

EXPLAINABLE MULTI-TASK VISION TRANSFORMER FOR AUTOMATED AUTISM SPECTRUM DISORDER DETECTION FROM FACIAL IMAGES

Dr. Hira Chaudhry¹, Abu Bakar Shabbir², Faisal Amjad^{*3}, Khadija Amber⁴

¹MBBS, MSc. (Psychological Medicine), Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Malaysia

¹Assistant Professor, Tashkent State Medical University, Uzbekistan

²School of Systems and Technology, University of Management and Technology, Lahore - 54770, Punjab, Pakistan

^{*3}PhD Candidate (Special Education), Department of Special Education (DSE), Division of Education (DoE), University of Education (UoE), Lahore, Punjab, Pakistan

⁴Department of Computer Science, Bahauddin Zakariya University, Bosan Road, Multan - 60800, Punjab, Pakistan

¹hiraa.co.uk@gmail.com, ²abubakarshabbir64@gmail.com, ^{*3}amjadfaisal40@gmail.com/
pds24000237@ue.edu.pk, ⁴khadijaamber44@gmail.com

¹<https://orcid.org/0009-0009-2640-738X>, ³<https://orcid.org/0009-0002-7617-5395>

DOI: <https://doi.org/10.5281/zenodo.23102399>

Keywords

Autism Spectrum Disorder (ASD); Early Diagnosis; Autism Screening; Multi-Task Learning; Vision Transformer (ViT); DeiT; Facial Image Analysis; Emotion Recognition; Explainable Artificial Intelligence (XAI); Attention-Based Explainability; Deep Learning; Medical Artificial Intelligence.

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 02 October 2026

Copyright @Author

Corresponding Author: *
Faisal Amjad

Abstract

Timely interventions and cognitive, social, and emotional growth among children with Autism Spectrum Disorder (ASD) can only be achieved by early and accurate diagnoses of the condition. Traditional methods of diagnostic like behavioral testing and clinical observations are usually subjective, lengthy and require specialists. This paper suggests a Multi-Task Vision Transformer (ViT) model of automated ASD recognition in face images. The model classifies and recognizes auxiliary emotions together with ASD through a common DeiT based transformer backbone, which improves feature representations and generalization. Attention based explainability identifies the regions of a face that make the predictions and enhance transparency and clinical understandability. The model is compared to the conventional convolutional neural networks (VGG16, ResNet18) and single-task ViTs, which prove worse performance and stability. The proposed approach should be offered as a scalable, non-invasive, and interpretable solution to screening ASD, integrating multi-task learning and transformer-based feature extraction with explainable AI, which can be used both in the academic and medical contexts.

INTRODUCTION

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder that is marked by remainder difficulties in social communication, limited interests, and repetitive

behavior. The recent global estimates indicate that the occurrence of ASD has grown much in the last twenty years, which shows the acute necessity of creating early and accurate diagnostic tools that will be helpful in providing

timely intervention and enhanced developmental outcomes Farhat et al., 2024. Early diagnosis is important in facilitating individual behavioral therapy, educational planning, and social support, which positively impact the cognitive, emotional, and social development of the afflicted children (Dass et al., 2024). Diagnostic processes of ASD traditionally are mainly based on behavioral scales, questionnaires given to the caregivers, and clinical observations by the trained specialists. Although these approaches are still the clinical gold standard, they are time-consuming, subjective, resource-intensive, and extremely reliant on the availability of experts (Elbattah et al., 2022). In addition, trained clinicians are not available in most areas, especially the countries that are low- and middle-income, hence delay in diagnosis and intervention. Such restrictions have encouraged scientists to identify automated, objective, and scalable methods of screening ASD with the help of artificial intelligence (AI) and machine learning. Recent developments in deep learning (DL), especially in the field of computer vision and pattern recognition, have made it possible to analyze behavioral and physiological indicators of ASD using data. Of them, facial image analysis is a potentially compelling instance of a non-invasive modality because ASD-influenced neurodevelopmental processes are manifested through facial morphology and expressions (Ibadi & Lakizadeh, 2025). Research has proven that latent craniofacial characteristics like facial symmetry, eye spacing, mouth curvature may be used as discriminative biomarkers to identify ASD through deep neural network analysis (Dass et al., 2024; Rezaee et al., 2025). Convolutional Neural Networks (CNNs) have been popular in the field of facial-based ASD classification because they are effective in pointing out hierarchical spatial features. VGG16, ResNet, and Xception are models that have demonstrated promising results and reported accuracy of over 90 percent on medium sized datasets (Dass et al., 2024; Awaji et al., 2023). The CNNs, however, also clearly have their limitations due to local receptive field, and hence, they might not be able to effectively capture long-range dependencies and global face relationships needed to capture fine-grained

ASD-related patterns. This has led to the investigation of transformer-based architecture in the ASD research. Vision Transformers (ViTs) represent a solution that creates a paradigm shift, as images are represented by arrays of patches and the mechanism of self-attention is used to depict the contextual connections between the entire facial structure. Recent works demonstrated that ViTs can be more effective than CNNs in detecting facial ASD because they learn more holistic and richer features. Ibadi and Lakizadeh (2025) suggested ASDvit, a ViT boosted with Squeeze-and-Excitation blocks, which demonstrated the state-of-the-art performance on a public ASD facial dataset. Likewise, Rezaee et al. (2025) showed that multi-scale ViT architectures can be useful in capturing global and fine-grained craniofacial cues, and they are computationally efficient to deploy as an edge. The findings also highlight the appropriateness of transformer-based models in the analysis of faces related to ASD. Regardless of these developments, most current methods deal with the single-task learning procedure, in which models are trained to discriminate between autistic and non-autistic people. These methods might ignore complementary behavioral signals that would add to the feature representations and generalization. The idea of multi-task learning (MTL) provides a strong solution as models can acquire several related tasks at once with shared representations. Previous studies have demonstrated that auxiliary tasks may serve as a type of regularization, which minimizes overfitting and increases robustness especially in medical imaging models with small amounts of data (Altomi et al., 2025). Emotion recognition is a useful auxiliary task in terms of ASD detection. Facial emotions and emotional reactions are not typical in patients with ASD and, thus, facial expressions have a significant role in the classification of emotions (Kasri et al., 2025). By adding emotion recognition to ASD detection systems, models are able to appreciate subtle patterns of emotions and behaviour that may be not explicitly described in ASD datasets. Although the multimodal methods based on eye-tracking, speech, and gaze patterns have shown high effectiveness (Elbattah et al., 2022; Kasri et al., 2025), they are often evaluated as expensive

in terms of hardware and data processing, which hinders scalability. Conversely, there is an alternative, which is a lightweight and scalable form of auxiliary emotion learning based on stationary facial images. Interpretability is another important need of the AI-based ASD screening systems. Generally, black-box prediction may be harmful because clinical trust and adoption are impeded in sensitive healthcare and education settings. To achieve transparency, explainable AI (XAI), including attention visualization and saliency mapping, show the areas that affect model choices in an image (Hossain et al., 2025; Alam et al., 2023). Vision Transformers. Vision Transformers naturally provide explainability via self-attention largely making them especially appealing to clinically meaningful deployment. It is also promoted by ethical aspects, as transparency is necessary after doubts about the validity of studies on unverified or unethical datasets were expressed (Atlam et al., 2025). These findings inspired the research proposed in this paper, which aims to present an end-to-end Multi-Task Vision Transformer (ViT) architecture to find ASD in facial images. The offered method learns both the classification of ASD and auxiliary emotion recognition on a common DeiT-based transformer backbone. The framework program is expected to enhance classification performance, decrease false negatives, and increase interpretability without using extra data modalities or the use of specialized acquisition devices by using multi-task learning and attention-based explainability.

Literature Review

Facial Image-Based ASD Detection

Facial morphology offers non-invasive biomarkers that can detect ASD, and they can be assessed automatically by analyzing the images. Several studies have researched on deep learning methods with facial databases to derive minute craniofacial patterns related to ASD. ASDvit is a Vision Transformer (ViT) framework proposed by Ibadi and Lakizadeh (2025), which includes the addition of Squeeze-and-Excitation (SE) blocks to the original model to predict facial features in their spatial and channel-wise dependencies. ViT transfer learning on ImageNet-21k enhanced the quality

of convergence and feature representation showing an AUC of 97.77% on a publicly available dataset of 2,926 images of autistic and typically developing (TD) children. On the same note, Rezaee et al. (2025) used a multi-scale Vision Transformer (MS-ViT) to store both the global craniofacial features and fine details of the faces, which had 96.85% accuracy upon after-train optimization to deploy on Raspberry Pi devices. These investigations point to the usefulness of architecture transformer-based architectures in terms of rich feature analysis of faces with regard to computational efficiency and data scarcity. This was complemented by the work of Dass et al. (2024), who used the already trained convolutional neural networks (CNNs), namely Xception and VGG16, as the detectors of ASD using 2,940 facial images. The authors proved that CNNs are able to identify discriminative facial biomarkers effectively, with Xception having an accuracy of around 91-94 percent. Equally, a comparative study was performed by Farhat et al. (2024) on deep neural network classifiers to detect ASD based on facial image experimenting on VGG16 and MobileNet architectures used on a home-based dataset collected in Pakistan. The analysis obtained a validation and test accuracy rate of 99 and 87 respectively, thus pointing to the significance of culture and demographically aligned datasets in screening ASD. Awaji et al. (2023) also investigated hybrid CNN-ML systems, in which the features of VGG16, ResNet101, and MobileNet were used with classical machine learning classifiers, including Random Forest and XGBoost. The best performance was obtained with feature fusion strategies with an accuracy of 98.8% and AUC of 99.25, highlighting the advantages of using more than one CNN architecture to obtain better discriminative ability. Other hybrid transformer-CNNs, such as Z.A. Altomi et al. (2025) used outputs with NasNetMobile and DeiT and attentional feature fusion before a bagging based SVM classifier. According to their framework, recall and accuracy were 95.77 and 95.67 respectively, showing that global-local combination of representation is better at making models robust. Hossain et al. (2025) and Alam et al. (2023) also incorporated explainable AI (Grad-CAM, LIME) to show salient areas of

the face, which they claim can give clinicians interpretable information about the decision made by the model.

Multi-Modal and Eye-Tracking Approaches

Although facial image analysis gives structural information, eye-tracking results give behavioral and attention patterns that are associated with ASD. Elbattah et al. (2022) proposed a pipeline that converts raw scanpaths of eye-tracking to synthetically generated visual images to classify them using CNNs (VGG-16, ResNet-50, DenseNet-121). The approach has an ROC-AUC of 0.78 using a dataset of 59 children, which shows that facial analysis can be complemented by eye movement patterns. Kasri et al. (2025) expanded it by introducing a hybrid spatiotemporal deep learning model based on Vision Transformer and Vision Mamba. Their framework enhanced the classification accuracy and interpretability by merging multimodal cues (gaze, facial and speech features) through attention-based fusion. These multi-modal methods overcome one of the major challenges of facial analysis in single modality, which is the frames of dynamic behavioral patterns, which improve generalization and strength. They are however usually sensitive to specialized data acquisition equipment hence they are not scalable in real-world screening situations.

Domain Adaptation, Ensemble Learning, and Data Generalization

One of the drawbacks of the AI implementation of ASD detection is the demographic, environmental, and image generalization of models. Alam et al. (2025) have solved this challenge with ASD-UANet, a domain-adaptive ensemble network that consists of Xception and ResNet50V2 models with a Fifty Percent Priority (FPP) weighting algorithm. The model also showed strong results in domain shift resistance, which generates the significance of multi-domain training, as it is trained on multiple datasets (Kaggle ASD, YTUIA) and evaluated on a clinically validated UIFID dataset. Such ensemble learning techniques are useful to decrease the bias and overfitting of models, as well as combine the complementary advantages of diverse structures. The use of transfer learning, which is especially with large

pretrained Vision Transformers, to counter small datasets is also highlighted by Rezaee et al. (2025) and Ibadi and Lakizadeh (2025). Together with data augmentation and cross-validation, these methods will make the learning process stable even when the amount of labeled medical images is low.

Explainable AI for ASD Detection

The applicability and reliability of AI models to diagnose ASD are dependent on open decision-making. Agrawal et al. (2025) overviewed the explainable AI (XAI) methods, including Grad-CAM, LIME, SHAP, and saliency maps, in early ASD detection systems. The visualization of the main parts of the face allows the clinicians to confirm that the models focus on the diagnostically significant elements of the face, including eyes, mouth, and facial symmetry. Hossain et al. (2025), Alam et al. (2023), and Altomi et al. (2025) used XAI as a method to validate the model predictions and justify the interpretability, which is especially essential when the models are used in sensitive healthcare or teaching settings. Shabbir (2025) suggested a SHAP-weighted hybrid ensemble based framework to make clinical risk predictions, including explainable AI to make interpretable predictions, in a related context. Through this method, it is evident that XAI-enhanced ensemble models can provide credible, explainable results in the healthcare domain, which can be used to inform analogous approaches to the ASD detection systems. On the contrary, Atlam et al. (2025) tried to explain explainable ASD screening via the use of facial images that were scraped online but due to ethical and methodological flaws, unconfirmed diagnoses, and consent, they were withdrawn. This highlights the importance of strong ethical gathered datasets to make credible XAI results.

Comparative Reviews and Trends

In addition to model-specific works, Rakotomanana & Rouhafzay (2025) performed a comprehensive scoping review of the AI-based ASD detection in relation to the behavioral and voice modalities in 158 studies. Their review revealed eight key modalities, with voice biomarkers, face gestures, attention to visual objects, and multimode combination. The

review has highlighted the relevance of large and balanced datasets, ethical concerns, and model interpretability. The results indicate that voice and visual attention hold potentials as scalable ASD screening because data collection is easy, and reported accuracies are high but points to the systemic issues of heterogeneity of data and low generalizability.

Limitations and Gaps

Even with the important advancements, there are a number of gaps in the research regarding ASD detection using AI. Most research uses medium-sized datasets (2,500-3,000 images), and this may be limiting to deep model generalization (Dass et al., 2024; Ibadi and Lakizadeh, 2025). Single-modality measures can

be effective but can miss behavioral signals as measured by eye-tracking, speech or multimedia (Kasri et al., 2025). The issue of ethics and privacy is crucial, as the retraction of Atlam et al. (2025) pointed out. Also, there are not many studies on the evaluation of the real world deployment, edge computing viability, or cross-population generalization, though Rezaee et al. (2025) and Alam et al. (2025) partially cover the matters. Table 1 is a summary of current research on the topic of ASD detection, including datasets, methodology, and AI/deep learning tools. It also points towards the variety of methods, such as Vision Transformers, CNNs, hybrid models, and explainable AI methods.

Table 1: Summary of recent studies on ASD detection and analysis, including datasets, methodologies, and tools used.

Author(s)	Dataset	Methodology	Tools / Frameworks
Ibadi & Lakizadeh (2025)	2,926 facial images (ASD vs TD)	ViT-B + Squeeze-and-Excitation, transfer learning	PyTorch, ViT (ImageNet-21k), CUDA, scikit-learn
Rakotomanana & Rouhafzay (2025)	158 studies (voice, speech, video, eye-tracking, behavioral)	Scoping review of AI-based ASD detection	SVM, RF, CNN, RNN, Transformers, NLP, signal processing
Elbattah et al. (2022)	Eye-tracking data, 59 children	Scanpath visualization + CNNs (VGG16, ResNet50, DenseNet121)	Python, Matplotlib, OpenCV, Keras, TensorFlow
Kasri et al. (2025)	Saliency4ASD (eye-tracking)	Hybrid spatiotemporal DL + multimodal attention fusion	ViT, Vision Mamba, Attention Mechanism
Rezaee et al. (2025)	2,940 facial images (Kaggle)	ROI preprocessing + MS-ViT, edge deployment optimization	Deep Learning, ViT, Edge AI, Raspberry Pi
Dass et al. (2024)	2,940 Kaggle facial images	CNN + transfer learning	Xception, VGG16, TensorFlow, Keras
Awaji et al. (2023)	2,940 facial images (Kaggle)	Hybrid CNN + ML classification	VGG16, ResNet101, MobileNet, XGBoost, RF, t-SNE
Altomi et al. (2025)	Autism Facial Image Dataset	Dual-network fusion + attention + SVM	NasNetMobile, DeiT, Bagging-SVM
Alam et al. (2025)	Kaggle ASD, YTUIA, UIFID	Domain-adaptive deep ensemble + FPPR weighting	Xception, ResNet50V2, MobileNetV2, MTCNN, ASD-UANet
Atlam et al. (2025)	Internet-scraped facial images	DL-based facial analysis + XAI	CNNs + XAI techniques

Agrawal et al. (2025)	Behavioral, genetic, neuroimaging, clinical	Literature review of AI & XAI for ASD	SVM, RF, DT, ANN, CNN, XGBoost, SHAP, LIME, Grad-CAM
Hossain et al. (2025)	2,950 facial images	Transfer learning with 12 pre-trained CNNs, DenseNet121 selected, XAI	Python, TensorFlow, Keras, Grad-CAM, LIME
Alam et al. (2023)	3,014 Kaggle facial images	Data-centric DL + augmentation + XAI	Python, Keras/TensorFlow, MTCNN, MobileNetV2, ResNet50V2
Farhat et al. (2024)	2,550 images (Kaggle + local)	VGG16 + MobileNet, 5-fold CV	Python, Keras, Orange, Google Colab
Shabbir, A.B. (2025)	Brain Stroke dataset, n=4,981	SHAP-weighted hybrid ensemble (XGBoost + RF + TabNet)	Python, Scikit-learn, XGBoost, PyTorch-TabNet, SHAP, LIME

Dataset Description

The current study makes use of available in the public facial image data set of children meant to be used in Autism Spectrum Disorder (ASD) classification tasks. The data set is comprised of frontal facial images that have been obtained under controlled and semi-controlled settings, and they have found extensive applications in the recent computer vision research studies that relate to ASD. All images are fixed RGB images, which guarantees consistency of images in samples and the applicability of the image-based deep learning models. Anonymization and non-identifiability of subjects are ethical concerns that are inherent to the guidelines of public release and use of the dataset. The data is sorted into two major categories: Autistic and Non-Autistic that denote children with ASD and children that develop normally, respectively. The dataset has a perfectly balanced class distribution to avoid bias more on the class

imbalance. There are 2,936 images in total, 1,468 images in each of them. The sample images of frontal faces of the ASD and Non-ASD classes are presented in Figure 1. The images are anonymized and emphasize minor craniofacial variations that will be used to detect ASD. The idea of a strict and non-overlapping data split strategy is used to support strong model evaluation and generalization assessment. The dataset is split into training, validation and test datasets. Training set consists of 2,536 images (1,268 each class) which represent roughly 86 percent of the entire data, and on which the models are learnt. Validation set: This set consists of 100 images (50 images per category) to be used in tuning hyperparameters and early selection of models. The test set will include 300 images (150 per class) that will be used as independent test set and will only be used to examine its final performance to prevent any kind of leakage of data.

Table 2: Class-wise distribution of the Autism Spectrum Disorder (ASD) facial image dataset across training, validation, and test splits.

Split	ASD (Autistic)	Non-ASD (Typical)	Total
Training	1268	1268	2536
Validation	50	50	100
Test	150	150	300
Total	1468	1468	2936

Table 2 presents the detailed class-wise distribution across all dataset splits. The balanced nature of the dataset enables unbiased learning and reliable metric computation,

particularly for recall and F1-score, which are critical in medical and behavioral screening tasks. While the dataset size is moderate, the use of transfer learning with large-scale pretrained

Vision Transformer models effectively mitigates data scarcity challenges. This dataset configuration reflects realistic constraints

common in healthcare and educational AI research while supporting meaningful and reproducible experimental outcomes.



Figure 1: Sample facial images from the ASD dataset. Left column: ASD class; Right column: Non-ASD class. Images are anonymized for privacy.

Data Preprocessing and Augmentation

Before model training, all images go through a preprocessing pipeline that has been standardized so that they can be compatible with the Vision Transformer architecture and enhance training stability. All pictures are scaled to a constant resolution of 224×224 pixels, which corresponds to the size of the input of the pretrained DeiT-based Vision Transformer backbone. This transformation tool maintains the same spatial dimension to all of the dataset and maintains key features of the faces. After the resizing process, each of the images is transformed into normalized tensors. ImageNet mean and standard deviation statistics are used to normalise pixel values, as is the case with ImageNet pretraining setup of the models used. The step of normalization facilitates faster convergence, stabilizes the gradient change, and enables the successful reuse of pretrained weights without any distribution mismatch. Lightweight data augmentation methods are only employed to the training set in order to achieve better generalization and reduce overfitting. They are presented with the introduction of horizontal flipping, which is introduced with a predetermined probability in order to model the natural left-right facial orientation variations normally seen in the real-life situation. This augmentation plan promotes the model to acquire the specialized insight of

the representations of faces instead of storing spatial preferences. It does not involve any aggressive augmentations, including random cropping or geometric distortions, since over-transforming can change minor facial features that can be used to detect ASD. The validation and test sets are carefully maintained without being augmented so that performance indicators reflect actual generalization of the model when facing unknown data. This division of training-time augmentation and evaluation-time preprocessing maintains the integrity of the experiments and is consistent with the best practice in analyzing medical images. The selected preprocessing and augmentation scheme has a tradeoff between robustness and realism such that the model can acquire discriminative facial patterns without sacrificing clinically meaningful visual features. It is specifically appropriate to transformer-based architectures, which have the advantage of input representations that are always the same and require only slight augmentation to be trained over moderately-sized datasets.

Proposed Methodology

Overall System Architecture

Our suggested end-to-end Multi-Task Vision Transformer (ViT) model can be used to detect Autism Spectrum Disorder (ASD) in facial images. The system uses a shared transformer-

based feature extractor and 2 task heads, one ASD classification head and another auxiliary emotion recognition head. The auxiliary emotion task promotes more detailed facial representations, as it involves observing finer patterns of emotion that are likely to be associated with ASD behavioral indicators and, thus, improves the classification task. The explainability based on self-attention also offers visual explainability, where a clinician and an educator can identify critical areas of the face that drive the decisions of the model. Preprocessing of raw facial images, such as resizing, normalization, and horizontal flipping, to augment the data, is the first step in the processing pipeline. Instead, the images are preprocessed and fed to the DeiT-base backbone, which encodes the images into high-dimensional feature representations by patch

embedding and multi-head self-attention. The common embeddings are subsequently inputted into ASD head to be categorized as binary and emotion head to predict auxiliary emotion. The two heads are jointly trained with a weighted loss to be able to optimize the primary and auxiliary tasks equally. A comparative evaluation model will compare the proposed approach with baseline architectures of deep learning, such as VGG16, ResNet18, and a single-task ViT that will be trained only on ASD labels. Each and every model is trained with the same preprocessing, batch size, and optimizer, which provides a fair and strict evaluation of the proposed system. Figure 2 depicts the general structure of the proposed Multi-Task Vision Transformer system, where the pre-processing of the input leads to two task-based heads and comparative analysis.

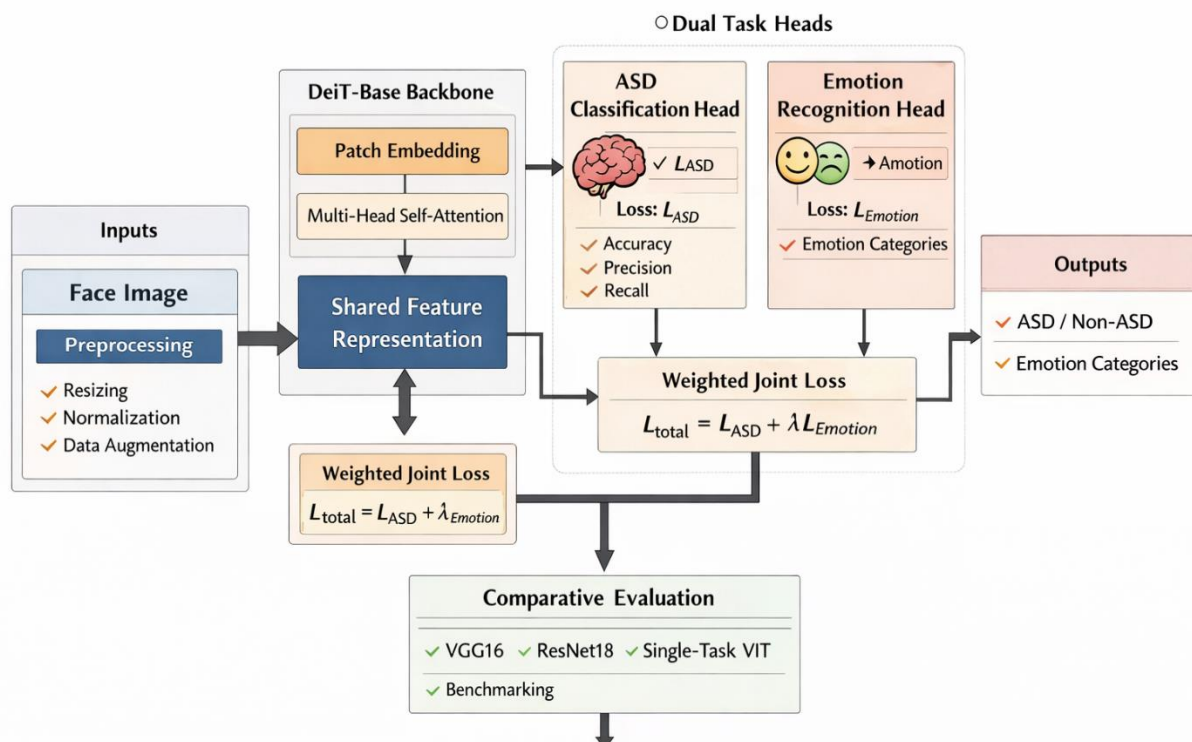


Figure 2: Overall system architecture of the proposed Multi-Task Vision Transformer (ViT) model for ASD and emotion detection, showing preprocessing, shared feature extraction, dual-task heads, weighted joint loss, and comparative evaluation. *Vision Transformer Backbone*

The system leverages the **DeiT-base (Data-efficient Image Transformer)** backbone, pretrained on ImageNet, to extract both local and global contextual information from facial images. Unlike conventional CNNs, ViTs divide

each image into non-overlapping patches (16×16 pixels in this study) and project them into a sequence of embeddings. These embeddings are processed via **multi-head self-attention layers**, which model interactions between all patches in

a global context. This approach enables the model to capture subtle facial features and long-range dependencies critical for distinguishing autistic from non-autistic facial expressions.

The transformer backbone produces a high-dimensional feature embedding shared across multiple tasks, encoding nuanced facial geometry, texture, and emotional cues. Importantly, the self-attention layers inherently provide attention weights, facilitating **visual explainability** by highlighting regions most relevant to the model's predictions.

Multi-Task Learning Framework

The multi-task framework consists of a **shared encoder** (the ViT backbone) and two task-specific heads:

1. **ASD Head:** A fully connected layer maps the shared embedding to a binary output

where $\mathcal{L}_{\text{ASD}}$ represents the cross-entropy loss for ASD classification, $\mathcal{L}_{\text{Emotion}}$ denotes the cross-entropy loss for auxiliary emotion prediction, and $\lambda = 0.5$ balances the contribution of the auxiliary task.

During the early training phase, a warm-up strategy is employed in which optimization focuses exclusively on $\mathcal{L}_{\text{ASD}}$. This allows the transformer backbone to learn robust ASD-specific representations before incorporating auxiliary emotion supervision, thereby improving convergence stability and generalization performance.

Training Strategy

The model is trained with the following specifications:

- **Optimizer:** AdamW, for adaptive learning and weight decay to reduce overfitting
- **Learning Rate:** $1e-4$, empirically determined for stable convergence
- **Batch Size:** 16, balancing GPU memory constraints with gradient estimation reliability
- **Epochs:** 20, sufficient for convergence without overfitting given the dataset size
- **Device:** NVIDIA GPU (CUDA) to accelerate transformer computations

During training, the backbone and both task-specific heads are updated jointly. Forward passes compute ASD and emotion predictions, which are combined via the weighted loss

(autistic vs. non-autistic). The primary task is optimized using a **weighted cross-entropy loss**, addressing potential class imbalance and ensuring balanced learning across ASD and non-ASD samples.

2. **Emotion Head (Auxiliary):** This head predicts emotion categories (e.g., neutral, happy, sad, surprised). While emotion labels are not perfectly granular, this auxiliary task provides additional supervision, promoting more discriminative and generalized feature learning. Cross-entropy loss is applied with a lower weight than the ASD loss to prevent overfitting.

Multi-Task Loss Function

The proposed framework jointly optimizes ASD classification and auxiliary emotion recognition using a weighted multi-task loss function. The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASD}} + \lambda \mathcal{L}_{\text{Emotion}}$$
 Gradients are backpropagated through both the shared encoder and heads to jointly optimize feature representations.

To prevent overfitting, data augmentation (horizontal flips) is applied to the training set, while validation and test sets remain unchanged for accurate generalization assessment. Checkpointing is implemented to save the best model based on validation accuracy for final evaluation.

Attention-Based Explainability

Attention visualization is a major characteristic of the methodology. The ViTs self attention mechanism generates per patch attention weights, which can be decoded to generate an attention overlay map on the original image. These overlays identify areas of the face that have the most impact in model predictions including the eyes, eyebrows, and the mouth. The explainability mechanism is used in two purposes:

1. **Clinical Trust:** Visualizations can enable clinicians and educators to confirm that the model is centered on purposeful facial expressions as opposed to unproductive areas.
2. **Model Debugging:** Attention maps: These maps can give details about misclassified samples and guide refinements to preprocessing, augmentation or auxiliary tasks.

Inference extracts attention overlays as the average of attention weights of all heads and layers, and scales them to a 01-range and overlays them onto the original image with a jet colormap.

Comparative Baselines

In order to confirm the usefulness of the proposed Multi-Task ViT, we conduct an extensive comparative analysis with three powerful baselines:

1. **VGG16:** This is a common CNN applied in the image classification process. It has local receptive fields and a hierarchical structure whose properties give a standard to the traditional convolution-based methods.
2. **ResNet18:** Residual CNN eliminates the problem of the vanishing gradient and enables the network to learn more, and it is a representation of modern CNN architecture.
3. **Single-Task ViT:** A Vision Transformer that is purely trained on the labels of ASD without additional emotion guidance is a direct ablation of the multi-task architecture design.

Every baseline is trained with the same preprocessing settings, optimizers, batch sizes

and epochs to make everything fair. Measurements of evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices identify the framework proposed as well as allow the evaluation of the framework to take quantitative and qualitative measures.

Experimental Setup

Evaluation Metrics

To assess the performance of the proposed Multi-Task Vision Transformer (ViT) for Autism Spectrum Disorder (ASD) detection and facial emotion prediction, a comprehensive set of standard classification metrics is employed. These metrics provide both overall predictive accuracy and detailed class-level performance insights.

Accuracy

Accuracy measures the proportion of correctly classified instances among all samples. While it provides an overall measure of model correctness, it may be insufficient for imbalanced datasets. Accuracy is defined as:

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

where TP denotes true positives (correctly predicted ASD cases), TN denotes true negatives (correctly predicted non-ASD cases), FP denotes false positives, and FN denotes false negatives.

Precision

Precision measures the proportion of correctly predicted positive samples among all predicted positives and is particularly important when the cost of false positives is high. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

Recall evaluates the proportion of correctly predicted positive samples among all actual

positives. High recall indicates fewer false negatives, which is critical in medical and educational screening tasks:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced metric when

both false positives and false negatives are important:

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

Confusion Matrix

A confusion matrix provides a detailed class-wise breakdown of model predictions, enabling

deeper analysis of misclassification patterns. For binary ASD classification, the confusion matrix is structured as shown in Table 3.

Table 3: Confusion matrix for binary ASD classification showing true/false positives and negatives

	Predicted ASD	Predicted Non-ASD
Actual ASD	TP	FN
Actual Non-ASD	FP	TN

The combination of these metrics ensures that model evaluation is **robust**, considering both overall accuracy and error distribution across classes.

Baseline Models

To validate the effectiveness of the proposed Multi-Task ViT, its performance is compared against several **widely-used CNN and transformer-based models**, which serve as baselines.

1. VGG16:

VGG16 is a traditional convolutional neural network (CNN) with 16 layers, which is characterized by a uniform architecture similar to small 3×3 convolutional kernels and deep sequencing of the network. Although it is effective in general image classification, VGG16 is mostly sensitive to local spatial features and it might not easily learn long-range dependencies spanning facial regions, which are usually important in identifying delicate ASD-related features. VGG16 in this study is trained on ASD data with extra fully connected layers to classify them as binary.

2. ResNet18:

ResNet18 is a residual CNN model that uses skip connections to overcome vanishing gradient issues to allow deeper network learning. The model is efficient in storing both low-level and high-level features due to its residual blocks. Although better than VGG16, ResNet18 also uses local receptive fields and cannot explicitly capture global dependencies, which restricts its capability to identify subtle emotional patterns related to ASD. Fair comparison is also done by fine-tuning ResNet18 on the same dataset.

3. Single-Task ViT:

A Vision Transformer used without the auxiliary task of emotion recognition, and this time it was

trained to classify ASD. It has an identical DeiT-base backbone as the suggested model, and it applies to images as a sequence of patches and learns the global relationships between contexts using multi-head self-attention. This control removes the effect of multi-task learning and proves the performance gain that occurs with the inclusion of the auxiliary emotion task.

4. Proposed Multi-Task ViT:

The proposed model is an end-to-end Multi-Task Vision Transformer that jointly learns ASD classification and facial emotion recognition. By leveraging shared embeddings, the model benefits from auxiliary supervision, which improves representation learning and generalization. Attention-based explainability further highlights critical facial regions influencing predictions, enhancing trustworthiness and interpretability. Comparative evaluation against the above baselines quantitatively and qualitatively demonstrates the effectiveness of the multi-task approach.

Results and Analysis**Quantitative Performance**

The given Multi-Task Vision Transformer (ViT) is shown to be more effective than the traditional CNN architectures (VGG16, ResNet18) and a one-task ViT baseline in detecting Autism Spectrum Disorder (ASD). In Table 4, quantitative analysis of the held-out test set is presented, and Accuracy, Precision, Recall, and F1-score are reported. VGG16 has an accuracy of 83.7, precision of 93.2, recall score of 72.7 and F1-score of 81.6. ResNet18 outperforms VGG16 in terms of accuracy of 87.7, precision of 94.5, recall of 80.0 and F1-score of 86.6 because of residual connections which enable the extraction of deeper features. Single-task ViT has an accuracy of 87.0,

precision of 90.5, recall of 82.7 and F1-score of 86.4 which is somewhat lower than the multi-task version. The Multi-Task ViT has the best overall result, an accuracy of 89.0, and a

precision of 94.0, a recall of 83.3 and F1-score of 88.3, which indicates better precision and recall balance, which are critical to sensitive clinical and educational screening.

Table 4: Quantitative performance comparison of ASD classification models, including Accuracy, Precision, Recall, and F1-Score.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.837	0.932	0.727	0.816
ResNet18	0.877	0.945	0.800	0.866
ViT (Single-Task)	0.870	0.905	0.827	0.864
ViT (Multi-Task)	0.890	0.940	0.833	0.883

Multi-Task ViT training dynamics Figure 3 demonstrates that the ability to optimize the gradient during training, illustrated by a progressive decrease in training loss over 20 epochs, appears to have occurred. Bigger oscillations are noted, although convergence is stable with AdamW optimization. Figure 4 shows that validation accuracy increases quickly

during early epochs and levels off at 84-89 percent indicating rapid convergence and stable overfitting. The initial ASD-only warm-up, which lets the model concentrate on the principal classification task and subsequently existence together to optimize the auxiliary emotion task, facilitates the secure and generalized learning.

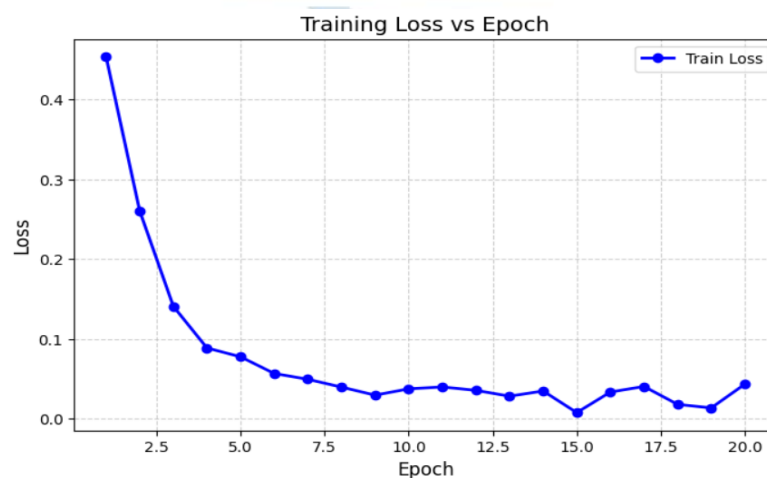


Figure 3: Training loss curve across epochs, showing model convergence during multi-task ViT training.

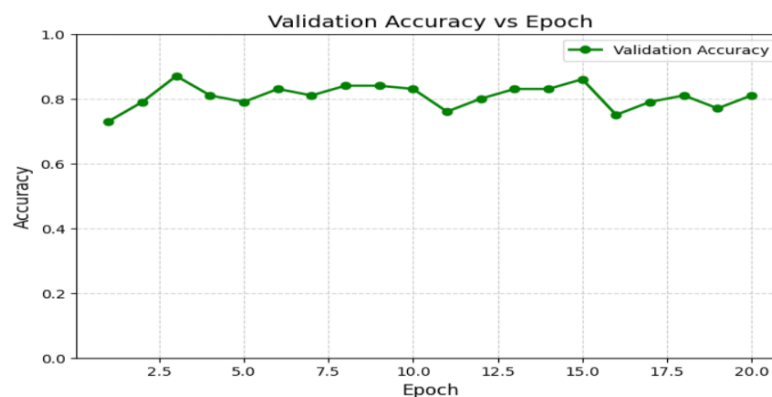
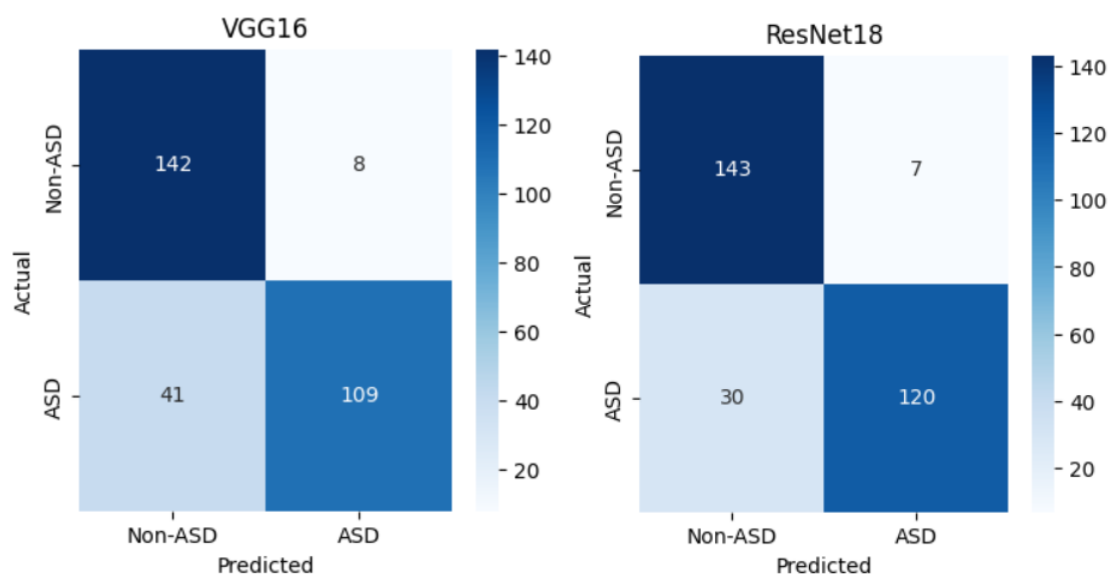


Figure 4: “Validation accuracy over epochs, illustrating the performance improvement of the proposed model on unseen data.”

Confusion Matrix and Robustness

Class-level performance is further analyzed using confusion matrices Figure 5. VGG16 exhibits higher false negatives, resulting in reduced recall, whereas ResNet18 reduces false negatives but retains minor false positives. The single-task ViT shows balanced performance but slightly lower recall than the multi-task variant. The Multi-Task ViT achieves the most balanced

confusion matrix, effectively minimizing false negatives while maintaining high precision, ensuring that fewer autistic cases are misclassified a critical requirement for early ASD detection. Attention-driven interpretability confirms that the model focuses on key facial regions such as the eyes, mouth, and forehead, validating the clinical relevance of predictions.



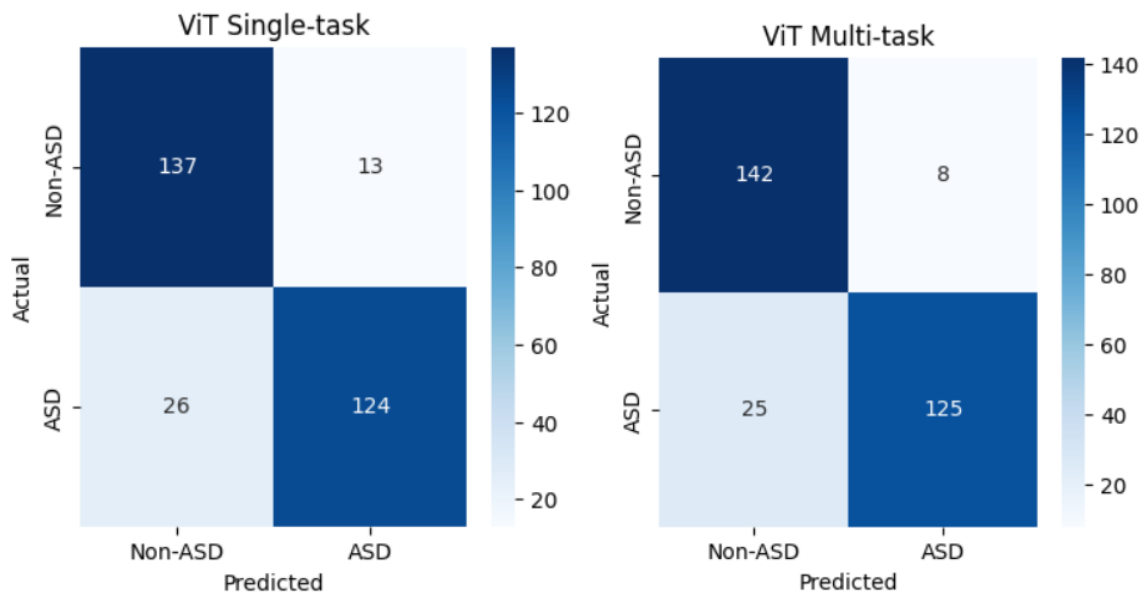


Figure 5: Confusion matrices of VGG16, ResNet18, single-task ViT, and Multi-Task ViT showing class-wise prediction performance for ASD detection.

To assess reliability and reproducibility, the Multi-Task ViT was trained three times with different random seeds, with performance metrics recorded in Table 5. Mean accuracy across runs is 88.7% (std dev 2.2%), precision 93.8% (std 1.4%), recall 82.9% (std 1.5%), and

F1-score 88.2% (std 1.3%). The low standard deviations indicate robust convergence, demonstrating consistent high performance under different initialization conditions and validating the effectiveness of multi-task learning with self-attention-based feature extraction.

Table 5: Performance metrics of the Multi-Task ViT over multiple training runs, showing mean values and standard deviations for Accuracy, Precision, Recall, and F1-Score.

Metric	Mean	Std Dev
Accuracy	0.887	0.022
Precision	0.938	0.014
Recall	0.829	0.015
F1-Score	0.882	0.013

Discussion

Interpretation of Key Findings

The suggested Multi-Task Vision Transformer (ViT) model shows greater advances in the detection of Autism Spectrum Disorder (ASD) in facial images, having a high level of performance compared to traditional convolutional neural networks (CNNs) and single-task transformer frameworks. Multi-Task ViT had an overall accuracy of 89.0, precision 94.0, recall 83.3, and F1-score 88.3, which means that it was able to perform equally well on both positive and negative classes. This enhancement shows the effectiveness of the

combination of auxiliary emotion recognition tasks into ASD detection systems that agree with previous works by Ibadi and Lakizadeh (2025) and Rezaee et al. (2025) that stated that transformer-based models are able to both learn global and local facial information. Auxiliary emotion task also leads to the increase in the discriminative power and promotes the shared transformer backbone to encode subtle emotional patterns which are usually distorted in children with ASD. The multi-task design has a significant effect on false negatives compared to the single-task ViT, which also enables a recall of 82.7 only, which is a serious factor in clinical

and educational screening, as an autistic child left undiagnosed can postpone interventions with potentially detrimental developmental effects. These results support the theoretical assumption that multi-task learning is a regularizer, that it favours more rich features and reduces overfitting as the evidence provided by Awaji et al. (2023) and Altomi et al. (2025) supports.

Transformer-Based Feature Learning

The effectiveness of ViT backbone as compared to the conventional CNNs like VGG16 and ResNet18 can be explained by the fact that the self-attention mechanisms have an innate capacity to capture long-range relationships among the facial parts. ViTs, unlike CNNs, which use hierarchical local receptive fields, allow global contextual learning, where even subtle craniofacial patterns are connected with ASD (Ibadi and Lakizadeh, 2025; Rezaee et al., 2025). This is especially beneficial in the extraction of ASD-related complex but distributed facial biomarkers that cannot be entirely realized with localized convolutional operations, including eye spacing, mouth curvature, and facial symmetry. In addition, the model can be improved by adding Squeeze-and-Excitation (SE) block, suggested by Ibadi and Lakizadeh (2025) to give the model greater sensitivity to channel-wise feature dependencies, which will enable the network to highlight salient features of the face. This addition of such architectural improvements is probably one of the reasons why there is high precision and low false positive rate in the multi-task model. These results confirm the possibility of real-time performance on medically relevant computer vision tasks with transformer-based frameworks augmented with attention-guided feature calibration to reach state-of-the-art results.

Multi-Modal and Auxiliary Learning

Advantages

Although this research mainly uses the static images of faces, the past research highlights the role of multi-modal integration. The study by Kasri et al. (2025) showed that the combination of data features (eye-tracking, speech, and facial expression) increases the accuracy and strength of a classification. Multi-modal acquisition is

logistically difficult, but the auxiliary emotion recognition task in the present study provides a computationally efficient approximation such that the model can generate dynamic emotion evidence using only static images. This approach somewhat overcomes the restrictions of single-modality analysis as pointed out by Elbattah et al. (2022) and supplements the results of Dass et al. (2024), who posited that more effective feature representations enhance the generalization of the model without necessitating extra data gathering systems. The auxiliary task also helps in explainability whereby it shows facial areas that are important in ASD classification as well as emotion recognition. The concept of attention overlays suggests that the model centers on the diagnostically meaningful features, i.e., eyes, eyebrows, and mouth, which are supported by the previous studies that used explainable AI (XAI) mechanisms (Hossain et al., 2025; Alam et al., 2023; Altomi et al., 2025). This visual interpretability does not only improve clinical trust, but also allows model debugging and refinement, especially in sensitive healthcare environments.

Robustness, Generalization, and Domain Adaptation

One major problem of AI-based ASD detection is the need to provide strong generalization to a wide variety of populations and imaging conditions. The multi-task ViT shows a low variance between repeated training processes (standard deviation of under 0.02 in terms of accuracy) which is an indication that the convergence is consistent and that the performance can be trusted in the presence of various random initializations. This stability is consistent with the Alam et al. (2025), who stated that ensemble and domain-adaptive architectures like ASD-UANet can alleviate the performance loss caused by domain changes. Moreover, lightweight data augmentation coupled with transfer learning based on DeiT-base that has been trained on ImageNet can be used to extract the features effectively, even though the dataset size (2,936 images) is moderate. The methods overcome small-sample constraints mentioned by Dass et al. (2024) and Ibadi and Lakizadeh (2025), and they allow

obtaining reproducibility and generalization without large-scale, data-consuming data collection. However, cross-population validation is a field yet to be attempted by researchers in the future as there is potential demographic and environmental error that might affect the performance of the models in real-world applications.

Clinical Implications and Ethical Considerations

The clinical utility of AI-based ASD screening is based on not only the predictive capacity but also the interpretability and ethical soundness. Explainability through attention gives clinicians clear information, which could be used to make informed decisions in addition to other conventional methods of diagnosing a patient. Additionally, the use of ethically obtained and anonymized data will guarantee the compliance with the privacy and consent policies, which were identified by Atlam et al. (2025) as an issue when dealing with retracted research using unconfirmed online images. The proposed framework is an opportunity to become a viable addition to the traditional ASD evaluation pipelines as it combines high accuracy, interpretability, and data-conscious approaches. The early diagnosis allows timely interventions, which may positively affect the growth of children with ASD and their learning achievements, as several studies have proven (Dass et al., 2024; Farhat et al., 2024).

Limitations and Future Directions

Although the current study has showed good performance, there are a number of limitations associated with it. The size of the dataset is adequate to run the transformer-based models with transfer learning, but is still moderate, and thus cannot be used to assess the models with different demographic groups. The use of facial images in the single-modes leaves out behavioral and physiological cues, including eye-tracking or speech, which have been found to complement each other in predictive accuracy (Kasri et al., 2025; Elbattah et al., 2022). The future work is to consider multi-modal, longitudinal datasets to understand the developmental patterns and introduce lightweight and edge-computing solutions to deploy them in real-time, which

Rezaee et al. (2025) proposed. Also, it is possible to include domain adaptation and ensemble solutions to augment resistance to population heterogeneity and real-world imaging variability, as shown by Alam et al. (2025).

Conclusion

This research is an end-to-end Multi-Task Vision Transformer (ViT) architecture that is used in the early and accurate detection of Autism Spectrum Disorder (ASD) in the face image. The proposed system can combine both structural and subtle emotional stimuli due to the introduction of an auxiliary emotion recognition task into the main ASD classification objective, which leads to better discriminative feature representations. When compared to traditional CNN designs (VGG16, ResNet18) and a single-task ViT, it is possible to note that the multi-task design is more accurate (89.0%), more precise (94.0%), more recall (83.3%), and more F1-score (88.3%), which underscores its capability to balance sensitivity and specificity which is a critical consideration in clinical and education screening setting. The pretrained DeiT-base transformer backbone is useful to extract global and local facial features in an efficient manner, whereas lightweight data augmentation eliminates overfitting in spite of moderate data size. Attention-based explainability also increases clinical trust and offers visual clues on the parts of the face that have the most influence in model predictions, e.g. the eyes, mouth, and eyebrows. This interpretability (along with high predictive accuracy) highlights the prospect of the transformer-based multi-task learning structures in supporting automated, non-invasive screening of ASD. The limitation of the study in the application of the sole modality (static facial images) and moderate data set restricts the ability to generalize the results to a variety of populations and settings. Future research ought to address the aspects of multi-modal integration, domain adaptation, and real-world deployment conditions to enhance its resistance, scale, and inclusiveness. Moreover, longitudinal research with developmental pathways might give more information about early ASD predictors. On the whole, the suggested Multi-Task ViT framework is an appropriate boost to

the field of AI-assisted ASD detection, as it proves that the synergies of transformer architectures, multi-task learning, and explainable AI can produce accurate, interpretable, and clinically significant results. This will provide a solid basis on the future research towards scalable, ethical and trustworthy AI-driven neurodevelopmental disorder screening.

Declarations:

The authors declare that there is no conflict of interest regarding the publication of this paper.

Data Availability:

The dataset used in this study is publicly available from Kaggle: *Autistic Children Facial Dataset* (Fazlayrabbi3, 2022.).

Ethical Approval Statement:

This study did not involve direct interaction with human participants, animals, or clinical trials. All experiments were conducted on publicly available facial image data, and therefore ethical approval was not required.

References

- Farhat, T., Akram, S., AlSagri, H. S., Ali, Z., Ahmad, A., & Jaffar, A. (2024). Facial Image-Based Autism Detection: A Comparative Study of Deep Neural Network Classifiers. *Computers, Materials and Continua*, 78(1), 105–126. <https://doi.org/10.32604/cmc.2023.045022>
- Altomi, Z. A., Alsakar, Y. M., El-Gayar, M. M., Elmogy, M., & Fouda, Y. M. (2025). Autism Spectrum Disorder Diagnosis Based on Attentional Feature Fusion Using NasNetMobile and DeiT Networks. *Electronics (Switzerland)*, 14(9), 1–25. <https://doi.org/10.3390/electronics14091822>

- Alam, M. S., Rashid, M. M., Jazlan, A., Alahi, M. E. E., Kchaou, M., & Alharthi, K. A. B. (2025). Robust Autism Spectrum Disorder Screening Based on Facial Images (For Disability Diagnosis): A Domain-Adaptive Deep Ensemble Approach. *Diagnostics*, 15(13), 1–32. <https://doi.org/10.3390/diagnostics15131601>
- Hossain, S. S., Al-Islam, F., Islam, M. R., Rahman, S., & Parvej, M. S. (2025). Autism Spectrum Disorder Identification from Facial Images using Fine Tuned Pre-trained Deep Learning Models and Explainable AI Techniques. *Semarak International Journal of Applied Psychology*, 5(1), 29–53. <https://doi.org/10.37934/sijap.5.1.2953b>
- Atlam, E. S., Aljuhani, K. O., Gad, I., Abdelrahim, E. M., Atwa, A. E. M., & Ahmed, A. (2025). Automated identification of autism spectrum disorder from facial images using explainable deep learning models. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-11847-5>
- Rezaee, K., Javanmardi, T. S., & Heydari, A. M. (2025). A deep learning framework for classifying autism spectrum disorder from children's facial images using a multi-scale ViT architecture and edge computing. *Exploration of Medicine*, 6. <https://doi.org/10.37349/emed.2025.1001376>
- Elbattah, M., Guérin, J.-L., Carette, R., Cilia, F., & Dequen, G. (2022). Vision-based Approach for Autism Diagnosis using Transfer Learning and Eye-tracking. *Semanticscholar*, 5(Biostec), 256–263. <https://doi.org/10.5220/0010975500003123>

- Dass, J. M., Sirisha, K., Hema, B., Girisha, A., Hareesh, K., & Jyothishkumar, P. (2024). Autism Spectrum Disorder Detection using Face Features Based on Deep Neural Network. *Proceeding of 2024 International Conference on Communication, Computing and Energy Efficient Technologies, I3CEET 2024*, 1008–1012.
<https://doi.org/10.1109/I3CEET6172.2024.10994079>
- Awaji, B., Senan, E. M., Olayah, F., Alshari, E. A., Alsulami, M., Abosaq, H. A., Alqahtani, J., & Janrao, P. (2023). Hybrid Techniques of Facial Feature Image Analysis for Early Detection of Autism Spectrum Disorder Based on Combined CNN Features. *Diagnostics*, 13(18).
<https://doi.org/10.3390/diagnostics13182948>
- Ibadi, H., & Lakizadeh, A. (2025). ASDvit: Enhancing autism spectrum disorder classification using vision transformer models based on static features of facial images. *Intelligence-Based Medicine*, 11(August 2024), 100226.
<https://doi.org/10.1016/j.ibmed.2025.100226>
- Kasri, W., Himeur, Y., Copiaco, A., Mansoor, W., Albanna, A., & Eapen, V. (2025). Hybrid Vision Transformer-Mamba Framework for Autism Diagnosis via Eye-Tracking Analysis. *International Conference on Communication, Computing, Networking, and Control in Cyber-Physical Systems, CCNCPS 2025, DI*, 343–348.
<https://doi.org/10.1109/CCNCPS66785.2025.11135843>
- Rakotomanana, H., & Rouhafzay, G. (2025). A Scoping Review of AI-Based Approaches for Detecting Autism Traits Using Voice and Behavioral Data. *Bioengineering*, 12(11), 1136.
<https://doi.org/10.3390/bioengineering12111136>
- Agrawal, R., & Agrawal, R. (2025). Explainable AI in early autism detection: a literature review of interpretable machine learning approaches. *Discover Mental Health*, 5(1).
<https://doi.org/10.1007/s44192-025-00232-3>
- 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. (2023). Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI. *Technologies*, 11(5), 1–27.
<https://doi.org/10.3390/technologies11050115>
- Shabbir, A. B. (2025). SHAP-Weighted Hybrid Ensemble for Interpretable Stroke Risk Prediction Using Tabular Clinical Data. *Robotics and Artificial Intelligence Review*, 1–19.
- ButAutistic Children Facial Dataset. The <https://www.kaggle.com/datasets/fazlayrabbi3/autistic-children-facial-dataset>