scientific reports

OPEN

An attention-based multi-residual and BiLSTM architecture for early diagnosis of autism spectrum disorder

Rami S. Alkhawaldeh[1,2], Jamil AlShaqsi[2], Bilal Al-Ahmad[1,3], Samar M. Alkhawaldeh[1] & Osama Drogham[4]

The prevalence of autism as a neurological disorder affecting approximately 1 in 54 children diagnosed with the condition and the absence of definitive biomarkers forces clinicians to depend on behavioral observations and history information. The variety of symptom presentations and the dependence on subjective clinical evaluations make the diagnostic process still difficult and require long periods of observing behavior and analysis. Effective and automated methods for early detection of patients are crucial to reducing adverse outcomes. Therefore, we propose a framework model that combines the features of attention layers, Residual layers, and the BiLSTM models as a promising transfer learning model with a multi-phase pipeline that significantly improves detection and recognition performance. The experimental results show that the proposed model obtains effective performance, achieving average values for precision, recall, F1 score, and accuracy scores of 87.5%, 87%, 87.5%, and 87.7%, respectively. These values indicate a balanced performance across the metrics, emphasising the model's ability to precisely and consistently classify autism-related features. Regarding ROC AUC values, the class-specific ROC AUC values are close to 95%, ensuring the robustness of the model to distinguish autism among images effectively.

Keywords Autism spectrum disorder, Attention mechanism, BiLSTM, Deep transfer learning, Early diagnosis, Multi-residual architecture

Autism spectrum disorder (ASD) is a neurological disorder characterized by a range of symptoms, such as difficulties in cognition, physical abilities, repetitive behavior, and social interactions[1–3]. While there is no specific medication for ASD, early intervention is critical. There are challenges in diagnosing ASD due to the lack of standard medical tests, forcing clinicians to rely on behavioral observations and historical information. During recent decades, the prevalence of ASD has increased significantly, with current estimates suggesting that approximately 1 in 54 children are diagnosed with the condition. Effective methods for early detection of patients are crucial for mitigating long-term challenges[3–5]. However, the variety of symptoms and the reliance on subjective clinical evaluations make the diagnostic process still difficult and require long periods of observation of behavior and analysis[2].

Among the many applications of machine learning, deep learning models have produced promising results in medical diagnosis through the analysis and interpretation of diverse datasets. The models recognise patterns within diverse data, including those used in behavioral examinations and medical imaging. In particular, the models are used in a wide range of contexts, including the detection of cancer by analyzing histopathological images, the diagnosis of cardiovascular diseases by analyzing electrocardiograms, and the diagnosis of various neurological disorders by combining brain imaging with behavioral evaluations[6–8]. Deep learning offers a powerful tool for improving the diagnostic process in the complex and diverse field of ASD, where symptoms can differ greatly among individuals. Deep learning models can analyze and interpret massive datasets that include biomarkers, speech patterns, behavioral data, and imaging studies to find subtle patterns that might otherwise go unnoticed by conventional diagnostic methods. Such features allow for earlier detection, which is crucial for timely and effective interventions, and it also improves diagnostic accuracy. When it comes to

[1]Department of Computer Information Systems, The University of Jordan, Aqaba 77110, Jordan. [2]Information Systems Department, Sultan Qaboos University, Muscat 123, Oman. [3]Computer Science and Information Technology, Saint Cloud State University, Minnesota, USA. [4]Prince Abdullah bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Al-Salt, Jordan. ✉email: r.alkhawaldeh@ju.edu.jo

improving outcomes for individuals affected by ASD, deep learning is the way to go because of its adaptability to handle diverse data modalities.

While medical imaging like fMRI offers direct neural insights[9–11], its cost and accessibility limit its use for early, large-scale screening. This study instead leverages facial morphology as a practical biomarker, grounded in the shared embryological pathways of craniofacial and neural development. Our approach uses deep learning not to replace definitive diagnosis, but to provide a scalable and non-invasive screening tool that can identify at-risk individuals for earlier, more comprehensive clinical evaluation. Thus, this study investigates how deep learning can be used to detect ASD in facial images. We plan to evaluate the performance of various deep learning models when analyzing image datasets related to ASD. The study also investigates how these models might be used in clinical settings to help and enhance the diagnostic process. Our ultimate goal is to contribute to the development of better and more accessible diagnostic tools that can lead to an earlier and more accurate diagnosis of ASD, improving the outcomes for individuals affected by the disorder. *Our contribution focuses on the construction of a model that combines the features of the attention layers, the residual layers, and the BiLSTM model as a framework. The framework contains a multiphase pipeline that significantly affects detection and recognition performance*. The first phase is pre-processing the input images as kernel-based patches. In the second phase, we employ a residual layer model for transfer learning to extract deep hierarchical features from images. Due to the significance of detecting the semantic sequences among features, we integrated the BiLSTM model to handle long dependencies in the third phase. We utilise an attention mechanism to further enhance the features in the output from the residual and BiLSTM layers, thereby determining the most pertinent parts of the input features. Lastly, we use the refined output features to determine whether the input leads to ASD during the classification phase.

Related work

As already stated, autism is a complex brain disorder that results in social isolation, problems with eye contact, and redundant and stereotypical behaviors. Although there is no standard cure for autism, early intervention and continued therapy can help manage the condition and enhance the quality of life for those who have it. Therefore, prioritizing adequate care and assistance for each individual with autism is of utmost importance. Early detection is associated with more positive outcomes. Several studies attempt to detect ASD by using various deep learning techniques[12]. In this study, we examined the research that used the AFID dataset as shown in Table 1.

The study in[13] created a refined framework using transfer learning models, specifically enhancing the MobileNetV1 model, to identify the faces of children with ASD. Researchers used images from a Kaggle dataset and implemented a variety of classifiers and pre-trained models for analysis. The enhanced MobileNetV1 model achieved an accuracy of 90.67%, which was the highest recorded, along with the lowest error rates. Furthermore, the model successfully distinguished between various subtypes of ASD through k-means clustering, achieving a 92.10% accuracy rate for two specific ASD subtypes. The authors believe that this model has the potential to assist doctors in accurately identifying ASD in children at a young age.

The study conducted by[14] examines ASD by utilizing social media data and biomedical images, with a specific emphasis on facial recognition technology. The research suggests a system that utilizes deep learning methods to recognize facial landmarks. Three pre-trained models were estimated using a dataset consisting of 2,940 face images from Kaggle; those are Xception, VGG19, and NasNetMobile. The Xception model provides superior results compared to others with an accuracy of 91%, while VGG19 gains an accuracy of 80% and NasNetMobile with an accuracy of 78%. The research is focused on assisting communities and psychiatrists in identifying ASD through a user-friendly web application.

The authors[15] proposed a deep learning model having an accuracy of 94.6% in correctly classifying those who are healthy or may have autism. The model was trained and evaluated on 3,014 child images using MobileNet and dense layers for image classification. Alam et al.[16] use CNN transfer learning to identify ASD in children by facial images as biomarkers. The optimized Xception model gains a superior accuracy result of 95% compared to other models such as VGG19 and ResNet50V2. This method might help with early screening and identification of ASD in children.

Mujeeb Rahman and Subashini[20] investigate five transfer learning models: MobileNet, Xception, and EfficientNet (B0, B1, and B2) to extract features. The models were then used to diagnose autism in children accurately. The Xception model has the best results, with an ROC AUC of 96.63%, a sensitivity of 88.46%, and a negative predictive value (NPV) of 88%. On the other hand, Rabbi et al.[18] identifies autistic children by using the VGG19, InceptionV3, and DenseNet201 models on a facial image dataset. Further, Alkahtani et al.[19] carried out an empirical investigation to identify the CNN model's ideal optimizer settings and hyperparameters to improve prediction accuracy. The transfer learning techniques VGG19 and MobileNetV2 are used. The results show that the MobileNetV2 outperformed the baseline models with an accuracy of 92% on the test set.

Cao et al.[17] proposed a Vision Transformer (ViT) for computational analysis of pediatric ASD. The ViTASD provides a transferable model structure by distilling knowledge from massive facial datasets. A lightweight decoder with a Gaussian Process layer is used for robust ASD analysis, and a vanilla ViT is used to extract features from patients' face images. According to the results of the extensive experiments, ViTASD is the best method for ASD face analysis, and ViTASD-L particularly sets new standards.

Using a dataset of 2,940 facial images, the authors Khan et al.[21] identify autism using pre-trained models such as MobileNetv2, Xception, ResNet-50, VGG16, and DenseNet-121. In terms of performance evaluation, the DenseNet-121 model outperformed other models with a superior accuracy of 96%.

The study by Singh and Kakkar[22] proposed an efficient model for diagnosing autism using electroencephalogram (EEG) signals. They remove noise with a Gaussian filter to extract statistical and spectral-based features for processing using a deep residual network. The chronological sewing training optimization algorithm is used by

Author	Method	Activiation function	Objective function	Evaluation
Cao et al.[17]	VIT	Unknown	Mean Square Error	Accuracy: 94.5%; ROC AUC: 97.9%
Rabbi et al.[18]	VGG19, InceptionV3, DenseNet201	Unknown	Unknown	Accuracy: 85.0%; ROC AUC: 92.3%, Accuracy: 78.0%; ROC AUC: 85.9%, Accuracy: 83.0%; ROC AUC: 91.0%
Alkahtani et al.[19]	MobileNet, VGG-16	softmax	Cross Entropy	Accuracy: 92.0%; Recall: 92.0%; F1 score: 92.0%, Accuracy: 82.1%; Recall: 82.0%; F1 score: 82.0%
Alam et al.[16]	Xception, ResNet-50	Unknown	Cross Entropy	Accuracy: 95.0%; ROC AUC: 98.0%; Precision: 95.0%, Accuracy: 94.0%; ROC AUC: 96.0%; Precision: 94.0%
Mujeeb Rahman and Subashini[20]	Xception, EfficientNetB1	softmax	Cross Entropy	Accuracy: 90.0%; Recall: 88.4%; Specificity: 91.6%; ROC AUC: 96.6%, Accuracy: 89.6%; Recall: 86.0%; Specificity: 94.0%; ROC AUC: 95.0%
Alsaade and Alzahrani[14]	Xception, VGG-19	softmax	Unknown	Accuracy: 91.0%; Recall: 88.0%; Specificity: 94.0%, Accuracy: 80.0%; Recall: 78.0%; Specificity: 83.0%
Hosseini et al.[15]	MobileNet	softmax	Unknown	Accuracy: 94.6%
Akter et al.[13]	MobileNet, DenseNet-121	Unknown	Unknown	Accuracy: 92.1% Accuracy: 83.6%; Recall: 83.6%; Specificity: 83.6%

Table 1. Summary of research works on the AFID dataset focuses on patch-based VIT and transfer learning models.

the ASD detection model to pre-train models and optimize weights, resulting in high accuracy rates of 88.6%. The approach minimizes costs and human intervention while improving detection efficiency, with a negative predictive value of 87.8%, a positive predictive value of 89.4%, a true negative rate of 85%, a true positive rate of 88.9%, and an f-measure of 87.5%. To identify autism in children, the study of[23] trained transfer learning VGG16, VGG19, and EfficientnetB0 models on a Kaggle dataset consisting of 3014 photos. The accuracy of the results was 84.66%, 80.05%, and 87.9%, in that order.

Using a variety of diagnostic categories, the study by Bawa et al.[24] evaluates how well various machine learning algorithms detect patterns in data about ASD. As a result, logistic regression yielded accuracy rates of 99.3% for adolescents and 94.3% for children in binary classification. The Support Vector Machine (SVM) outperformed other models in adult binary classification, achieving a more reasonable accuracy rate of 98.5%. Both SVM and logistic regression demonstrated high accuracy rates of 99.55% for multiclass classification and 97.2% for binary classification across a range of age groups.

Several studies have successfully applied deep learning to ASD detection using facial images, achieving promising results. For instance, Akter et al.[13] enhanced the MobileNetV1 model to achieve 90.67% accuracy, while Alam et al.[16] utilized an optimized Xception model to reach 95% accuracy. Others, like Cao et al.[17], have explored Vision Transformers (ViT) for robust feature extraction. These studies validate the premise that facial phenotypes contain sufficient information for ASD screening.

However, a closer analysis of existing literature reveals several recurring gaps that may limit the clinical translation and robustness of current models. Many approaches rely on standard CNN architectures (e.g., VGG, ResNet) or their direct variants, which may not be optimally designed to capture the temporal or sequential dependencies between facial features. Furthermore, many models do not explicitly incorporate mechanisms to weigh the importance of different facial regions, treating all features with equal significance.

To systematically highlight these gaps, Table 2 provides a comparative analysis of prominent prior models. It evaluates each approach based on its core architecture, its ability to handle sequential data, and its inclusion of feature-weighting mechanisms like attention.

As illustrated in Table 2, a significant opportunity exists to develop a hybrid architecture that synergistically combines the strengths of different model types. There is a clear need for a framework that not only extracts robust spatial features (like a CNN) but also explicitly models the contextual relationships between them (like an RNN/LSTM) and dynamically focuses on the most diagnostically relevant features (like an attention mechanism). Our

Study	Core Model Architecture	Handles Sequential/Temporal Dependencies?	Includes Feature Attention/Weighting?	Parameter Count (Approx.)	Identified Gap
Akter et al.	MobileNetV1	No	No	4.2M	Lacks focus on feature importance and sequential relationships.
Alam et al.	Xception & ResNet50	No	No	25.6M	Standard architecture; does not prioritize specific facial regions.
Cao et al.	ViT	Yes (Implicitly)	Yes (Self-Attention)	-	While powerful, ViT can be data-hungry and computationally intensive.
Rabbi et al.	VGG19 & InceptionV3	No	No	143.7M	Relies on conventional feature extraction without specialized focus.
Mujeeb Rahman et al.	Xception & EfficientNet	No	No	-	Focuses on architectural comparison, not on modeling feature dependencies.
Our Proposed Model	Multi-Residual + BiLSTM + Attention	Yes (BiLSTM)	Yes (Attention Layer)	28.5M	Addresses all identified gaps in a unified framework with moderate efficiency.

Table 2. Comparative analysis of prior models and identification of research gaps.

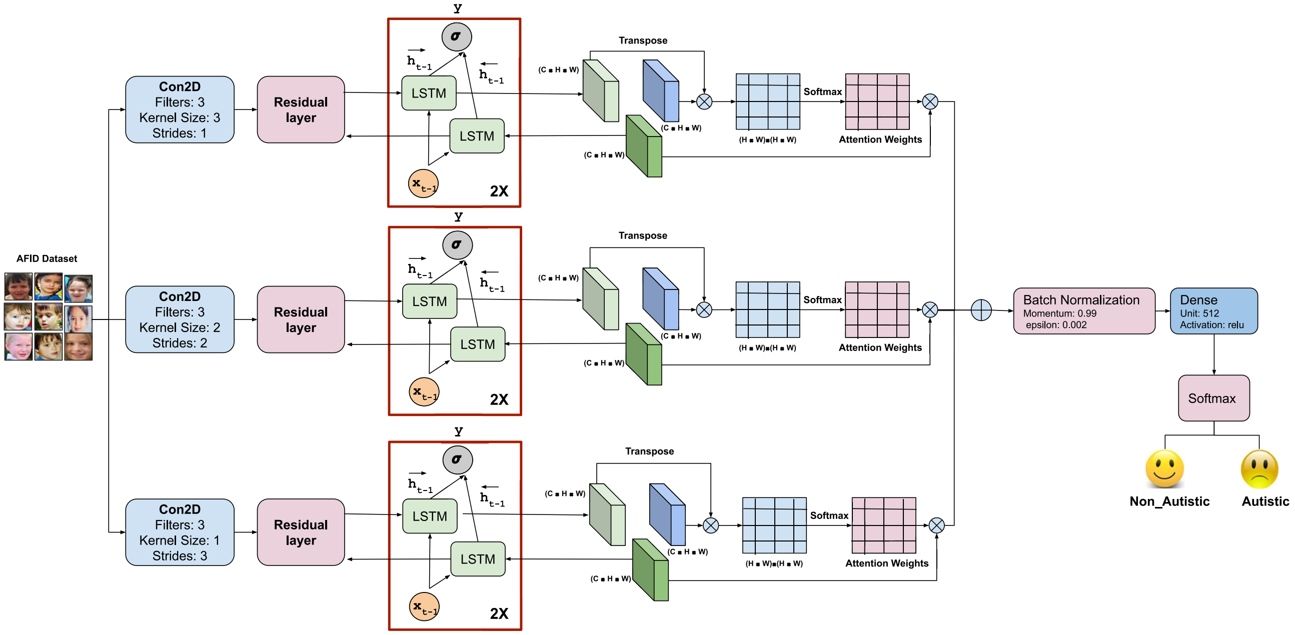


Fig. 1. Attention-enhanced framework utilizing multi-residual and Bidirectional LSTM architecture.

research directly addresses these gaps by proposing an attention-based multi-residual and BiLSTM framework. This hybrid approach is designed to enhance diagnostic accuracy by creating a more comprehensive and context-aware representation of facial phenotypes, thereby improving upon the limitations of prior models.

Attention-based multi-residual BiLSTM model design

The proposed framework aims to detect and recognise ASD using an attention-enhanced residual and BiLSTM model. The framework comprises a multi-phase pipeline, shown in Fig. 1, which is designed to significantly improve detection and recognition performance. The pipeline phases begin with preprocessing the input images, where the input images are preprocessed to enhance features relevant to the model. In the second phase, we employ residual layers to extract deep hierarchical features from images. This allows leveraging knowledge gained from large-scale image recognition tasks, enhancing performance while reducing the need for a substantial ASD-specific dataset. Due to the significance of detecting the semantic sequences among features, in the third phase, we integrate the BiLSTM model to handle long dependencies. To further enhance the features, we utilize an attention mechanism on the output from the residual and BiLSTM layers to determine the most pertinent parts of the input features. Finally, in the classification phase, we utilize the refined output features to classify the input as indicative of ASD.

Kernel-based patch approach for image preprocessing

The input image may contain both extraneous elements (noise) and essential features required for accurate facial expression recognition. To address this issue, the framework refines the input image using three convolution-based kernels with different parameters to capture the critical information accurately. The process of determining the kernels depends on the number of filters, which refers to the number of channels for the output feature maps, the size of the kernel for having more spacing area, and the strides approach that identifies the size of the window through the image.

During the preprocessing phase, we employ multiple convolutional operations using three different kernels to extract significant features from the input image data. Using these convolutional kernels, the framework provides smaller feature maps, allowing the model to focus on and capture distinct aspects of the image. These convolutional operations yield feature maps that function as image patches, highlighting salient features within the images. The first convolutional kernel of (3 × 3) dimension and a stride of (1 × 1) slides across images in single-pixel increments, ensuring comprehensive analysis across the entire spatial domain of the image, resulting in dense feature maps that preserve the spatial structure of the original image. Next, with a window size of 2 and a stride of (1 × 1), the second convolutional kernel scans consecutive pixel pairs. By grouping two adjacent pixels, it enables the extraction of more complex relationships between neighboring pixel values. This step is particularly effective in enhancing the representation of subtle patterns and finer details that may not be fully captured by the first kernel. Finally, we use a third convolutional layer of kernel size 1 and a stride of (3 × 3) to further downsampling images at three-pixel intervals. This helps to maintain important features while decreasing the dimensionality of the feature maps. Specifically, a larger stride guarantees that the convolution abstracts higher-level properties of inherited components by concentrating on wider regions of images. Furthermore, this layer successfully achieves a compromise between computational efficiency and feature extraction, allowing the model to handle the image more effectively in subsequent steps.

Residual transfer learning model

The features produced during the preprocessing phase act as the input for the further stage, where a deep learning model is employed to extract low-dimensional, informative feature representations. In the proposed framework, we use the residual layer model, which is employed to capture and refine patterns from prior patches because of its robust feature extraction capabilities. In particular, the architecture of the residual layer comprises a sequence of convolutional layers that learn the hierarchical shapes of the input data and generalise to further problems with minimal training. The core component of the residual part is the residual layers with a skip connection mechanism that addresses the issue of vanishing gradients as challenges typically encountered in training deep neural networks[25]. In traditional deep networks, the gradient often decreases during backpropagation as the number of layers rises, resulting in diminished convergence. The skip connection (or identity block) allows the gradient to flow more directly through the network by enabling the input of a block to be added to its output, which helps mitigate the vanishing gradient problem.

The residual layers consist of 20 layers disseminated as two blocks, each composed of several convolutional layers with residual connections, as shown in Fig. 2. In particular, the architecture begins with an initial block consisting of a sequence of operations: Conv2D (7 × 7), Conv2D (7 × 7), ReLU activation, and a final Conv2D (7 × 7). A residual (skip) connection bypasses these layers, enabling the direct addition of the input to the block's output, terminating at a summation node. A recurrent residual block, outlined by a dashed box, is repeated three times. Each iteration of this block comprises ReLU activation, MaxPooling2D for reducing the temporal dimension, a Conv2D (7 × 7) layer, another ReLU activation, and a concluding Conv2D (7 × 7) layer. This recurrent block also integrates a skip connection, allowing the input of the block to be directly added to its output.

While optimising, we also apply L2 regularisation, also called weight decay, to all the parameters in the convolutional layers. An often-used technique in CNNs, L2 regularisation helps prevent overfitting by sparsity induction or penalties on large weights. The loss function incorporates the regularisation term within the context of a 2D convolutional layer. The standard equation for a 2D convolutional layer is:

$$z_i = \sum_{j=1}^{k-1} x_{i+j} \cdot w_j + b \tag{1}$$

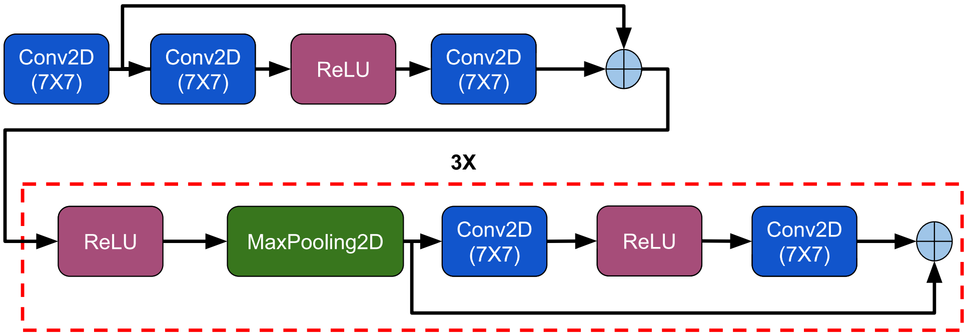


Fig. 2. The residual layer architecture.

where the z_i represents the outcome at position *i*, x_{i+j} represents the input at position $i + j$, position *j* has weight w_j, the bias term is *b*, and the kernal convolutional size is *k*.

The regularization is crucial to alleviate the overfitting during training; hence, we deploy the L2 regularization term with the loss function:

$$L = \frac{1}{2N}\sum_{i=1}^{N}(y_i - z_i)^2 + \frac{\lambda}{2}\sum_{j=0}^{k-1} w_j^2 \quad (2)$$

where *N* represents sample size, the output at position *i* is y_i, and λ represents the regularization parameter.

The loss function in our model comprises two components: the standard mean squared error (MSE) and an L2 regularization term, with MSE as the first component and L2 regularization as the second. The regularization term helps mitigate overfitting by penalizing large weights, thereby reducing their values. In our work on face images, we opted not to apply data augmentation. This decision was based on several potential limitations: (1) Aggressive data augmentation can heighten the risk of overfitting when it is too closely compared to the original, thus restricting generalization; (2) data augmentation can introduce semantically inconsistent samples, potentially impacting the dataset's overall quality; and (3) the effectiveness of augmentation techniques varies, with some methods unintentionally altering the original data and reducing the accuracy of the resulting samples.

Bidirectional LSTM model and attention-based layers

The proposed framework employs a BiLSTM architecture to address long-term dependencies between features. The features in bidirectional structure propagate in both directions as forward and backward, thereby enabling a more comprehensive information flow[26]. The basis of the BiLSTM is the LSTM network with an activation function to preserve consistent information. The LSTM layers are applied iteratively, preserving feature maps generated by the CNN model and using them as sequential input features[27]. As a specialized type of Recurrent Neural Network (RNN), the LSTM network consists of a chain of recurring modules, with information progressing through time steps so that the output from one step serves as input for the next. Thus, the model uses the LSTM cell as a single unit to evaluate the preceding features' impact in a sequence addressing long-term dependencies. As shown in Fig. 3, this structure maintains two components as a state and carry cells that ensure gradient-based information propagates during sequence processing.

As illustrated in Fig. 3, the LSTM model relies on two key values: the hidden state *H*(*t*) updated on a time sequence, and the cell state *C*(*t*) manages long-term memory. A long short-term memory (LSTM) cell's function

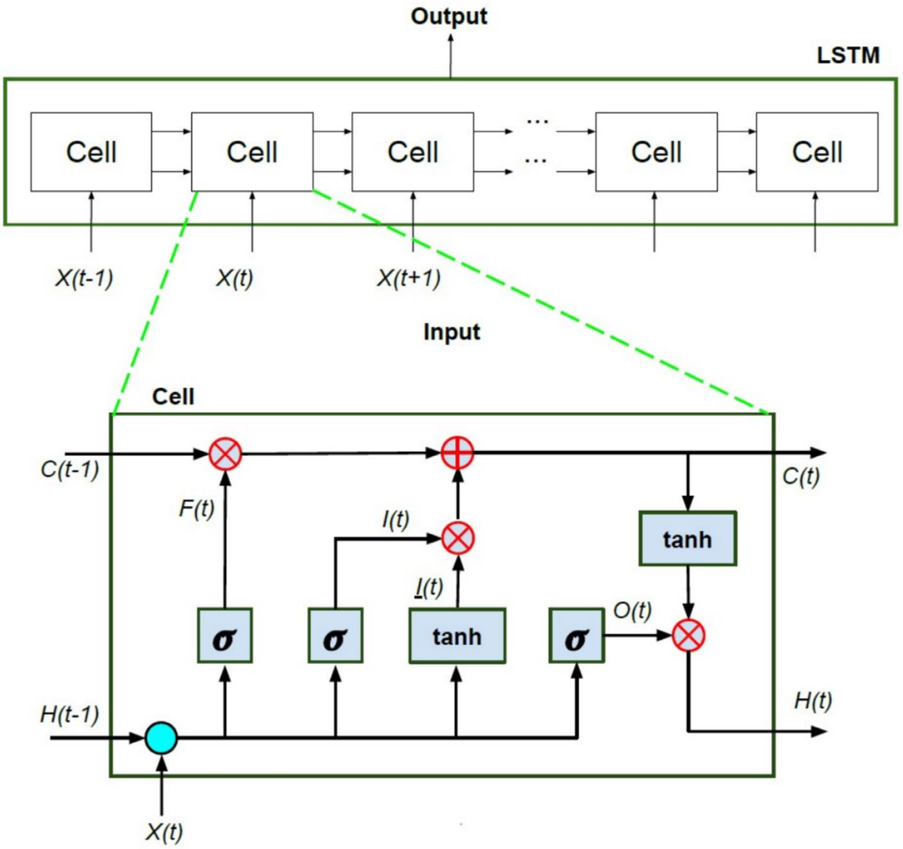


Fig. 3. The core components of LSTM architecture.

is defined by the weights (parameters) learnt during training, and the cell state is updated by updating the information. The input/output gates, target cell, and main cell make up LSTM cells. For the cell to retain or discard $X(t)$ based on the binary output of the Sigmoid function, the forget gate $F(t)$ regulates the influence of the input $X(t)$ and the prior hidden state $H(t-1)$ on the cell state $C(t)$. The input value is fed into the cell state $C(t)$ according to the control of the input gates $I(t)$ and $\underline{I}(t)$. At every time step, the output gate $O(t)$ uses the cell state $C(t)$ to generate predictions, with the last value being specified by the cell. This process is governed by equations presented in Equations 3 - 8. LSTM models employ primary functions such as sigmoid (σ), hyperbolic tangent ($tanh$), multiplication ($\times$), and addition ($+$) to simplify the weight update process during backpropagation.

$$F(t) = \sigma(W_f[H(t-1), X(t)] + B_f) \tag{3}$$

$$I(t) = \sigma(W_i[H(t-1), X(t)] + B_i) \tag{4}$$

$$O(t) = \sigma(W_o[H(t-1), X(t)] + B_o) \tag{5}$$

$$\underline{I}(t) = \tanh(W_i[H(t-1), X(t)] + B_i) \tag{6}$$

$$C(t) = F(t) \cdot C(t-1) + I(t) \cdot \underline{I}(t) \tag{7}$$

$$H(t) = O(t) \cdot \tanh(c(t)) \qquad (\tanh(x) = \frac{\sinh}{cosh} = \frac{e^x - e^{-1}}{e^x + e^{-1}}) \tag{8}$$

Here W is the weight matrix, while B represents the bias parameter. The values W_f, W_i, and W_o represent the weights of the forget, input, and output gates, respectively. LSTM layers are employed due to their ability to handle sequential signals effectively without encountering the vanishing gradient problem, where the gradient values become so small that the weight updates appear negligible during training. In backpropagation, gradient descent updates the weights across the entire network, including those in the LSTM layers.

Integrating BiLSTM significantly enhances the model's predictive performance by leveraging both forward and backward information, substantially accelerating the learning process. We utilize two layers of the BiLSTM model before fully connected networks to manage dependencies within the residual network's features, thus mitigating the long-term dependency issue. These dependencies form contextual features that boost the model's relevance and performance. To focus on the most salient features within the sequence, an attention mechanism is applied, particularly in sequence modeling tasks like natural language processing. In particular, the model weights the significance of each element in the input relative to each other, thereby enabling it to capture more meaningful associations in features. In attention mechanisms, the computation of attention weights involves transposing matrices and applying the softmax function to normalize the weights. The process computes raw attention scores. For two matrices, the query Q and the key K, the raw attention scores are computed as:

$$Scores = Q \times K^T \tag{9}$$

Where K^T represents the transpose of the key matrix. This dot product measures the relevance of each query element for each key element.

The process, in the second step, scales the scores to avoid large values when using high-dimensional values, the scores are scaled by the square root of the input dimension

$$Scores_{scaled} = \frac{Scores}{\sqrt{d_k}} \tag{10}$$

The scaled scores, in the third step, are passed through the softmax function along the last dimension, turning them into a probability distribution across each query-key pair:

$$\text{Attention Weights} = \text{Softmax}(Scores_{scaled}) \tag{11}$$

Finally, these attention weights are used to take a weighted sum of the value matrix V.

$$\text{Attention Output} = \text{Attention Weights} \times V \tag{12}$$

This process enables the model to focus on specific parts of the input sequence by adjusting attention weights, which are normalized using softmax after the transpose and scaling steps. Finally, in the classification phase, the network of three layers is a batch normalization layer, a dense (fully connected) layer, and a final softmax layer. The batch normalization layer stabilizes and accelerates the training process by standardising activations, thus reducing internal covariate shifts. The dense layer maps the filtered features into a lower-dimensional space tailored to the target task. Finally, the classification phase of three layers is dense, batch normalization, and final softmax layer. The dense layer is an important feature extractor, mapping filtered features into a lower-dimensional space specific to the target task, the batch normalization layer ensures stability and speeds up the training process by standardizing filtered features. The softmax layer improves interpretability by converting the dense layer's outputs into normalized probability distribution across two classes.

Experimental setup and results analysis

Dataset setting

The Autism Facial Image Dataset (AFID) is a curated collection of images intended to serve as a benchmark for developing and evaluating models for ASD diagnosis and analysis. It consists of facial images from individuals with autism. It is designed to aid in developing and evaluating machine learning models that can identify facial patterns associated with ASD and is made available under the Creative Commons CC0: Public Domain Dedication license, which permits its use for research. The dataset is primarily used to investigate facial phenotypes as potential biomarkers for earlier and more accurate autism diagnosis. AFID is a valuable resource for medical research and technological advancements in detecting and intervening in autism.

The dataset employed in this research was provided by Piosenka (2021). The majority of images are from online pages, forming face images of both autistic and non-autistic children. While most of the images represent children from Europe and the United States, there is less representation from other regions. The dataset includes images of both boys and girls within each group, and the number of non-autistic children's images is equal to that of autistic children. The children's dataset containing 2,940 images is split into three subsets: training, validation, and test sets. The training set consists of 2,352 images of 80% of the total dataset. The validation and test sets include 294 images for each set, representing 10% of the dataset.

We acknowledge the ethical sensitivity of using facial data, especially of children. Our use of the AFID dataset adheres strictly to the terms of its public domain license. The data is anonymized, containing no personally identifiable information linked to the images. Our work utilizes this valuable public resource with the goal of advancing medical science in a responsible and ethical manner.

Prior to model training, a series of standard preprocessing steps were applied to the AFID dataset to ensure consistency and optimize performance. First, all images were resized to a uniform input dimension of 224×224 pixels to be compatible with the architectures of the various models tested. Second, the pixel values of each image were normalized from the integer range of to a floating-point range of by dividing by 255. This normalization step is crucial for stabilizing the learning process.

Notably, we made a deliberate decision not to apply data augmentation techniques (such as random rotations, flips, or zooms). This choice was made to preserve the integrity of the subtle morphological features present in the facial images, as aggressive augmentation could potentially introduce semantically inconsistent artifacts and negatively impact the dataset's quality. The potential limitations of this decision are further discussed in the "Limitations and Future Work" section.

Evaluation metrics and hyper-parameters

We conducted experiments for training the proposed model using Keras 2.14.0 infrastructure designed for deep learning. The TensorFlow machine learning framework serves as its foundation. The experiments were conducted on the Kaggle platform, employing a GPU P100. The training process lasted for 123.7 seconds. The model's performance was assessed using the following evaluation metrics:[28].

- **Accuracy:** represents the model's percentage of positive predictions compared to total predictions. In an imbalanced dataset, accuracy can be misleading if it is high for the majority class but fails to make meaningful predictions for the minority class.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{13}$$

- **Precision:** is a measure that focuses on the accuracy of positive labels. The mathematical notation of the precision metric divides the number of true positive labels by adding true and false positive predictions.

$$Precision = \frac{True\ Positive}{True\ Postitive + False\ Positive} \tag{14}$$

- **Recall:** is a portion of positive instances detected accurately by the classifier. Divide the total number of accurate positive predictions by the total number of inaccurate positive and negative predictions to perform the computation. When the consequences of a false negative are significant, the recall is extremely valuable.

$$Recall = \frac{True\ Positive}{TruePositive + False\ Negative} \tag{15}$$

- **F1 score:** is a compromising mathematical average between precision and recall, 1 indicates an optimal value while a zero value suggests an imbalance. The F1 score is valuable in situations with imbalanced datasets where the goal is to minimize false negatives.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{16}$$

- **ROC Under-curve:** is an acronym for Receiver Operating Characteristic. It is a visual representation that illustrates the efficacy of a classification model using 1-specificity and recall metrics at different threshold values. The false positive rate (or 1-Specificity) is the proportion of negative cases that are mistakenly forecast as positive relative to the total number of real negative instances.

$$\text{FP-Rate (1-Specificity)} = \frac{False\ Positive}{False\ Postitive + True\ Negative} \tag{17}$$

The ROC curve provides an overview of the model performance at different threshold values. In addition, it is useful on medical diagnostic datasets of imbalanced class distribution. In such scenarios, traditional accuracy can be a misleading metric. For instance, a model that exclusively predicts the majority class would appear highly accurate while completely failing to identify the minority class. Moreover, another key advantage of the AUC ROC Curve is its ability to provide a single number that acts as a composite measure of the classifier performance and an efficient measure of its discriminatory power. It is a measure of the strength of this relationship and takes a value between 0 (no discrimination) and 1 (perfect discrimination). This single score makes it easy to compare models and helps researchers and practitioners quickly see the winners. In addition, the ROC AUC is class balance invariant and is a suitable benchmark for assessing classifiers across different clinical conditions. It is particularly useful in medical settings where creating balanced datasets can be difficult. Because of its relative insensitivity to class imbalances, the ROC AUC is useful in evaluating diagnostic tools; in this role, it often provides an interpretation and validation of improvement.

The hyperparameters used during the training phase, shown in Table 3, were determined through a systematic grid search methodology. This process, conducted on the validation set, ensures that our reported results are based on an optimized and robust model configuration. From this search, a batch size of 32 of search space (16, 32, 64) was selected to balance the efficiency of batch gradient descent with the robustness of stochastic gradient descent. Furthermore, we established the learning rate at 0.001 (from 0.0001 to 0.01), which was found to improve the network's functionality by stabilizing the findings. For the optimization algorithm, we selected Adamax, as its parameters (built on Adadelta and RMSprop) effectively preserve the increasingly declining average of past gradients, comparable to momentum. Other ideal parameters, such as vector dimensionality and the number of multi-head attention units, were also finalized through this experimental optimization process.

Model evaluation

To assess the predictive performance and generalizability of the proposed model, we employed a stratified 10-fold cross-validation methodology. This robust evaluation framework ensures that the available data is used comprehensively, providing reliable performance estimates while minimizing potential bias and variance.

The cross-validation procedure systematically partitions the dataset into ten equally-sized, mutually exclusive subsets (folds) while preserving the original distribution of target classes within each fold through stratification. This stratification process is particularly crucial for maintaining representativeness across all folds, especially when dealing with imbalanced datasets where certain classes may be underrepresented. During each iteration of the cross-validation process, nine folds are used as the training set for model development, while the remaining fold serves as an independent validation set for performance evaluation. This procedure is repeated ten times, with each fold serving exactly once as the test set, ensuring that every data instance contributes to both model training and validation across different iterations. For each fold, the model undergoes complete training on the designated training subset, followed by evaluation on the corresponding test subset. Performance metrics are computed for each individual fold, and the final model performance is determined by averaging these metrics across all ten iterations. This averaging process provides a more stable and reliable estimate of the model's true performance compared to single train-test splits.

The stratified 10-fold cross-validation approach offers several methodological advantages. First, it maximizes the utilization of available data by ensuring that each instance participates in both training and testing phases. Second, it reduces the risk of overfitting by evaluating the model on multiple independent test sets, thereby providing a more realistic assessment of generalization capability. Third, the stratification component maintains class balance across folds, preventing potential bias that could arise from uneven class distributions in individual subsets. This evaluation strategy enables the identification of potential overfitting scenarios, where the model demonstrates superior performance on training data but exhibits degraded performance on unseen test data.

Hyper-parameters	Setting
Learning rate (lr)	0.001
Optimizer	$Adamax(lr = 0.001, beta_1 = 0.9,$
	$beta_2 = 0.999, epsilon = 1e - 08)$
Kernal size	5
Num_layers	4
D_model (dimensionality of vectors)	64
Num_units	64
$Batch\_size$	32
Epochs	50

Table 3. The hyper-parameter values in this study.

Discrepancies between training and validation performance are important indicators of potential overfitting and can inform decisions regarding model complexity and regularization strategies.

Experiment results and discussion

To validate the proposed model's performance and generalization capabilities, we conducted a series of robust experiments comparing it against several baseline models. Thus, we conduct extensive experiments on varied transfer learning models and compare the outcomes with the proposed model. The experimental design has two phases: the training and validation phase and the other phase is model evaluation and testing phase.

Training and validation phase

The training phase is essential for optimizing the classifier's parameters by learning from the input training set. The model receives data in the form of batches of images, and its weights are updated according to the values of hyperparameters during a sequence of epochs (or iterations) that comprise the learning process. Thus, each epoch provides the model with a new possibility to learn from the data, incrementally improving its understanding of the underlying patterns in the training set. Consequently, the model can accurately predict the target labels and invisible instances. To mitigate overfitting during training, the model's performance was monitored on a separate validation set. Hyperparameters were tuned based on this validation performance.

The comprehensive results for a set of transfer learning models during the training and validation phases are presented in Fig. 4. This figure includes two subfigures, each representing a key evaluation metric—accuracy and loss—for each transfer model. Specifically, the y-axis in the first subfigure depicts accuracy values. In contrast, the y-axis in the second subfigure illustrates loss values, with both metrics plotted against the x-axis, representing epoch values across 50 epochs. Most baseline models began to converge around epoch 10, eventually reaching near-100% accuracy on the training data. Notably, the proposed model and ResNet50 demonstrated faster convergence, reaching near-100% training accuracy as early as epoch 5. From the validation perspective, a similar pattern appears across models, with most attaining a validation accuracy near 83%. However, the proposed model and ResNet50 stand out by reaching exceptional validation accuracy earlier in the training process—around epoch 5—averaging an impressive 85.4% accuracy. This indicates that these models not only converge more rapidly but also maintain robust validation performance from an early training stage. The experimental results obtained demonstrate a notable performance benefit of the proposed model, highlighting its effectiveness in accurately identifying autism in children at early stages. The model's architecture, which integrates attention layers with a BiLSTM network, confirms particularly the robustness in separating and prioritizing the most relevant features that are crucial for early autism detection. In particular, the attention mechanisms focus on critical input features, while the BiLSTM component allows for capturing temporal dependencies. This combination contributes to improving the performance of the model with an effective ability to detect early signals of autism with high accuracy, establishing a promising early diagnosis approach.

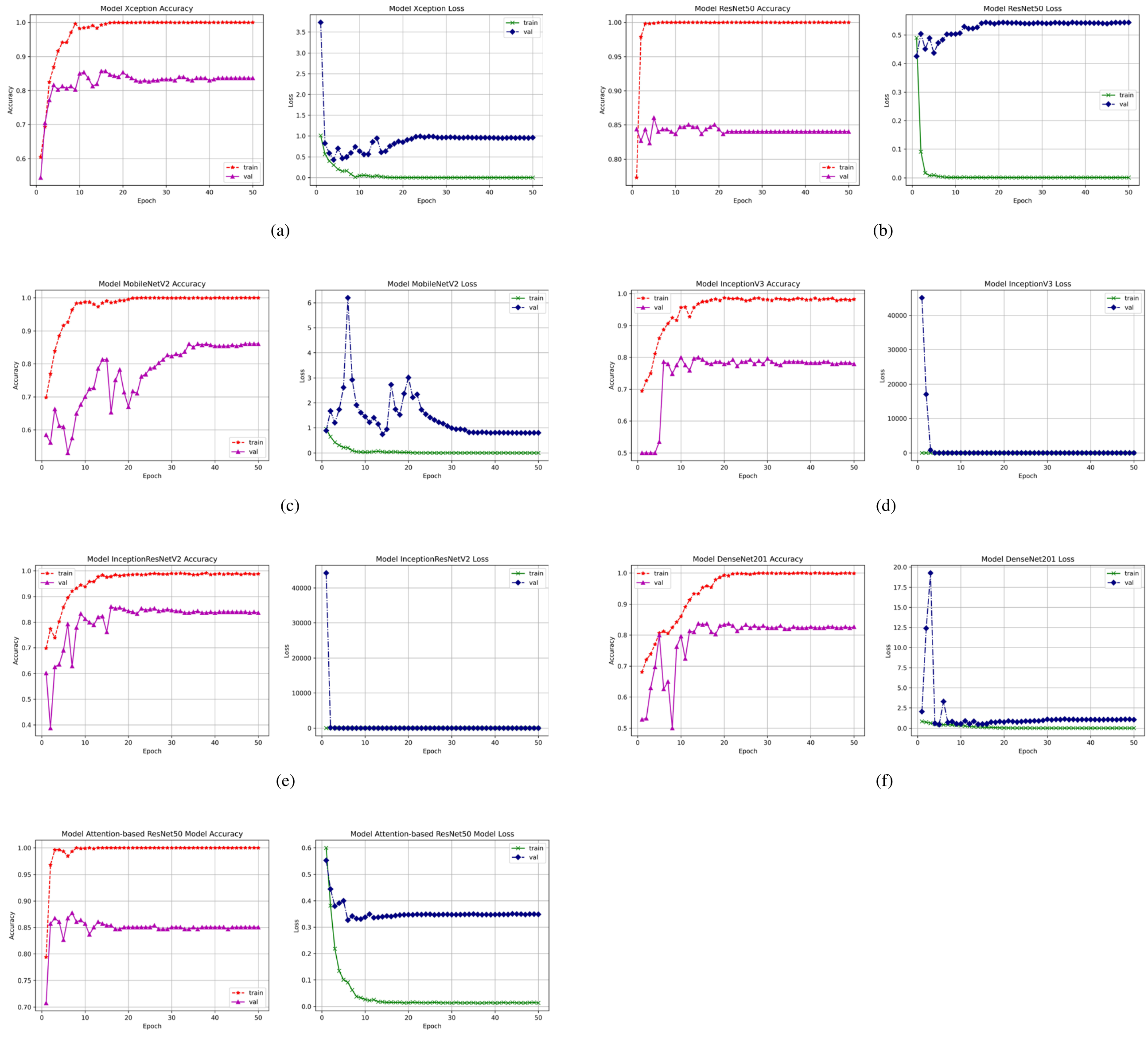
Evaluation and testing phase

The robustness of a machine learning model is defined by its ability to generalize and perform consistently on unseen data, achieving high performance across diverse test sets when evaluated. This study's evaluation and testing phases focus on the model's ability to generalize well beyond the training data. A robust model effectively manages the bias-variance tradeoff, demonstrating low error on both the training data and unseen test data. However, the images in the test set may contain invisible patterns leading to a struggle for the model to recognise autism. A failure to generalize to the feature distribution of the test set results in lower evaluation metrics, such as precision, recall, and F1 score, as the representations in the trained model are aligned with the training set deviated from the test set. Hence, it provides insights into the model's ability to balance sensitivity to new patterns with consistent accuracy across varied data.

In extensive experiments on a test set, an unseen collection of images, we evaluate the effectiveness of the proposed model and compare its performance with baseline techniques. The validation approach estimates the robustness of the model and its generalization ability on data that is not seen in training. The performance of our proposed model is summarized in Table 4 using five important evaluation metrics: accuracy, precision, recall, F1-score, and a supplementary metric for robustness comparison between two different classes (autism-present and autism-absent). Compared to baseline models, the proposed model shows an effective performance improvement. In particular, it achieves average precision, recall, F1 score and accuracy scores of 87.5%, 87%, 87.5%, and 87.7%, respectively. These values indicate a balanced performance across the metrics, emphasising the model's ability to precisely and consistently classify autism-related features. For comparison, MobileNetV2, one of the strongest baseline models, achieved an average performance of approximately 86.5% across all metrics. Although MobileNetV2 exhibits competent performance, the proposed model's higher evaluation scores indicate its enhanced ability to precisely identify autism, offering a more accurate tool that could result in improvements in early detection and diagnostic support. To formally validate these results, a paired t-test was conducted on the F1-scores from the 10-fold cross-validation between our proposed model and the top-performing baseline, MobileNetV2. The test confirmed that the improvement is statistically significant ($p< 0.05$), providing strong evidence that our architecture offers a tangible advantage for this task.

These results indicate that the proposed architecture, which combines attention layers with a BiLSTM network, demonstrates robust performance by effectively identifying pivotal features for early autism detection and presenting a promising approach for early diagnosis.

Figure 5 illustrates the ROC (Receiver Operating Characteristic) curves for the proposed model and baseline models, revealing their performance in terms of recall (true positive rate) against fall-out (false positive rate, calculated as 1-specificity) across various thresholds. The plot visually represents how well each model discerns between the two classes—autistic and non-autistic. The proposed model achieves high performance, as indicated



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 4. Training and validation phase results. (**a**) Xception, (**b**) ResNet50, (**c**) MobileNetV2, (**d**) InceptionV3, (**e**) InceptionResNetV2, (**f**) DenseNet201, (**g**) Attention-based Residual Model.

by its ROC AUC values; however, the micro-average ROC curve, represented by a dashed green line, has an ROC AUC of 0.94, indicating that the model performs well across individual thresholds regardless of class imbalance. The macro-average ROC curve, depicted by a dashed cyan line, has a ROC AUC of 0.95, calculated by taking the average performance across all classes. This high value shows the model's ability to maintain stable discrimination performance across autistic and non-autistic classes. Examining the ROC curves for each class, the autistic class ROC curve, represented by a blue line, and the non-autistic class ROC curve, shown with a red line, each yield an ROC AUC of 95%. These class-specific ROC AUC values, close to 100%, indicate the model's robust ability to distinguish between autistic and non-autistic images effectively. In addition, the model proposed showed high performance based on its ROC AUC values. The micro-average ROC curve had an ROC AUC of 94%, indicating good performance across individual thresholds despite class imbalance. The macro-average ROC curve had an ROC AUC of 95%, showing consistent discrimination performance across autistic and non-autistic classes. When examining the ROC curves for each class, both the autistic and non-autistic classes had ROC AUC values of 95%, suggesting the model's robust ability to effectively differentiate between autistic and non-autistic images.

Computational efficiency analysis

In addition to predictive accuracy, the computational efficiency of a model is a critical factor for its potential deployment in clinical or edge-computing environments. We analyzed the parameter count of our proposed model in comparison to the baseline architectures. Our model consists of approximately 28.5 million parameters.

Reference/Model	Class	Precisison	Recall	f1-score	Accuracy
Mujeeb Rahman and Subashini[20]	Autistic	0.85	**0.86**	0.85	0.85
Xception	Non_Autistic	0.86	0.85	0.85	0.854
Alam et al.[16]	Autistic	0.87	0.83	0.85	0.85
ResNet50	Non_Autistic	0.84	0.88	0.86	0.85
Akter et al.[13]	**Autistic**	**0.89**	0.83	**0.86**	**0.86**
MobileNetV2	**Non_Autistic**	**0.84**	**0.9**	**0.87**	**0.864**
Rabbi et al.[18]	Autistic	0.89	0.8	0.84	0.85
InceptionV3	Non_Autistic	0.82	0.9	0.86	0.85
InceptionResNetV2	Autistic	0.87	0.83	0.85	0.85
	Non_Autistic	0.84	0.87	0.85	0.85
Rabbi et al.[18]	Autistic	0.84	0.81	0.82	0.83
DenseNet201	Non_Autistic	0.82	0.84	0.83	0.83
Our Proposed Model	**Autistic**	**0.91**	**0.84**	**0.88**	**0.88**
	Non_Autistic	**0.85**	**0.91**	**0.88**	**0.874**

Table 4. Comparison Evaluation Performance of the Proposed Model based on Hyper-parameters in Table 3.

As shown in Table 2, this is significantly more efficient than heavy models like VGG19 (143.7M) and is broadly comparable to other powerful baselines such as ResNet50 (25.6M). This analysis demonstrates that our model's performance gains are not achieved through brute-force scaling but through an efficient and strategic architectural design. It strikes a favorable balance between high performance and manageable computational complexity, making it a viable candidate for real-world application.

Limitations and future work

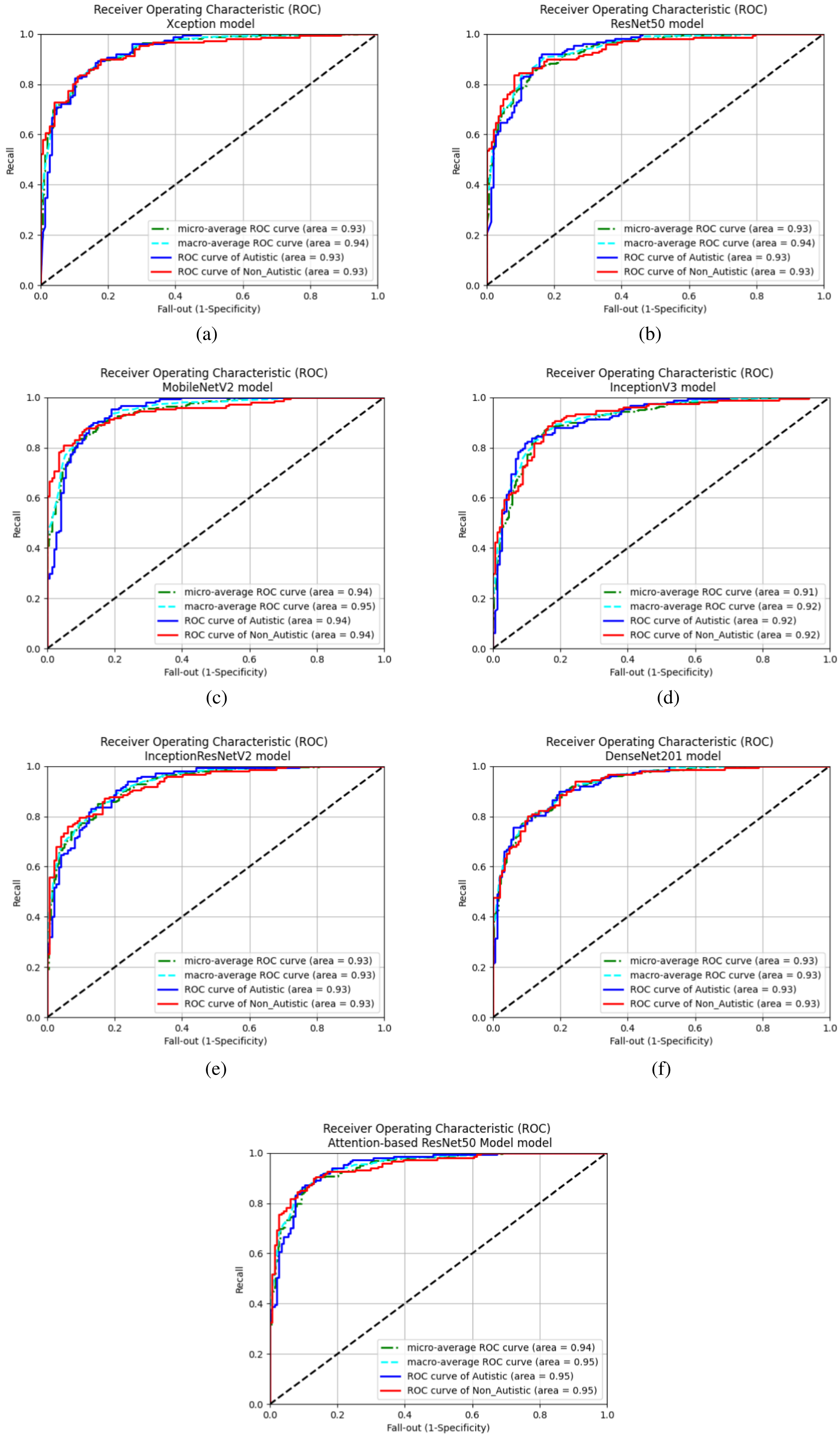
While the proposed framework achieves promising results, several limitations should be acknowledged. The dataset used in this research has significant geographic and demographic limitations, as the majority of the 2,940 facial images are sourced from children in Europe and the United States, with limited representation from other global regions. This limitation restricts the model's generalizability to diverse populations and may introduce sampling bias in the diagnostic framework across different cultural and ethnic backgrounds. Furthermore, the limited data volume may not provide sufficient samples for comprehensive model training and reliable performance estimation, particularly when compared to larger datasets used in similar deep learning applications for medical diagnosis.

The decision to forego data augmentation, while made to avoid introducing low-quality or semantically inconsistent samples, may have limited the model's robustness to variations in image conditions. The potential limitations of data augmentation include: (1) aggressive data augmentation can heighten the risk of overfitting when it is too closely compared to the original, thus restricting generalization; (2) data augmentation can introduce semantically inconsistent samples, potentially impacting the dataset's overall quality; and (3) the effectiveness of augmentation techniques varies, with some methods unintentionally altering the original data and reducing the accuracy of the resulting samples.

The clinical applicability of using facial features as biomarkers for ASD diagnosis requires further validation in real-world clinical settings. Although the proposed model achieves an accuracy of 87. 7%, a misclassification rate of 12.3% is a significant concern for real-world clinical applications. Furthermore, the underlying premise—that facial phenotypes can serve as reliable biomarkers for ASD—requires further validation through large-scale clinical trials.

Future research should address these limitations through several key avenues:

1. **Dataset Enhancement and Diversity:** To overcome data limitations, future work should focus on developing larger, more demographically diverse datasets. To address privacy and accessibility challenges, federated learning presents a powerful solution. This approach would enable collaborative model training across multiple institutions without centralizing sensitive patient data, leading to a more robust and generalizable model by leveraging diverse, multi-institutional datasets while upholding strict privacy standards.
2. **Advanced Data Augmentation:** To improve model robustness, advanced data augmentation strategies should be explored. Beyond simple transformations, techniques like Generative Adversarial Networks (GANs) could be used to synthesize realistic facial images, particularly for underrepresented demographics, thereby improving dataset diversity and model fairness without compromising data quality.
3. **Clinical Validation and Explainability:** For safe clinical implementation, enhancing model transparency through Explainable AI (XAI) is crucial. Future work should employ tools like Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the model's decision-making process. By generating heatmaps that highlight the most influential facial regions for a given prediction, we can build clinical trust, provide insights for clinicians, and potentially uncover novel, data-driven information about the specific facial phenotypes of ASD. This, combined with comprehensive clinical validation studies, will be essential for the responsible translation of this technology into practice.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 5. Evaluation and testing phase results. (**a**) Xception, (**b**) ResNet50, (**c**) MobileNetV2, (**d**) InceptionV3, (**e**) InceptionResNetV2, (**f**) DenseNet201, (**g**) Attention-based Residual Model.

Conclusion

Autism spectrum disorder (ASD) is a neurological condition that affects cognitive abilities, physical skills, and social engagement. In the absence of a specific medication for ASD, early intervention is crucial. The difficulty in diagnosing ASD, primarily due to the lack of objective biomarkers, underscores the need for effective and automated detection methods. Deep learning models have shown promise in medical diagnosis by analyzing diverse contexts of datasets, such as detecting cancer through histopathological images and diagnosing cardiovascular diseases using electrocardiograms. In the field of ASD, deep learning is crucial due to the diverse symptoms among individuals, and it can analyze biomarkers, speech patterns, behavioral data, and imaging to detect subtle patterns for an earlier and more accurate diagnosis. Hence, in this study, we proposed a framework that aims to detect ASD using an attention-enhanced residual and BiLSTM model. The process involves preprocessing input images to enhance clarity. Residual layers extract deep features, enabling high performance without requiring an excessively large training dataset. A BiLSTM model is integrated to handle semantic sequences among features. An attention mechanism is applied to determine the most relevant parts of input features. The refined features are then used in the classification phase to identify ASD. This multi-phase pipeline significantly improves detection and recognition performance.

The proposed model showed significant performance on the training set, achieving 100% accuracy by epoch 5. Most models reached a validation accuracy near 83%, but the proposed model stood out by reaching 85.4% accuracy early on. This indicates that it converges faster and maintains strong validation performance. The proposed model shows promise and effectively diagnoses autism in children early on with high accuracy in the training and validation phase. In experiments using a test set of images, the proposed model was compared to baseline techniques to assess its performance. The validation process tested the model's ability to work on data it had not seen during training. Results showed the proposed model outperformed the baseline, with average scores of 87.5% for precision, 87% for recall, 87.5% for F1 score, and 87. 7% for accuracy. Compared to the top-performing baseline, MobileNetV2 (86.5% average accuracy), our proposed model demonstrated superior accuracy (87.7%) in identifying autism-related features, suggesting its potential to enhance early detection and diagnosis. The true significance of this work lies in its potential for real-world applicability. The high performance and manageable computational complexity of our model make it a strong candidate for deployment as a low-cost, non-invasive, and highly accessible screening tool. Such a technology could be integrated into mobile health (mHealth) applications, allowing for preliminary screening by pediatricians, parents, or community health workers using a standard smartphone camera. By facilitating earlier identification of at-risk children, our work aims to bridge the gap between initial concern and formal diagnosis, ultimately enabling access to crucial early interventions and improving long-term developmental outcomes.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Code availability

The code generated during and/or analysed during the current study is available from the corresponding author on reasonable request.

Materials availability

The materials generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Received: 8 July 2025; Accepted: 5 September 2025
Published online: 29 September 2025

References

1. Rahmadianto, R. C., Febriyana, N. & Irmawati, M. Autism spectrum disorder: A literature review. *Autism Spectrum Disorder: A Literature Review* **140**, 6–6 (2024).
2. Licea, A. D. A Historical examination of the autism spectrum disorder diagnosis: a systematic literature review. Ph.D. thesis, Angelo State University (2024).
3. Nazir, A., Hussain, A., Singh, M. & Assad, A. Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis. Multimedia Tools and Applications 1–64 (2024).
4. Cheekaty, S. & Muneeswari, G. Early detection of autism spectrum disorder in children: A review. In AIP Conference Proceedings, vol. 3044 (AIP Publishing, 2024).
5. Uddin, M. Z. et al. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Engineering Applications of Artificial Intelligence* **127**, 107185 (2024).
6. Alkhawaldeh, R. S. & Al-Dabet, S. Unified framework model for detecting and organizing medical cancerous images in iomt systems. *Multimedia Tools and Applications* **83**, 37743–37770 (2024).
7. Alkhawaldeh, R. S. et al. Convolution neural network bidirectional long short-term memory for heartbeat arrhythmia classification. *International Journal of Computational Intelligence Systems* **16**, 197 (2023).
8. Khawaldeh, S., Pervaiz, U., Rafiq, A. & Alkhawaldeh, R. S. Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. *Applied Sciences* **8**, 27 (2017).
9. Khan, K. & Katarya, R. Mcbert: a multi-modal framework for the diagnosis of autism spectrum disorder. *Biological Psychology* **194**, 108976 (2025).
10. Khan, K. & Katarya, R. Ws-bitm: Integrating white shark optimization with bi-lstm for enhanced autism spectrum disorder diagnosis. *Journal of Neuroscience Methods* **413**, 110319 (2025).
11. Khan, K. & Katarya, R. Aff-bpl: An adaptive feature fusion technique for the diagnosis of autism spectrum disorder using bat-pso-lstm based framework. *Journal of Computational Science* **83**, 102447 (2024).

12. Uddin, M. Z. et al. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Engineering Applications of Artificial Intelligence* **127**, 107185 (2024).
13. Akter, T. et al. Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain sciences* **11**, 734 (2021).
14. Alsaade, F. W. & Alzahrani, M. S. [retracted] classification and detection of autism spectrum disorder based on deep learning algorithms. *Computational Intelligence and Neuroscience* **2022**, 8709145 (2022).
15. Hosseini, M.-P., Beary, M., Hadsell, A., Messersmith, R. & Soltanian-Zadeh, H. Retracted: Deep learning for autism diagnosis and facial analysis in children. *Frontiers in Computational Neuroscience* **15**, 789998 (2022).
16. Alam, M. S. et al. Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach. *Bioengineering* **9**, 710 (2022).
17. Cao, X. et al. Vitasd: Robust vision transformer baselines for autism spectrum disorder facial diagnosis. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5 (IEEE, 2023).
18. Rabbi, M. F. et al. Autism spectrum disorder detection using transfer learning with vgg 19, inception v3 and densenet 201. In International Conference on Recent Trends in Image Processing and Pattern Recognition, 190–204 (Springer, 2023).
19. Alkahtani, H., Aldhyani, T. H. & Alzahrani, M. Y. Deep learning algorithms to identify autism spectrum disorder in children-based facial landmarks. *Applied Sciences* **13**, 4855 (2023).
20. Mujeeb Rahman, K. & Subashini, M. M. Identification of autism in children using static facial features and deep neural networks. *Brain Sciences* **12**, 94 (2022).
21. Khan, B. et al. Autism spectrum disorder detection in children via deep learning models based on facial images. Bulletin of Business and Economics (BBE) **13** (2024).
22. Singh, J. . K. & Kakkar, D. Chronological sewing training optimization enabled deep learning for autism spectrum disorder using eeg signal. *Multimedia Tools and Applications* **83**(30), 1–28 (2024).
23. Reddy, P. Diagnosis of autism in children using deep learning techniques by analyzing facial features. *Engineering Proceedings* **59**, 198 (2024).
24. Bawa, P., Kadyan, V., Mantri, A. & Vardhan, H. Investigating multiclass autism spectrum disorder classification using machine learning techniques. e-Prime-Advances in Electrical Engineering, Electronics and Energy **8**, 100602 (2024).
25. Amelio, A. et al. Representation and compression of residual neural networks through a multilayer network based approach. *Expert Systems with Applications* **215**, 119391 (2023).
26. Febrian, R., Halim, B. M., Christina, M., Ramdhan, D. & Chowanda, A. Facial expression recognition using bidirectional lstm-cnn. *Procedia Computer Science* **216**, 39–47 (2023).
27. Misgar, M. M., Mushtaq, F., Khurana, S. S. & Kumar, M. Recognition of offline handwritten urdu characters using rnn and lstm models. *Multimedia Tools and Applications* **82**, 2053–2076 (2023).
28. Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports* **14**, 6086 (2024).

Author contributions

"Conceptualization, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh; methodology, Rami S. Alkhawaldeh, Samar M. Alkhawaldeh; software, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh; formal analysis, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh, Jamil AlShaqsi; investigation, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh, Osama Drogham; resources, Rami S. Alkhawaldeh Samar M. Alkhawaldeh; data curation, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh, Jamil AlShaqsi, Osama Drogham; writing—original draft preparation, Rami S. Alkhawaldeh, Samar M. Alkhawaldeh; writing—review and editing, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh, Jamil AlShaqsi, Osama Drogham; visualization, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh; supervision, Rami S. Alkhawaldeh, Bilal Al-Ahmad, Samar M. Alkhawaldeh, Jamil AlShaqsi, Osama Drogham; project administration, Rami S. Alkhawaldeh; funding acquisition, Rami S. Alkhawaldeh, Jamil AlShaqsi. All authors have read and agreed to the published version of the manuscript."

Declarations

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.S.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.