on ImageNet. The top (classification) layers of the pretrained model are removed to extract features from the convolutional layers.

Resizing and Normalization: The first step in Stream-2 involves resizing the image to a fixed size (e.g., 224×224) and normalizing the pixel values. This ensures that the image has a consistent dimension and range, making it suitable for input into the CNN backbone:

$$I'' = \text{Preprocess}(I') \quad (9)$$

where I' is the resized input image and I'' is the normalized image ready for further processing.

CNN Feature Extraction: The normalized image I'' is then passed through a pretrained CNN backbone. The network is used for feature extraction, leveraging the learned filters and weights from ImageNet. This produces feature maps from the convolutional layers:

$$F_{\text{CNN}} = \text{Backbone}(I'') \quad (10)$$

where F_{CNN} represents the raw feature maps obtained from the CNN backbone. These feature maps capture high-level semantic information about the input image.

Squeeze-and-Excitation (SE) Block: To enhance the representational capacity of the CNN, we apply a Squeeze-and-Excitation (SE) block to recalibrate the channel-wise feature responses. The SE block performs global average pooling (GAP) on the feature maps and applies a lightweight attention mechanism to adaptively recalibrate the features:

$$F_{\text{SE}} = \text{SE}(\text{GAP}(F_{\text{CNN}})), \quad (11)$$

where $\text{GAP}(F_{\text{CNN}})$ computes the global average pooling of the feature maps F_{CNN} , and F_{SE} is the recalibrated feature set.

Global Average Pooling (GAP): Next, the output of the SE block undergoes Global Average Pooling (GAP), which reduces the spatial dimensions of the feature maps into a single vector per channel:

$$\text{GAP}(F_{\text{SE}}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{\text{SE}}(i, j), \quad (12)$$

where H and W are the height and width of the feature map, and $F_{\text{SE}}(i, j)$ are the values at each spatial location. GAP summarizes the spatial information into a single vector.

Dense Layer Projection: The pooled feature vector is then passed through a Dense layer with 256 units and dropout for regularization:

$$F_{\text{Dense}} = \text{Dense}(F_{\text{GAP}}, 256) \quad \text{with dropout}, \quad (13)$$

where F_{Dense} is the output of the Dense layer, and dropout is applied to prevent overfitting.

Finally, a Dense layer with 64 units is applied to reduce the dimensionality and produce the final output features for Stream-2:

$$F_{\text{Stream-2}} = \text{Dense}(F_{\text{Dense}}, 64), \quad (14)$$

where $F_{\text{Stream-2}}$ is the final feature set representing the processed RGB image, ready to be fused with Stream-1 for the final classification.

In Stream-2, we used a CNN backbone to extract features from the normalized RGB image. The SE block enhances the feature map representation, and GAP summarizes the spatial features into a global feature vector. A Dense layer reduces the dimensionality,

and the final Dense layer produces the Stream-2 feature set. These features are then concatenated with the features from Stream-1 for final classification.

4.4. Feature Concatenation and Classification Module

In the classification module, features extracted from Stream-1 (landmark-based features) and Stream-2 (RGB-based features) are fused using a softmax-gated weighting mechanism followed by concatenation. Let F_1 and F_2 denote the feature vectors obtained from Stream-1 and Stream-2, respectively. A softmax gate is applied to generate two scalar weights, α_1 and α_2 , corresponding to the landmark and RGB streams:

$$[\alpha_1, \alpha_2] = \text{Softmax}([s_1, s_2]), \quad \alpha_1 + \alpha_2 = 1 \quad (15)$$

where s_1 and s_2 are learnable stream-level importance scores. These scalar weights reflect the relative contribution of each modality. Each stream's feature vector is then scaled by its respective weight:

$$F'_1 = \alpha_1 \cdot F_1, \quad F'_2 = \alpha_2 \cdot F_2 \quad (16)$$

The weighted feature vectors are concatenated to form a unified fused representation:

$$F_{\text{fused}} = [F'_1 \parallel F'_2] \quad (17)$$

The fused feature vector is subsequently passed through fully connected layers for feature refinement, including a Dense layer with 128 units and dropout for regularization. The final classification is performed using a softmax layer:

$$\mathbf{y} = \text{Softmax}(W F_{\text{fused}} + \mathbf{b}), \quad \mathbf{y} \in \mathbb{R}^2 \quad (18)$$

where W and $\mathbf{b}$ denote the weight matrix and bias vector, respectively. The predicted class label is determined using a validation-optimized threshold τ :

$$\hat{y} = \begin{cases} \text{ASD} & \text{if } y_{\text{ASD}} \geq \tau \\ \text{Non-ASD} & \text{otherwise} \end{cases} \quad (19)$$

For multi-backbone ensemble configurations, outputs from different CNN backbones are combined using probability averaging or majority voting, with thresholds optimized on the validation set to enhance robustness and classification performance.

5. Experimental Evaluation

We conducted experiments to evaluate the performance of the proposed autism classification model based on facial image analysis. The dataset consists of two classes: ASD (Autistic), which includes facial images of diagnosed individuals, and non-ASD, which contains facial images of individuals without ASD. The objective was to assess how well the model can differentiate between these two classes using deep learning techniques applied to facial images. Following the protocol from previous work [14], the dataset was split into training, validation, and test sets. The Kaggle dataset was used with its inherent split, while for the YTUIA dataset, the same splitting approach was followed to ensure consistency in class distribution across both subsets.

5.1. Environmental Setting

The experimental evaluation focused on the performance of the proposed autism classification model using facial image analysis. The dataset consisted of two classes: ASD (Autistic) and non-ASD, where each class contained facial images from individuals

diagnosed with ASD and those without the condition. Following the evaluation protocol of Alam et al. (2024) [14], the dataset was split into training, validation, and testing sets. Specifically, for the Kaggle dataset, the predefined split of 86.38% for training, 10.22% for testing, and 3.41% for validation was used. The YTUIA dataset followed the same general approach, with a split for training and testing, and the validation set was not explicitly defined in the original study. To ensure consistency, the images were resized to a standard size of 224×224 pixels, and RGB image tensors were normalized to a range of $[0, 1]$. In addition to the image data, each sample included a vector of facial landmark-based features, which incorporated both geometric features (such as distances and angles between key points) and blendshape features representing facial expressions. The model was trained using the Adam optimizer with a learning rate scheduler, and checkpoints were saved based on validation accuracy to mitigate the risk of overfitting.

5.2. Hyperparameter Setting

The model was trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32. For training, the Kaggle dataset was run for 100 epochs, while the YTUIA dataset was trained for 30 epochs. Binary cross-entropy was chosen as the loss function to suit the binary classification task. To enhance training efficiency and prevent overfitting, a learning rate scheduler was implemented using the ReduceLROnPlateau callback, which adjusts the learning rate when the validation accuracy plateaus. Additionally, the ModelCheckpoint callback was used to save the model with the best validation accuracy, ensuring that the most optimal model was retained during training. The model also employed feature extraction methods that included both facial landmark-based geometric and blendshape features, as well as CNN-based features derived from the RGB image data.

5.3. Performance Metrics

The performance of the model was assessed using several key metrics, including accuracy, the Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC). Accuracy measures the proportion of correctly predicted instances relative to the total number of instances. The ROC curve was used to visualize the model's diagnostic performance across classification thresholds, while the AUC score provided a single numerical summary of performance across all thresholds. Threshold optimization was performed by sweeping thresholds on the validation set to determine the optimal classification threshold, particularly for ensemble models. For ensemble predictions, final class labels were determined by combining individual model predictions through probability averaging or majority voting, with the thresholds optimized per model to enhance overall classification performance.

5.4. Experimental Result with Kaggle Dataset

Table 2 presents the results of the binary classification task using the Kaggle dataset. The proposed model was trained for 100 epochs with a threshold of 0.85. The model achieved an impressive accuracy of 96.43%, with a precision of 97.10%, a recall of 95.71%, and an F1 score of 96.40%. These results underscore the importance of proper feature alignment, particularly the inclusion of asymmetry and facial part information, in enhancing the model's performance. The model's ability to efficiently distinguish between ASD and non-ASD faces is evident; however, further optimization through model parameter tuning or additional training epochs may lead to even better performance. In comparison, Stream-1, which used only landmarks for training, achieved a lower accuracy of 73.57%, with precision at 78.45%, recall at 65.00%, and an F1 score of 71.09%. While the performance was acceptable, it highlights the necessity of incorporating additional features for more

accurate classification. Stream-2, which used only landmarks for RGB input, achieved substantial improvements, achieving an accuracy of 95.00%, precision of 95.65%, recall of 94.29%, and an F1 score of 94.96%. This result emphasizes the advantage of integrating RGB input in training, which significantly improved the model's ability to classify ASD and non-ASD faces accurately. These findings collectively demonstrate the effectiveness of the proposed method and the potential improvements that can be made by exploring additional features and training strategies.

Table 2. Experimental Results and SOTA Comparison with Kaggle Dataset.

Model	Acc.	Prec.	Recall	F1 Score	Epochs	Thres.	Comments
Stream-1 (Only Geometric Features)	76.64	81.65	63.57	74.90	100	-	Accuracy achieved with only landmarks with Only Geometric features for training.
Stream-1 (Only Blend Shape Features)	74.29	70.99	82.14	76.16	100	-	Accuracy achieved with only landmarks with Blend Shape features for training.
Stream-1 (Geometric + Blend Shape Features)	73.57	78.45	65.00	71.09	100	-	Accuracy achieved with only landmarks for training.
Stream-2: CNN + SE GAP + Dense	95.00	95.65	94.29	94.96	100	-	Accuracy achieved with only landmarks for RGB input.
Combined [Stream-1: Only Geometric Features + Stream-2: CNN + SE GAP + Dense]	93.93	96.95	90.71	93.73	100	-	Accuracy achieved with Stream-1: Only Geometric Features + Stream-2: CNN + SE GAP + Dense.
Combined [Stream-1: Only Blend Shape Features + Stream-2: CNN + SE GAP + Dense]	96.07	97.08	95.00	96.03	100	-	Accuracy achieved with Stream-1: Only Blend Shape Features + Stream-2: CNN + SE GAP + Dense.
Proposed Method Combined of [Stream-1 (Geometric + Shape Based) + Stream-2: CNN + SE GAP + Dense]	96.43	97.10	95.71	96.40	100	0.85	Efficient in classifying ASD and non-ASD faces, with potential for further improvement through parameter tuning or increased epochs.

5.5. Experimental Result with YTUIA Dataset

Table 3 presents the results of the binary classification task using the YTUIA dataset. The model was trained for 30 epochs with a threshold of 0.35. The proposed method achieved an impressive accuracy of 97.83%, with precision, recall, and F1 score all reaching 97.78%. This balanced performance across all metrics demonstrates the model's robustness and its ability to efficiently classify ASD and non-ASD faces. The higher performance on the YTUIA dataset compared to the Kaggle dataset suggests that the model benefits from a more diverse or representative set of facial images, which allows it to capture more discriminative features between ASD and non-ASD individuals. In comparison, Stream-1, which used only landmarks for training, achieved a lower accuracy of 72.83%, with precision at 76.32%, recall at 64.44%, and an F1 score of 69.88%. While the performance was relatively good, it highlights the potential for improvement by incorporating additional features or adjusting the training strategy. Stream-2, which utilized only landmarks for RGB input, achieved an accuracy of 88.00%, with precision at 91.30%, recall at 84.00%, and an F1 score of 87.50%. This result indicates that while RGB input provides some

benefit, it is still less effective compared to the proposed method that uses a combination of features for classification. These results collectively demonstrate the effectiveness of the proposed method in distinguishing between ASD and non-ASD faces, with potential for further enhancement through the exploration of additional features or more intensive training strategies.

Table 3. Experimental Results with YTUIA Dataset.

Model	Acc.	Prec.	Recall	F1 Score	Epochs	Thres.	Comments
Stream-1 (Only Geometric Features)	65.22	67.57	55.56	60.98	30	-	Accuracy achieved with only landmarks with Only Geometric features for training.
Stream-1 (Only Blend Shape Features)	65.17	80.95	38.64	52.31	30	-	Accuracy achieved with only landmarks with Blend Shape features for training.
Stream-1 (Geometric + Blend Shape Features)	72.83	76.32	64.44	69.88	30	-	Accuracy achieved with only landmarks with Geometric + Blend Shape Features for training.
Stream-2: CNN + SE GAP + Dense	88.00	91.30	84.00	87.50	30	-	Accuracy achieved with only landmarks for RGB input.
Combined [Stream-1: Only Geometric Features + Stream-2: CNN + SE GAP + Dense]	94.57	95.45	93.33	94.38	30	-	Accuracy achieved with Stream-1: Only Geometric Features + Stream-2: CNN + SE GAP + Dense.
Combined [Stream-1: Only Blend Shape Features + Stream-2: CNN + SE GAP + Dense]	96.00	96.00	96.00	96.00	30	-	Accuracy achieved with Stream-1: Only Blend Shape Features + Stream-2: CNN + SE GAP + Dense.
Proposed Method Combined of [Stream-1 (Geometric + Shape Based) + Stream-2: CNN + SE GAP + Dense]	97.83	97.78	97.78	97.78	30	0.35	Demonstrates balanced performance across all metrics with a higher threshold for improved classification.

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

As summarized in Table 4, recent state-of-the-art methods for facial image-based ASD classification have employed various evaluation protocols, dataset splits, and problem formulations. These differences must be carefully considered when comparing performance results to avoid overstating the competitiveness of the proposed approach. For instance, Alam et al. (2024) [14] used predefined splits for the Kaggle dataset (86.38% train, 3.41% validation, and 10.22% test), which balances the ASD and non-ASD classes. In contrast, the YTUIA dataset in the same study uses a simpler train–test split with 1068 images for training and 100 images for testing, without explicitly defining a validation set. These variations in dataset splits and evaluation protocols may influence the reported accuracy and performance metrics, making direct comparisons more complex. Furthermore, Alam et al. (2025) [16] applied a subject-independent evaluation on a combined dataset (Kaggle + YTUIA), which provides a more challenging evaluation setting by ensuring no overlap between training and test subjects. This differs from the evaluation protocols used by both the 2024 studies. As a result, the reported performance may differ based on the specific dataset and the inclusion or exclusion of validation and test splits. In contrast,

the proposed method integrates landmark-based geometric and facial asymmetry features with RGB-based deep learning representations using a dual-stream model. This approach processes the datasets using the predefined Kaggle split (86.38%/3.41%/10.22%) and the YTUIA train–test split (1068 images for training and 100 for testing). Our method explicitly balances precision, recall, and F1-score, which provides a more holistic view of performance compared to accuracy alone, and we evaluate the model’s robustness across both datasets using consistent protocols. The comparison is intended to highlight that while our approach achieves competitive performance, it operates under slightly different evaluation conditions than previous work. The significant improvements reported for our method, especially in terms of precision and recall, are notable in the context of these dataset-specific protocols, but we acknowledge the potential for variations in results due to the differing evaluation setups in other studies. In summary, although the numerical results in Table 4 are promising, it is crucial to consider the differences in evaluation protocols and dataset splits when interpreting the competitiveness of our proposed method. Future work will further validate the model under different datasets and evaluation protocols to ensure its robustness and generalizability across various settings.

Table 4. Comparison with State-of-the-Art (SOTA) Methods for ASD Classification Using Facial Images.

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Train/Val/ Test Protocol
Alam et al. (2024) [14]	Kaggle	95.00	95.00	–	94.00	Predefined split: 86.38/10.22/3.41 (ASD:NC = 1:1)
Alam et al. (2024) [14]	YTUIA	95.90	95.90	–	95.90	Train = 1068 images, Test = 100 images (Val not reported)
Alam et al. (2025) [16]	Kaggle + YTUIA	96.00	96.00	96.00	96.00	Domain adaptation: Test sets T1 (Kaggle) and T2 (YTUIA); evaluation on (T1 + T2)
Proposed Method (Ours)	Kaggle	96.43	97.10	95.71	96.40	Predefined split: 86.38/10.22/3.41 (Train/Test/Val)
Proposed Method (Ours)	YTUIA	97.83	97.78	97.78	97.78	Train = 1068 images, Test = 100 images (Val not specified)

5.7. Statistical Significance Analysis

To evaluate whether the observed performance improvements are statistically meaningful, we conducted hypothesis testing across different feature representations and datasets as summarized in the Table 5. Paired *t*-tests and Wilcoxon signed-rank tests were applied to ASD score outputs to assess distribution-level differences, while McNemar’s tests were employed to analyze per-sample classification correctness. On the Kaggle dataset, no statistically significant difference was observed in ASD score distributions ($p > 0.05$); however, McNemar’s test indicates that RGB-based and fusion models achieve significantly higher classification correctness compared to landmark-based representations ($p < 0.01$). In contrast, results on the YTUIA dataset show statistically significant improvements in both ASD scores and classification correctness ($p < 0.05$), highlighting the superior robustness of the proposed approach under more challenging and diverse conditions. These findings suggest that even when absolute accuracy gains are modest, the improvements are statistically validated and consistent at the sample level.

Table 5. Statistical significance analysis of RGB and landmark-based representations across datasets.

Dataset	Test	Metric	<i>p</i> -Value	Significance
Kaggle	Paired <i>t</i> -test	ASD score	0.234	No
Kaggle	McNemar	Correctness	0.0088	Yes
YTUIA	Paired <i>t</i> -test	ASD score	0.036	Yes
YTUIA	McNemar	Correctness	<0.001	Yes

5.8. Overfitting Analysis and Robustness Evaluation

Given the limited sample size of the YTUIA dataset, we further evaluated the robustness of the proposed model using stratified 5-fold cross-validation. This strategy ensures that class distributions are preserved across folds and reduces bias introduced by random train–test splits. As shown in Table 6, the model achieves a mean accuracy of 95.23% with a low standard deviation of 0.61%, while other metrics such as precision, recall, F1-score, and AUC also exhibit minimal variance. These results indicate stable performance across different data partitions, suggesting that the proposed model generalizes well and is not overly sensitive to training data selection. The low variance across folds provides empirical evidence that overfitting is effectively controlled despite the model’s architectural complexity. For the Kaggle dataset, similar cross-validation was conducted to ensure the robustness of the model. The mean accuracy achieved was 84.93% with a standard deviation of 0.62%, demonstrating consistent performance across folds. Precision, recall, F1-score, and AUC also show acceptable values, albeit with slightly higher variance compared to the YTUIA dataset. Specifically, precision ($84.90 \pm 0.87\%$) and recall ($85.00 \pm 1.50\%$) indicate that the model performs well, though with slight fluctuations in different folds. The F1-score and AUC further reinforce the overall model robustness, though the lower AUC of 92.48% (with $\pm 0.95\%$ variance) compared to the YTUIA dataset ($99.10 \pm 0.58\%$) suggests that the model faces slightly more challenges with the Kaggle dataset, likely due to the dataset’s inherent noise or variability. The overall results across both datasets suggest that the model demonstrates strong generalization capabilities. While there is some variance, especially in the Kaggle dataset, the overall performance metrics are robust and suggest that the model effectively avoids overfitting despite the complexity of the architecture. The stratified 5-fold cross-validation provides a solid method for evaluating the model’s performance while minimizing the risk of overfitting.

Table 6. Performance Evaluation Using Stratified 5-Fold Cross-Validation with YTUIA and Kaggle Datasets.

Metric	YTUIA Mean (%)	YTUIA STD (%)	Kaggle Mean (%)	Kaggle STD (%)
Accuracy	95.23	± 0.61	84.93	± 0.62
Precision	96.08	± 2.44	84.90	± 0.87
Recall	93.65	± 2.84	85.00	± 1.50
F1-score	94.78	± 0.85	94.94	± 0.70
AUC	99.10	± 0.58	92.48	± 0.95

6. Conclusions

In this study, we propose a dual-stream framework for classifying Autism Spectrum Disorder (ASD) from facial images by combining deep learning with facial asymmetry-based features. Using Google MediaPipe, we extracted facial landmarks and quantified the asymmetry between the left and right sides of key facial regions, such as the eyes, nose, and mouth. These asymmetry features were integrated with high-level visual features learned via deep learning to improve classification performance and interpretability.

The dual-stream model processes RGB images and facial landmarks independently, capturing complementary information from each stream. The first stream extracts geometric and blendshape features from facial landmarks, while the second stream processes RGB images using a Convolutional Neural Network (CNN) augmented with Squeeze-and-Excitation blocks. The outputs from both streams are then concatenated and refined using a softmax gate to ensure robust classification. Experimental results on the Kaggle and YTUIA datasets demonstrate that our approach can effectively distinguish between ASD and non-ASD individuals, achieving high accuracy and precision. The incorporation of asymmetry features significantly enhances the interpretability and performance of the model, showing the value of combining landmark-based features with deep learning techniques. The key contribution of this work lies in the integration of facial asymmetry features derived from MediaPipe into a deep learning framework for ASD classification. While our approach shows promising results, there are limitations to address. The datasets used for training and evaluation, particularly the Kaggle and YTUIA datasets, are small and exhibit imbalances in age and gender distribution, which may limit the model's generalizability. Future work should aim to expand the dataset with a more diverse sample to enhance the model's robustness and applicability across broader populations. Additionally, the reliance solely on facial features may not fully capture the complexity of ASD. Incorporating additional modalities, such as EEG or speech patterns, could offer a more comprehensive understanding and improve prediction accuracy. Another potential limitation is the risk of overfitting due to the small dataset size. To mitigate this, future research should implement more rigorous cross-validation techniques and utilize larger, more diverse datasets to better assess the model's real-world performance. Additionally, further refinement of the fusion mechanism between landmark-based and RGB features could enhance the model's capabilities. These challenges highlight areas for improvement, and future work will focus on expanding the dataset, integrating additional modalities, and refining the model to improve ASD detection performance and generalization.

Author Contributions: Conceptualization, A.S., A.S.M.M., K.H., N.H., M.A.M.H., Y.O., Y.T. and J.S.; methodology, A.S., A.S.M.M., K.H.; software, A.S., A.S.M.M., K.H.; validation, A.S., A.S.M.M., K.H., N.H., M.A.M.H. and J.S.; formal analysis, A.S., A.S.M.M., K.H., N.H., M.A.M.H. and J.S.; investigation, A.S., A.S.M.M. and K.H.; resources, J.S.; data curation, A.S., A.S.M.M., K.H.; writing—original draft preparation, A.S., A.S.M.M., N.H.; writing—review and editing, A.S., A.S.M.M., K.H., N.H., M.A.M.H., Y.O., Y.T. and J.S.; visualization, A.S., A.S.M.M., K.H., N.H. and J.S.; supervision, J.S.; project administration, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Competitive Research Fund of The University of Aizu, Japan.

Data Availability Statement: Kaggle dataset: <https://github.com/mm909/Kaggle-Autism> (accessed on 4 February 2025), <https://minerva-clinic.or.jp/geneticstesting/autismpanel/column/autism-face/#i> (accessed on 4 February 2025).

Conflicts of Interest: The authors declare no conflict of interest for this research.

References

1. Ghosh, T.; Al Banna, M.H.; Rahman, M.S.; Kaiser, M.S.; Mahmud, M.; Hosen, A.S.; Cho, G.H. Artificial intelligence and internet of things in screening and management of autism spectrum disorder. *Sustain. Cities Soc.* **2021**, *74*, 103189. [CrossRef]
2. Khodatars, M.; Shoeibi, A.; Sadeghi, D.; Ghaasemi, N.; Jafari, M.; Moridian, P.; Khadem, A.; Alizadehsani, R.; Zare, A.; Kong, Y.; et al. Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: A review. *Comput. Biol. Med.* **2021**, *139*, 104949. [CrossRef] [PubMed]
3. Abdou, M.A. Literature review: Efficient deep neural networks techniques for medical image analysis. *Neural Comput. Appl.* **2022**, *34*, 5791–5812. [CrossRef]

4. Wang, M.; Xu, D.; Zhang, L.; Jiang, H. Application of multimodal MRI in the early diagnosis of autism spectrum disorders: A review. *Diagnostics* **2023**, *13*, 3027. [[CrossRef](#)] [[PubMed](#)]
5. Esqueda-Elizondo, J.J.; Juárez-Ramírez, R.; López-Bonilla, O.R.; García-Guerrero, E.E.; Galindo-Aldana, G.M.; Jiménez-Beristáin, L.; Serrano-Trujillo, A.; Tlelo-Cuautle, E.; Inzunza-González, E. Attention measurement of an autism spectrum disorder user using EEG signals: A case study. *Math. Comput. Appl.* **2022**, *27*, 21. [[CrossRef](#)]
6. Alam, M.S.; Tasneem, Z.; Khan, S.A.; Rashid, M.M. Effect of Different Modalities of Facial Images on ASD Diagnosis Using Deep Learning-Based Neural Network. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2023**, *32*, 59–74. [[CrossRef](#)]
7. Cîrneanu, A.L.; Popescu, D.; Iordache, D. New trends in emotion recognition using image analysis by neural networks, a systematic review. *Sensors* **2023**, *23*, 7092. [[CrossRef](#)] [[PubMed](#)]
8. Shin, J.; Miah, A.S.M.; Kakizaki, M.; Hassan, N.; Tomioka, Y. Autism Spectrum Disorder Detection Using Skeleton-Based Body Movement Analysis via Dual-Stream Deep Learning. *Electronics* **2025**, *14*, 2231. [[CrossRef](#)]
9. Derbali, M.; Jarrah, M.; Randhawa, P. Autism Spectrum Disorder Detection: Video Games Based Facial Expression Diagnosis Using Deep Learning. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*. [[CrossRef](#)]
10. El Mouatasim, A.; Ikermane, M. Control learning rate for autism facial detection via deep transfer learning. *Signal Image Video Process.* **2023**, *17*, 3713–3720. [[CrossRef](#)] [[PubMed](#)]
11. Shi, C.; Xin, X.; Zhang, J. Domain adaptation using a three-way decision improves the identification of autism patients from multisite fMRI data. *Brain Sci.* **2021**, *11*, 603. [[CrossRef](#)] [[PubMed](#)]
12. Jung, S.K.; Lim, H.K.; Lee, S.; Cho, Y.; Song, I.S. Deep active learning for automatic segmentation of maxillary sinus lesions using a convolutional neural network. *Diagnostics* **2021**, *11*, 688. [[CrossRef](#)] [[PubMed](#)]
13. Ammari, A.; Mahmoudi, R.; Hmida, B.; Saouli, R.; Bedoui, M.H. Deep-active-learning approach towards accurate right ventricular segmentation using a two-level uncertainty estimation. *Comput. Med. Imaging Graph.* **2023**, *104*, 102168. [[CrossRef](#)] [[PubMed](#)]
14. Alam, M.S.; Elsheikh, E.A.; Suliman, F.; Rashid, M.M.; Faizabadi, A.R. Innovative strategies for early autism diagnosis: Active learning and domain adaptation optimization. *Diagnostics* **2024**, *14*, 629. [[CrossRef](#)] [[PubMed](#)]
15. Alam, M.S.; Rashid, M.M.; Roy, R.; Faizabadi, A.R.; Gupta, K.D.; Ahsan, M.M. Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach. *Bioengineering* **2022**, *9*, 710. [[CrossRef](#)] [[PubMed](#)]
16. Alam, M.S.; Rashid, M.M.; Jazlan, A.; Alahi, M.E.E.; Kchaou, M.; Alharthi, K.A.B. Robust Autism Spectrum Disorder Screening Based on Facial Images (For Disability Diagnosis): A Domain-Adaptive Deep Ensemble Approach. *Diagnostics* **2025**, *15*, 1601. [[CrossRef](#)] [[PubMed](#)]
17. Alam, M.S.; Rashid, M.M.; Faizabadi, A.R.; Mohd Zaki, H.F.; Alam, T.E.; Ali, M.S.; Gupta, K.D.; Ahsan, M.M. Efficient deep learning-based data-centric approach for autism spectrum disorder diagnosis from facial images using explainable AI. *Technologies* **2023**, *11*, 115. [[CrossRef](#)]
18. Ghazal, T.M.; Munir, S.; Abbas, S.; Athar, A.; Alrababah, H.; Khan, M.A. Early detection of autism in children using transfer learning. *Intell. Autom. Soft Comput.* **2023**, *36*, 11–22. [[CrossRef](#)]
19. Kaur, N.; Gupta, G. Refurbished and improvised model using convolution network for autism disorder detection in facial images. *Indones. J. Electr. Eng. Comput. Sci* **2023**, *29*, 883–889. [[CrossRef](#)]
20. Ikermane, M.; Mouatasim, A. Web-based autism screening using facial images and a convolutional neural network. *Indones. J. Electr. Eng. Comput. Sci* **2023**, *29*, 1140–1147. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.