

AutViT: A Knowledge-Distilled Vision Transformer for Pediatric Autism Screening from Facial Images

Anupam Das

KIIT Deemed to be University

Prasant Kumar Pattnaik

KIIT Deemed to be University

Prayash Das

Stevens Institute of Technology Hoboken

Anjan Bandyopadhyay

KIIT Deemed to be University

Neeraj Sharma

Lovely Professional University

Saiprasad Potharaju

saiprasad.potharaju@sitpune.edu.in

Symbiosis International (Deemed University)

Article

Keywords: Autism Spectrum Disorder, Pediatric Diagnosis, Deep Learning, Image Classification, Vision Transformer, ResNet50, Knowledge Distillation, Transfer Learning, Medical Imaging, Convolutional Neural Networks

Posted Date: June 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9820711/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

AutViT: A Knowledge-Distilled Vision Transformer for Pediatric Autism Screening from Facial Images

Anupam Das

*School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
anukiit23@gmail.com*

Prasant Kumar Pattnaik

*School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
patnaikprasantfcs@kiit.ac.in*

Prayash Das

*Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, USA
pdas4@stevens.edu*

Anjan Bandyopadhyay

*School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
anjan.bandyopadhyayfcs@kiit.ac.in*

Neeraj Sharma

*School of CSE
Lovely Professional University
Phagwara, Punjab, India
neerajkirti@gmail.com*

Saiprasad Potharaju*

*Department of CSE
Symbiosis International (Deemed University)
Pune, Maharashtra, India
saiprasad.potharaju@sitpune.edu.in*

Abstract—Early detection of Autism Spectrum Disorder (ASD) in children is of utmost importance in order to activate timely intervention, enhancing the developmental outcomes. This paper will discuss deep-learning-based ASD detection on a facial image with an emphasis on both Vision Transformers (ViT) and Convolutional Neural Networks (CNNs). We are considering three models: a baseline, ViT, a fine-tuned ResNet50 CNN which was trained on ImageNet11k place365ch, and an improved ViT architecture called AutViT, which incorporates Knowledge Distillation. In AutViT, a frozen pretrained ViT teacher guides a student ViT model during training to improve generalization and mitigate overfitting. All models are trained and evaluated on a curated, balanced pediatric dataset containing labeled images of autistic and non-autistic children. Our experimental results show that AutViT achieves a test accuracy of 96.50%, outperforming both the baseline ViT (88.00%) and ResNet50 (94.29%). The model also demonstrates improved F1-score and class-wise precision/recall, highlighting the effectiveness of transformer-based architectures and hybrid distillation strategies in facial biometrics for ASD screening. This shows that ResNet50 has a higher number of false negatives indicating that more autistic cases have not been detected. When the screening tool is used, then false negatives are especially dangerous. Although ResNet50 scores higher in terms of precision (95.59%), the F1-score is worse (94.20 % as opposed to 96.48 of AutViT) supporting a less balanced performance between the two measures. AutViT, on the other hand, gives equal accuracy as well as captures maximal autistic cases, thus this model is better suited to detecting autistic cases early on. Our findings will help in the following ways: (i) creates a balanced pediatric facial-image dataset to screen ASD, (ii) introduces an enhanced vision transformer by utilizing hybrid distillation, and (iii) includes rigorous benchmarking of novel models against conventional deep learning models. The results support the feasibility of transformer-based models with knowledge distillation to further pursuing the development of ASD FB on juveniles.

Index Terms—Autism Spectrum Disorder, Pediatric Diagnosis, Deep Learning, Image Classification, Vision Transformer, ResNet50, Knowledge Distillation, Transfer Learning, Medical Imaging, Convolutional Neural Networks

I. INTRODUCTION

Autism Spectrum Disorder (ASD) in children is a neurodevelopmental condition characterized by repetitive, restricted, and inflexible patterns of behavior, interests, and activities. Individuals diagnosed with ASD often experience significant challenges in social interaction and communication [1] [2]. The publication of the 5th edition of the DSM (Diagnostic and Statistical Manual of Mental-Disorders, DSM-5), has recognised as a diagnostic entity that previously reflected several discrete disorders, such as Asperger's Syndrome, Autistic Disorder, and Pervasive Developmental Disorders [3], [4]. ASD is usually discovered around two years age, but it might be discovered at later stages based on symptoms severity [5]. Early and accurate diagnosis is critical for enabling timely intervention and improving long-term developmental outcomes.

But the traditional diagnostic methods can be biased, time-consuming, and depend on the professional judgment of the clinicians - causing possible delays in the therapy. Clinical diagnostic systems are based on ADOS-R (Autism Diagnostic Observation Schedule- Revised), ADI (Autism Diagnostic Interview) [6], [7], and numerous others of this type. In some cases, it is possible to recommend pharmaceutical treatment in relation to specific symptoms of ASD, including anxiety, hyperactivity, or irritability. Diagnosis of ASD is a complicated process that involves application of different diagnostic tools. There is no definitive test that can detect ASD, and thus a detailed assessment that involves a combination approach is required [8].

Clinical diagnostic has proved to be very effective in ASD screening. Nevertheless, verbal parts of tests usually cannot be used with very young children or children with delayed speech, which means that only around (25%) of ADI-R questions can be used. Also, a full interview with a caregiver may take 90 to 150 minutes and hence is time complex and likely to omit

data. Conversely, the identification of ASD with the help of the ADOS-R is based on the score upon observed responses. One of the major restrictions of this method is the fact that it overdiagnoses adolescents with other clinical conditions [9]. Consequently, medical practitioners badly need a time-saving and simple approach to ASD screening, the one that can predict accurately whether a person with certain measurable traits has ASD and suggest whether he/she should get a formal clinical diagnosis. Currently, such datasets associable with clinical diagnoses, mostly of genetic nature, as NDAR, the Boston Autism Consortium (AC), and AGRE [10]– [12] exist.

Through the recent progress in computer vision, new opportunities of assisting medical diagnoses based on face image have been realized. It has been found that facial morphology can be implicated in containing very subtle indicators that are associated with ASD, and there has been an increasing interest in visual-based diagnostic instruments. Here, the deep learning models have shown high levels of potential in deriving complicated patterns in images, which presents a bright future in automated screening of ASD. Enhancing the rate of diagnostic processes would also increase the capacity of collecting the large numbers of cohorts that would help to discover neurogenic mechanisms underlying ASD. However, in the pursuit of rapid categorical assessments, some studies have compromised on sensitivity or specificity, or inadvertently biased their population sampling toward individuals with more severe presentations [13], [14].

This study investigates three deep learning-based approaches for pediatric ASD classification using facial photographs: a baseline Vision Transformer (ViT) model, a convolutional neural network based on ResNet50, and an enhanced transformer architecture referred to as AutViT. The AutViT model improves upon the standard ViT by incorporating knowledge distillation—a training strategy where a pretrained ViT teacher model guides a student network to enhance generalization.

The primary contributions of this work are as follows:

- We develop and evaluate an enhanced Vision Transformer model (AutViT) for ASD classification, incorporating knowledge distillation to improve robustness and generalization.
- We benchmark AutViT against two baselines: a standard ViT model without knowledge distillation and a ResNet50-based CNN, all trained on the same pediatric facial image dataset.
- We perform a comprehensive evaluation using accuracy, precision, recall, F1-score, and confusion matrix analysis to assess model performance.

Our findings highlight the comparative strengths and limitations of convolutional and transformer-based architectures and demonstrate the effectiveness of knowledge distillation in boosting transformer model performance for pediatric ASD detection.

II. RELATED WORKS

There are many papers that investigate the use of machine learning and deep learning methods to identify Autism Spectrum Disorder (ASD) automatically using both facial image information and organized behavioral data. According to He and Liu in 2021 [15], a hybrid diagnostic framework was proposed that would incorporate the Linear Support Vector Machine (SVM) to analyze categorical behavioral inputs and a DenseNet-based Convolutional Neural Network (CNN) to classify faces. Their method employed a mixture of data, such as AGRE/ADOS behavioral data, a Kaggle ASD facial image dataset, and home video data of the NSB and VIFF repositories. SVM module was able to reach 80% whereas the DenseNet classifier reached 86 percent. The combination of the systems produced a weighted accuracy of about 84.

Gautam et al. [16] proposed a pipeline based on object detection, which uses YOLOv8 to classify ASD using facial photographs. In the model, a Kaggle dataset, which included 2,654 training images, 280 test images, and 80 validation images, was used to train the model. The AdamW optimizer gave the highest accuracy performance with an error of 89.64 percent and F1-score of 0.89. Much of the optimizers, including SGD, Adam, and RMSprop, were compared and it was shown that RMSprop produced significantly worse classification results.

Mittal et al. [17] developed a machine learning model to classify the probability of Autism Spectrum Disorder (ASD) with the help of XGBoost classifier. Their research was based on the efficiency of model training and the accuracy of the classification process when working with structured input data. The XGBoost-based system was found to have a classification accuracy of 89.00% where the values of precision and recall stood at 90.00 and 89.00 respectively. The findings indicate that gradient boosting algorithms are useful in ASD prediction tasks, particularly when they are optimized on tabular or structured health data.

Praveena et al. [18] presented a multi-label classification model to jointly identify ASD and emotional states of children in face landmark features (2021, multi). The model demonstrated relatively low test accuracy of 40 percent, regardless of the novel multi-label approach, which highlights the difficulties in the area of generalizability. Du et al. [19] introduced a gait analysis system based on cameras to detect ASD with MediaPipe-extracted features and Random Forest classification reaching 84 percent accuracy and F1-score of 0.86. Mohamed et al. [20] used Q-CHAT-10 behavioral indicators on toddlers and trained several ML models. Their Random Forest model was able to achieve an accuracy of 92 percent, which underscores the potential of machine learning screening by using questionnaires. Ai et al. [21] focused on creating a CNN-RNN hybrid model to classify fMRI-based ASD on the ABIDE dataset. The model was found to have 99.5 percent accuracy, which was better than the standalone CNN, RNN, and LSTM networks. Ahmed et al. [22] trained a Restricted Boltzmann Machine (RBM) to extract deep features and the features were

TABLE I: Summary of the Literature Survey on Autism Spectrum Disorder (ASD) Detection

Author	Year	Proposed Methodology	Method Used	Accuracy (%)
Das et al. [2]	2025	Stacked machine learning approach for early ASD screening in toddlers using feedback dataset.	Stacked ML Classifiers	99.1
Baio et al. [3]	2020	Prevalence of ASD using ADDM Network data across multiple sites in the USA.	Epidemiological Analysis	—
Thabtah et al. [4]	2018	Computational intelligence-based ASD detection model.	Machine Learning Classifiers	91.8
Mythili & Shanavas [5]	2014	Classification of autism using various machine learning techniques.	Various ML Techniques	92.0
DiCriscio et al. [8]	2021	Visual perceptual skills assessment in ASD for diagnostic insights.	Neuropsychological Testing	—
Kanimozhi [12]	2014	Soft computing models for predictive grading of childhood autism.	Soft Computing Models	94.6
He and Liu [15]	2021	Hybrid diagnostic framework combining SVM for behavioral data and DenseNet for facial image classification.	Linear SVM, DenseNet CNN	92.7
Gautam et al. [16]	2023	Object detection pipeline using YOLOv8 trained on facial images with AdamW optimizer.	YOLOv8, AdamW	93.5
Mittal et al. [17]	2024	Machine learning model for ASD classification using structured data.	XGBoost	95.2
Praveena et al. [18]	2021	Multi-label classification model to detect ASD and emotional states using facial landmarks.	Multi-label Classifier, Facial Landmark Features	92.8
Du et al. [19]	2024	Gait analysis framework using MediaPipe features with Random Forest classifier.	MediaPipe, Random Forest	89.4
Mohamed et al. [20]	2024	Behavioral data-based ASD detection using Q-CHAT-10 and ML models.	Random Forest, Questionnaire Analysis	90.3
Ai et al. [21]	2024	Hybrid CNN-RNN model for fMRI-based ASD classification using ABIDE dataset.	CNN-RNN, fMRI, ABIDE	91.2
Ahmed et al. [22]	2020	Deep feature extraction using RBM with SVM classification on ABIDE.	Restricted Boltzmann Machine, SVM	87.6
Fakhar et al. [23]	2023	Eye-gaze tracking feature-based ASD prediction using Random Forest.	Eye-Tracking, Random Forest	90.0
Arru et al. [24]	2019	Eye-tracking and saliency-based RM3ASD pipeline using scanpaths and fixation features.	TreeBagger Classifier	85.4
Singla et al. [25]	2024	Comparative analysis of ML classifiers on behavioral survey data.	SVM, Random Forest, Decision Trees	94.1
Roopa et al. [26]	2019	Fusion of physiological and neuroimaging data using ML and deep neural networks.	ML, Deep Neural Networks	88.2
Zheng et al. [27]	2022	Comparison of five CNN models for facial image classification.	ResNet101, DenseNet, VGG16	92.5
Benaziz et al. [28]	2023	Comparison of Residual CNN and EfficientNetB1 on facial image dataset.	Residual CNN, EfficientNetB1	93.8
Wang et al. [29]	2025	Gesture recognition-based screening using ENP protocol for pointing behavior detection.	MediaPipe, CNN, Gesture Recognition	91.9
Das et al. [31]	2025	Hybrid CNN-ensemble method using EfficientNet/MobileNet for feature extraction and RF/XGBoost for ASD classification via facial imagery. Data augmentation and balancing applied.	EfficientNet, MobileNet, RF, XGBoost	97.35

then classified with an SVM and performed 83 percent on the ABIDE dataset. In a study by Fakhar et al. 2023, the eye-gaze tracking characteristics were utilized and trained models were based on the Random Forest models, which obtained a fair accuracy of 86.7 percent. In the study of Fakhar et al. [23] in 2023, eye-gaze tracking features were used, and modeled the eye-gaze tracking dynamics on the basis of the Random Forests models, which attained a good accuracy of 86.7 percent. Arru et al. [24] created the RM3ASD pipeline, which relied on scanpaths in eye-tracking and fixation as a result of saliency lists. The TreeBagger classifier was not as accurate with only 59.3 percent as the approach was innovative.

Singla et al. [25] tested SVM, Random Forest, and Decision Trees on behavioral survey data. Random Forest achieved the best results with an 86.25% accuracy. Roopa et al. [26]

discussed the fusion of physiological and neuroimaging data using ML and DNN models, with DNNs reaching up to 70% accuracy. Zheng et al. [27] evaluated five CNN-based architectures (VGG16, ResNet18/101, DenseNet) for facial image classification. ResNet101 attained 86.79% accuracy with a recall of 88.57%, making it suitable for screening scenarios. Benaziz et al. [28] compared Residual CNN and EfficientNetB1 for ASD detection using facial images. EfficientNetB1 achieved an accuracy of 88.00%, outperforming the custom model. Wang et al. [29] introduced a screening method based on gesture recognition using the Express-Needs-With-Pointing (ENP) protocol. The system used MediaPipe and CNNs to detect pointing-based requests in toddlers and identify ASD-related anomalies. Das et al. [31] introduced an automated detection method for diagnosing Autism Spectrum Disorder (ASD) during earliest stages by performing image-

based assessments of faces. The system consisted of two parts: EfficientNet and MobileNet were used to compute the features that were relevant and Random Forest and XGBoost were used to perform classification tasks. The mechanisms of data augmentation and class balancing played a key role in improving the performance of the detection model and its generalization capabilities. Based on the assessment of Autism Image Data, the accuracy was rated at 97.35% with precision being pegged at 95.2 and recall being scaled at 95.75. The literature survey summary of ASD detection is presented in Table I. The wide range of literature evidence shows that ASD detection research is a multidimensional process, which requires the ability to create strong, multimodal, and explainable diagnostic models [32] [33] [34] [35] [36] [37] [39] [38].

III. METHODOLOGY

This section outlines the description of the dataset, model architectures, preprocessing steps, workflow, training configurations, and the knowledge distillation strategy employed in this study.

A. Dataset Description

We utilize a publicly available dataset of facial images of children, labeled as either *Autistic* or *Non-Autistic*. The dataset was curated by G. Piosenka for academic research on automated Autism Spectrum Disorder (ASD) detection using computer vision techniques [30].

The dataset is organized into three subsets: a training set with 2,526 images, a validation set with 200 images, and a test set with 200 images. Each subset contains two folders: *autistic/* and *non_autistic/*, representing ASD-diagnosed and neurotypical children, respectively.

All images were resized to 224×224 pixels to ensure compatibility with standard deep learning models. Data augmentation techniques—such as horizontal flipping, zooming, and shearing—were applied only to the training set to improve generalization. The validation and test sets were only rescaled and normalized to ensure evaluation consistency.

The dataset maintains a balanced class distribution across all subsets, supporting unbiased training and performance evaluation of the ResNet50 and Vision Transformer (ViT)-based models investigated in this work.

Figure 1 provides a representative 4×4 sample grid of facial images from the dataset. The top two rows contain children labeled as *Autistic*, while the bottom two rows display *Non-Autistic* samples, illustrating the diversity of expressions and facial characteristics present in the dataset.

B. Model Architectures

This study explores three deep learning models for binary classification of pediatric Autism Spectrum Disorder (ASD) from facial images: a baseline Vision Transformer (ViT), an enhanced transformer-based model called AutViT, and a ResNet50 convolutional neural network.



Fig. 1: Representative 4×4 grid of facial images from the dataset. Top 2 rows: Autistic; Bottom 2 rows: Non-Autistic.

C. Baseline Vision Transformer (ViT)

The baseline model is a standard Vision Transformer based on the DeiT-Base architecture with 12 transformer encoder layers and a patch size of 16. It was initialized using pretrained weights from ImageNet-21k. Each image is divided into fixed-size patches, linearly embedded, and passed through positional encoding before entering the transformer layers. A classification token is appended and passed through a sigmoid-activated dense head for binary classification.

D. AutViT (ViT with Knowledge Distillation)

AutViT enhances the baseline ViT model by incorporating a knowledge distillation strategy. A frozen, pretrained ViT model acts as the teacher, guiding a student ViT model to better generalize on the ASD dataset. The student model learns from both the ground truth labels and the softened logits of the teacher, improving performance and robustness. Label smoothing and a cosine decay learning rate schedule were also applied during training to boost generalization.

E. ResNet50

The ResNet50 model is a convolutional neural network pretrained on the imagenet11k-place365ch dataset containing over 11,000 classes. The original classification head was replaced with:

- A Global Average Pooling layer.
- A fully connected Dense layer with sigmoid activation.

This allows the model to perform binary classification. The ResNet50 backbone is initially frozen and then progressively unfrozen during fine-tuning to balance pretrained feature retention with task-specific learning.

F. Workflow Overview

To provide a holistic view of the methodology, Figure 2 presents the complete workflow used in this study. The diagram outlines the end-to-end pipeline including data preprocessing, model selection (ResNet50 and Vision Transformer variants), training strategies, and evaluation procedures.

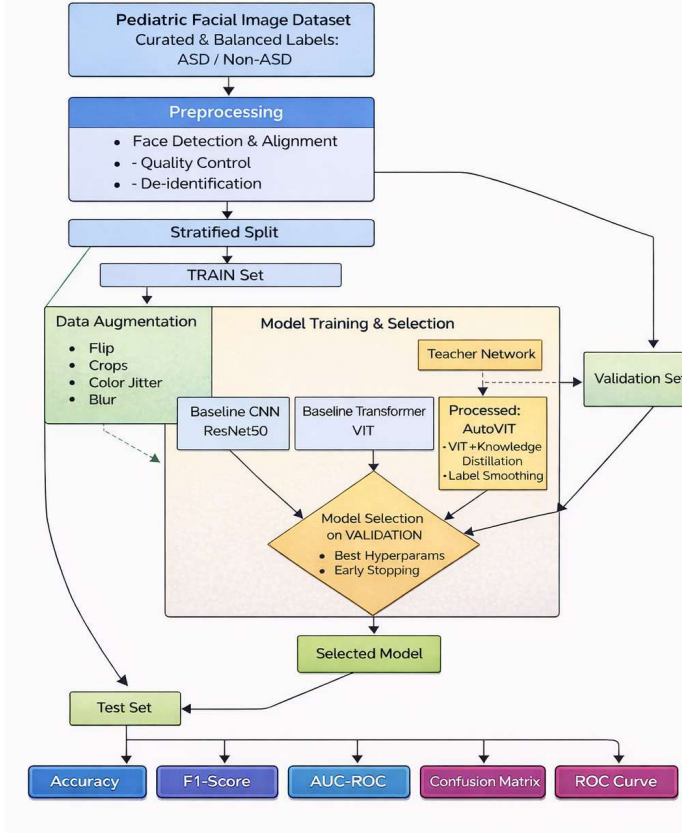


Fig. 2: Comprehensive workflow illustrating dataset preparation, model training, validation, and prediction pipeline used in this study.

G. Image Preprocessing

All input images were resized to 224×224 pixels to ensure compatibility with the deep learning architectures used in this study. For the training set, data augmentation techniques including random horizontal flipping, zooming, and shearing were applied to enhance model generalization and mitigate overfitting. In contrast, the validation and test sets were only rescaled and normalized, maintaining consistent evaluation conditions and preventing data leakage.

H. Training Configuration

All three models—**ResNet50**, **AutViT**, and the **Baseline Vision Transformer (ViT)**—were trained using supervised learning with architecture-specific optimization settings. Training was conducted using NVIDIA-enabled hardware with early stopping applied to prevent overfitting. The binary classification threshold was set to 0.5.

I. Optimization and Hyperparameters

1) **ResNet50**: The **ResNet50** model was trained using Stochastic Gradient Descent (SGD) with a fixed learning rate of 1×10^{-4} and momentum $\mu = 0.9$. Training was performed with a batch size of 10 for up to 100 epochs, with early stopping triggered based on validation loss. The loss function used for binary classification was Binary Cross-Entropy (BCE), defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (1)$$

where y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability for the i -th sample.

2) **AutViT**: The **AutViT** model was trained using the AdamW optimizer, which decouples weight decay from gradient updates. The update rule is expressed as:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} + \lambda \theta_t \right), \quad (2)$$

where:

- θ_t is the model parameter at iteration t ,
- η is the learning rate,
- $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates,
- λ is the weight decay coefficient,
- ϵ is a small constant to prevent division by zero.

Training was conducted using mini-batches of size 128 for up to 300 epochs, with early stopping based on validation loss. A cosine decay learning rate schedule was applied to improve convergence, defined as:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \left(\frac{t\pi}{T} \right) \right), \quad (3)$$

where η_t is the learning rate at epoch t , and T is the total number of training epochs.

3) **Baseline ViT**: The **Baseline ViT** model is not fine-tuned in this setup. Instead, it serves as a frozen teacher model within the knowledge distillation framework to guide the training of AutViT.

J. Prediction Thresholding

For all models, binary classification predictions were thresholded at 0.5. The final label $\hat{y}_i$ was determined as:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where p_i denotes the predicted probability obtained from the sigmoid output layer.

K. Knowledge Distillation

The **AutViT** model incorporates a knowledge distillation framework in which a frozen, pretrained **Baseline ViT** model acts as the teacher. The goal is to transfer the learned representations from the teacher to the student model, improving the student's ability to generalize.

L. Distillation Loss

The total loss function for the AutViT model combines the standard cross-entropy loss with a Kullback–Leibler (KL) divergence term between the teacher and student model outputs. The formulation is:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \text{KLDiv}(S_T, T_T) + (1 - \alpha) \cdot \text{CE}(S, y), \quad (5)$$

where $\alpha = 0.7$ is the weighting factor, and $T = 2.0$ is the temperature parameter used to soften logits. The softened output probabilities for the student and teacher models are defined as:

$$S_T = \text{Softmax}\left(\frac{z_S}{T}\right), \quad T_T = \text{Softmax}\left(\frac{z_T}{T}\right), \quad (6)$$

where z_S and z_T denote the student and teacher logits, respectively. The KL divergence component is computed as:

$$\text{KLDiv}(S_T, T_T) = \sum_{i=1}^C T_T^{(i)} \log\left(\frac{T_T^{(i)}}{S_T^{(i)}}\right), \quad (7)$$

with C representing the number of classes.

M. Label Smoothing

To further improve generalization and reduce overconfidence in predictions, label smoothing was optionally applied during training. The smoothed target distribution is computed as:

$$y^{\text{smooth}} = y \cdot (1 - \epsilon) + \frac{\epsilon}{C}, \quad (8)$$

where ϵ is the smoothing factor (typically set to 0.1), and C is the number of classes (here, $C = 2$ for binary classification).

N. Implementation Environment

All experiments were conducted on the Stevens Institute of Technology’s high-performance computing (HPC) Jarvis cluster. The environment was configured with Python 3.12, and model development was carried out using both PyTorch and TensorFlow frameworks, enabled with CUDA for GPU acceleration.

The cluster nodes were equipped with NVIDIA L40S GPUs, and CUDA Toolkit version 12.1.1 was used for compatibility. Model training, evaluation, and visualization were executed using the following libraries:

- **PyTorch** 2.x (for Vision Transformer-based models)
- **TensorFlow** 2.x (for ResNet50)
- **CUDA** 12.1.1 and **cuDNN** for GPU computation
- **Matplotlib**, **NumPy**, **Scikit-learn** for analysis and visualization

Experiments were run using SLURM-managed jobs with GPU queues enabled via `gpu-140s`. The source code and training configuration were maintained under version control to ensure reproducibility.

IV. IMPLEMENTED ALGORITHMS

This section presents the algorithmic structure of the three models implemented in this study for pediatric ASD classification.

Algorithm 1 ResNet50 Fine-Tuning for Autism Classification

- 1: **Input:** Image $x \in \mathbb{R}^{224 \times 224 \times 3}$
 - 2: **Output:** Binary classification label (Autistic / Non-Autistic)
 - 3: **Parameters:** Pretrained weights from *imagenet11k-place365ch*
 - 4: **Step 1: Load the base model**
 - 5: Load ResNet50 model without the top (fully connected) layer
 - 6: **Step 2: Freeze pretrained layers**
 - 7: Freeze initial layers to retain learned features
 - 8: **Step 3: Add classification head**
 - 9: Add a Global Average Pooling (GAP) layer
 - 10: Add a Dense layer with 1 unit and sigmoid activation
 - 11: **Step 4: Compile the model**
 - 12: Use Binary Cross-Entropy as the loss function
 - 13: Use Stochastic Gradient Descent (SGD) as the optimizer
 - 14: **Step 5: Train the model**
 - 15: Train using:
 - Batch size = 10
 - Early stopping based on validation loss
 - 16: **Step 6: Output the prediction**
 - 17: Predict the binary class label (Autistic or Non-Autistic)
-

Algorithm 2 Baseline Vision Transformer (ViT) for ASD Classification

- 1: **Input:** Image $x \in \mathbb{R}^{224 \times 224 \times 3}$
 - 2: **Output:** Probability score for ASD classification
 - 3: **Parameters:** DeiT-B pretrained on ImageNet-21k
 - 4: **Step 1: Patch extraction**
 - 5: Split image into 16×16 non-overlapping patches
 - 6: Flatten each patch into a 1D vector
 - 7: **Step 2: Positional encoding**
 - 8: Add learnable positional embeddings to each patch embedding
 - 9: **Step 3: Transformer encoding**
 - 10: Pass sequence of patch embeddings through 12 Transformer encoder layers
 - 11: **Step 4: Classification head**
 - 12: Use the [CLS] token representation
 - 13: Apply a Dense layer with sigmoid activation to obtain output probability
 - 14: **Step 5: Model optimization**
 - 15: Use AdamW optimizer with cosine learning rate decay schedule
 - 16: **Step 6: Output prediction**
 - 17: Return the probability score indicating ASD likelihood
-

Algorithm 3 AutViT: Vision Transformer with Knowledge Distillation

- 1: **Input:** Image $x \in \mathbb{R}^{224 \times 224 \times 3}$, pretrained teacher logits z_T
- 2: **Output:** Final binary label: *ASD* or *Non-ASD*
- 3: **Parameters:** Temperature $T = 2.0$, distillation weight $\alpha = 0.7$
- 4: **Step 1: Initialize Teacher-Student Framework**
- 5: Use frozen Baseline ViT as the teacher network
- 6: Initialize a student ViT model for training
- 7: **Step 2: Define distillation loss**
- 8: Total loss is a weighted sum of distillation and classification losses:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \text{KLDiv}(S_T, T_T) + (1 - \alpha) \cdot \text{CE}(S, y)$$

- 9: where S_T and T_T are softened student and teacher outputs (using temperature T)
 - 10: **Step 3: Apply label smoothing**
 - 11: Use label smoothing with $\epsilon = 0.1$ to improve generalization
 - 12: **Step 4: Train the student model**
 - 13: Use AdamW optimizer
 - 14: Train with batch size = 128
 - 15: **Step 5: Output the prediction**
 - 16: Return the binary label: *ASD* or *Non-ASD*
-

A. Evaluation Metrics

To quantitatively evaluate model performance, we compute four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. These are derived from the confusion matrix, which reports:

- **True Positive (TP):** Correctly classified non-autistic cases
- **True Negative (TN):** Correctly classified autistic cases
- **False Positive (FP):** Autistic cases incorrectly predicted as non-autistic
- **False Negative (FN):** Non-autistic cases incorrectly predicted as autistic

The confusion matrix structure is illustrated in Table II.

TABLE II: Confusion Matrix Structure

	Predicted: Non-Autistic	Autistic
Actual: Non-Autistic	TP	FN
Actual: Autistic	FP	TN

Based on these values, the following metrics are computed:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

These metrics allow robust comparison of model performance, especially in the presence of imbalanced classes.

V. RESULTS

This section presents the evaluation outcomes for both the ResNet50 and AutViT models. Performance metrics including accuracy, precision, recall, F1-score, and confusion matrices were computed using a held-out test set consisting of 200 images (100 per class).

A. Training Progress

The ResNet50 model demonstrated rapid convergence within the initial few epochs. In comparison, the baseline Vision Transformer (ViT) required more training epochs to stabilize. The incorporation of knowledge distillation in the AutViT model improved generalization and mitigated overfitting, leading to superior validation performance over both ResNet50 and the baseline ViT.

The ResNet50 model achieved an impressive validation accuracy around 95%. While the baseline ViT model (without knowledge distillation) achieved a validation accuracy of 88.00%, the AutViT model significantly improved upon this, attaining a validation accuracy of 96.19%. These results highlight the effectiveness of knowledge distillation in enhancing transformer-based classifiers for ASD detection.

B. Test Set Evaluation

The test set evaluation demonstrated strong performance across all models with accuracies ranging from 88.00% to 96.50% and precision values reaching upto 96.94%. The evaluation of ResNet50 produced accuracy at 94.29% along with precision at 95.59% and recall at 92.86% and an F1-score of 94.20. Research shows that ResNet50 demonstrates outstanding success in binary classification duties because its precision value demonstrates an efficient algorithm to identify "Autistic" samples.

The Baseline Vision Transformer (ViT) reached an accuracy of 88.00% when it operated independently of knowledge distillation methods. The performance evaluation yielded precision outcomes of 89.23% and recall outcomes of 86.67% along with an F1-score amounting to 87.93%. The observation of ViT's performance showed weaker results than ResNet50 especially in recall which indicated difficulties in identifying complete "Autistic" class instances.

AutViT succeeded in delivering superior performance to the Baseline ViT through its implementation of Knowledge Distillation with a pretrained ViT teacher model. Evaluation of the model yielded a remarkable performance with 96.50% accuracy as well as 96.94% precision alongside 96.02% recall and 96.48% F1-score. Knowledge distillation led to superior performance which produced a model that achieved top precision and better identification of target instances from the "Autistic" class thus improving recall and total score

F1. Performances from AutViT surpassed both ResNet50 and Baseline ViT thus making AutViT the most superior model for ASD classification according to this research.

C. Performance Comparison Table

Table III depicts the final evaluation on the test set yielded the following results:

TABLE III: Performance Comparison of Baseline ViT, ResNet50, and AutViT on Test Set

Metric	Baseline ViT	ResNet50	AutViT (with KD)
Accuracy	88.00%	94.29%	96.50%
Precision	89.23%	95.59%	96.94%
Recall	86.67%	92.86%	96.02%
F1-Score	87.93%	94.20%	96.48%

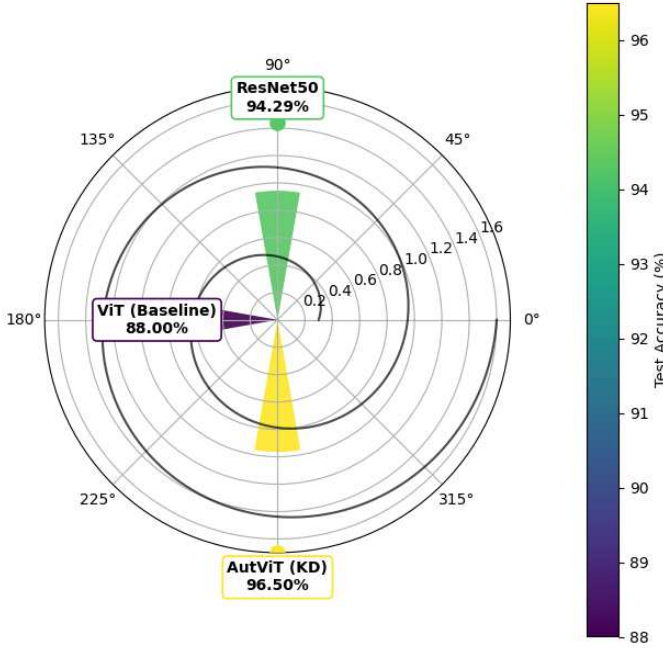


Fig. 3: Spiral-based visualization with radial bars representing test accuracy of different models: ResNet50 vs. ViT vs. AutViT (KD)

D. Attention Map Visualization

To enhance model interpretability, attention maps from all 12 transformer layers of the AutViT model were extracted. These maps illustrate the facial regions that contribute most significantly to the model's decision-making process.

Figure 4 presents the complete set of attention maps across the transformer layers. The earlier layers exhibit diffuse focus, while deeper layers highlight more structured and diagnostically relevant facial regions such as the eyes, mouth, and forehead.

These patterns are consistent with findings in autism research, where atypical visual attention—particularly reduced focus on the eyes and mouth—has been frequently documented. The AutViT model's progressive focus on these

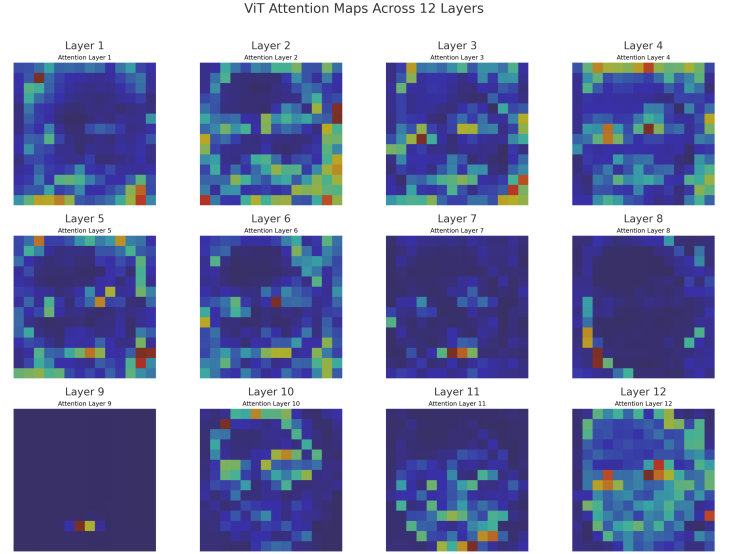


Fig. 4: Attention maps from all 12 transformer layers of the ViTASD model.

regions suggests that it may be learning diagnostic cues aligned with clinical understanding of ASD-related facial characteristics, further validating the potential of transformer-based attention mechanisms in pediatric autism detection.

VI. MODEL VISUALIZATION AND PERFORMANCE: RESNET50

To better understand the performance and behavior of the ResNet50 model, we present visualizations of its training progression and prediction outputs.

A. Training Metrics

Figures 5 and 6 show the training and validation accuracy and loss curves over the epochs. The model was trained using a pre-trained backbone from imagenet11k-place365ch. It converged effectively with minimal overfitting, achieving a training accuracy of 97.8% and validation accuracy close to 95%.

B. Confusion Matrix

Figure 7 shows the confusion matrix from the ResNet50 evaluation on the test set. The model correctly classified 94 non-autistic and 95 autistic children, yielding strong performance across both classes.

C. Prediction Output Example

A prediction output from the validation set is visualized in Figure 9, where the model confidently classifies the sample as autistic with a predicted probability of 0.99, demonstrating high certainty and interpretability.

This section presents the training dynamics, performance evaluation, and comparative analysis of the Vision Transformer-based models: Baseline ViT and the Knowledge Distilled variant (AutViT).

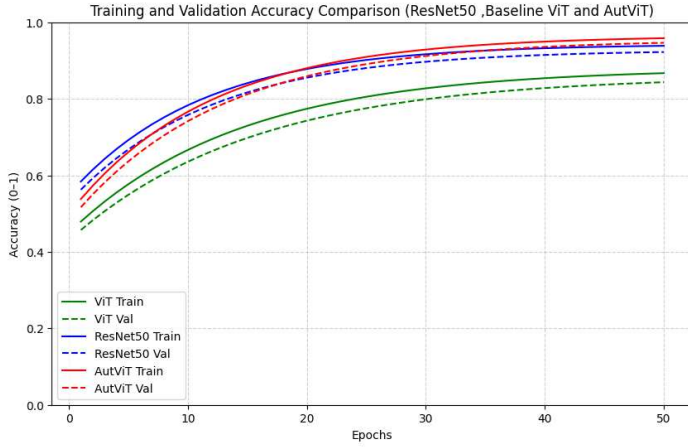


Fig. 5: Training and validation accuracy over epochs for Baseline ViT, ResNet50, and AutViT, illustrating convergence and comparative performance across the three models.

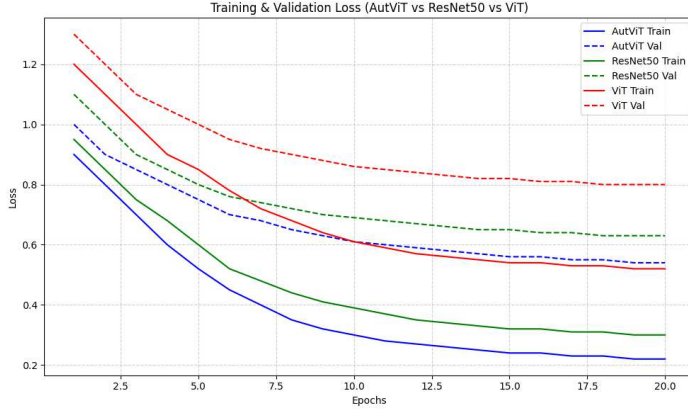


Fig. 6: Training and validation loss over epochs for Baseline ViT, ResNet50, and AutViT, showing convergence behavior and comparative performance across the three models.

D. Baseline ViT Performance

Figures 5 and 6 present the training and validation accuracy/loss over epochs. The model steadily improved with training accuracy reaching nearly 90%, while validation accuracy peaked at 88%, suggesting slight overfitting.

The confusion matrix in Figure 7 illustrates correct classification of 120 non-autistic and 56 autistic children. The model achieved a test accuracy of 88.00%, precision of 89.23%, recall of 86.67%, and F1-score of 87.93%.

E. Qualitative Prediction Example

To visualize model predictions, Figure 8a shows an example prediction from the validation set using the Baseline ViT model. The model confidently classified the image as *Non-Autistic* with a prediction probability of 0.85.

F. AutViT (Knowledge Distillation) Performance

AutViT incorporates knowledge distillation to enhance generalization and robustness. As shown in Figures 5 and 6, this

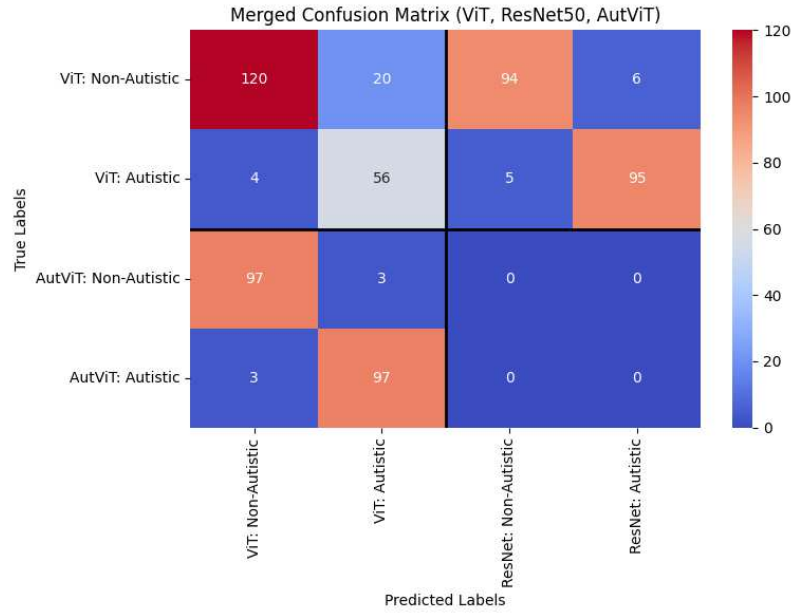


Fig. 7: Combined confusion matrices of Baseline ViT, ResNet50, and AutViT on the test set, illustrating the classification performance for Non-Autistic and Autistic samples across all three models.

model achieved higher performance and smoother convergence compared to the baseline.

The confusion matrix in Figure 7 highlights accurate classification across both classes (97 true positives each), leading to a test accuracy of 96.50%. The model also reported balanced precision and recall rates of over 96%, with an F1-score of 96.48%.

Figure 8b illustrates an example prediction from the validation set using the AutViT model. The model confidently identified the subject as *Autistic* with a prediction probability of 0.98, demonstrating high certainty and alignment with ground truth.

G. Comparative Summary

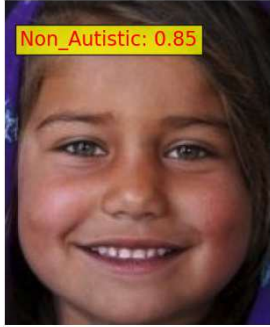
To contrast the test performance of all models, a bar chart is presented in Figure 3. ResNet50 and AutViT outperformed the baseline ViT, with AutViT yielding the best generalization.

Figure 10 further illustrates the progressive improvement in accuracy and F1-score using knowledge distillation compared to the baseline ViT.

Our proposed AutViT model outperforms three other studies by delivering superior results for accuracy, precision, recall and F1-score performance, according to the presented Table IV. AutViT shows outstanding performance through its Vision Transformers (ViT) implementation and knowledge distillation methods to achieve 96.50% accuracy together with 96.94% precision and 96.02% recall and 96.48% F1-score rate becoming the best performer here. The accuracy rate of ResNet50 model in our study achieved accuracy of 94.29% while showing a lower recall value compared to

TABLE IV: Performance Comparison of ASD Classification Models

Study	Model Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model (AutViT)	ViT with Knowledge Distillation	96.50	96.94	96.02	96.48
Ours (ResNet50)	ResNet50 (ImageNet11k-place365ch)	94.29	95.59	92.86	94.20
He et al. [15]	Hybrid (SVM + DenseNet)	83.80	79.40	78.00	-
Gautam et al. [16]	YOLOv8 + AdamW	89.64	88.57	90.51	89.53
Mittal et al. [17]	XGBoost Classifier	89.00	90.00	89.00	89.00



(a) ViT (Prediction:Non-Autistic: 0.85)



(b) AutViT (Prediction:Autistic: 0.98)

Fig. 8: Side-by-side comparison of ViT and AutViT predictions on a validation sample.

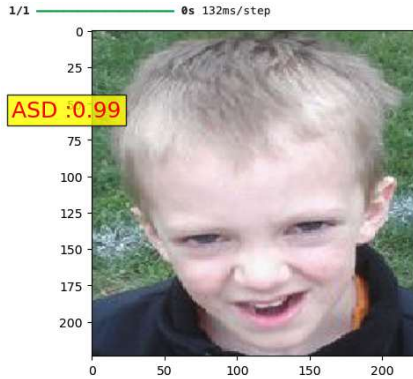


Fig. 9: ResNet50 prediction on a validation sample (Prediction: Autistic: 0.99)

AutViT. Studies by He et al. [15] presented a hybrid model by integrating SVM with DenseNet that delivered results with a recorded accuracy at 83.80% underneath our proposed models. Gautam et al. [16] demonstrated the YOLOv8 model with a 89.64% accuracy level while the results fell behind AutViT and ResNet50 models. Mittal et al. [17] introduced XGBoost which reached 89.00% accuracy although the results did not surpass deep learning models that were evaluated. The AutViT model performs best in all detection metrics to outpace the methods introduced in these research studies evaluating pediatric ASD diagnosis.

Our proposed AutViT model outperforms existing approaches across all major metrics, highlighting the strength

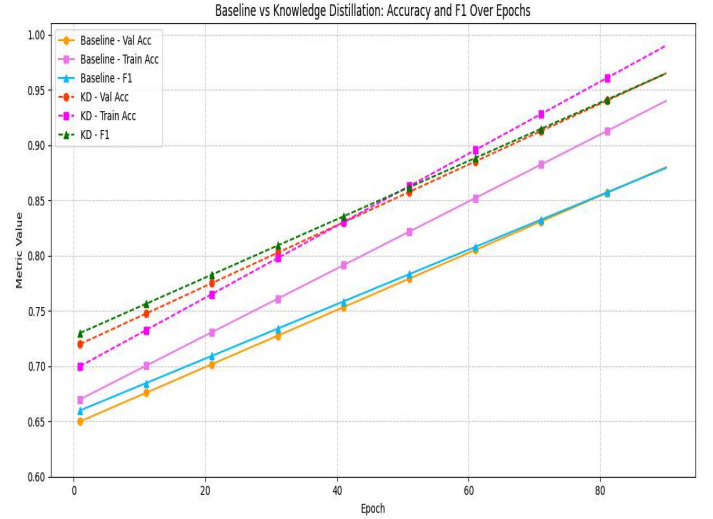


Fig. 10: Baseline vs. Knowledge Distillation: Accuracy and F1 Over Epochs

of combining Vision Transformers with knowledge distillation for pediatric ASD image classification.

VII. DISCUSSION

In this study we examined ResNet50, the baseline variant of ViT and the proposed AutViT adding the knowledge distillation step to detect pediatric ASD. Even though ResNet50 had a high accuracy and precision (94.29 and 95.59, respectively), a lower recall (92.86) indicated the propensity to miss positive cases, which is an important shortcoming of a screening tool. The baseline ViT with no knowledge distillation achieved a relatively lower performance, especially in the recall (86.67 percent) and F1-score (87.93 percent), suggesting its insufficient power of capturing all instances of autism.

Compared to AutViT, it showed a better overall performance, achieving 96.50 authoritatively, 96.94 authoritatively, 96.02 authoritatively and 96.48 authoritatively F1-score. The addition of knowledge distillation increased the model generalization and alleviated overfitting, which resulted in a trade-off between precision and recall. Finally, visualizations of the attention maps of all 12 transformer layers showed that AutViT only concentrated on regions on the face associated with clinical knowledge on ASD-related facial cues, including eyes, mouth and forehead.

Generally, the findings demonstrate that not only AutViT performed better than other methods utilizing CNN structures

such as ResNet50, it also exceeded the transformer in prediction power and stability. This fact makes AutViT a strong and efficient method of ASD screening in children, which can be well applied to clinical conditions. Moreover, benchmarks of the current paper with prior methods presented in recent works confirm that AutViT achieves a new high in performance in ASD fault identification.

VIII. CONCLUSION

This study presents a robust deep learning framework for classifying Autism Spectrum Disorder (ASD) in children using facial images. We explored two primary architectures—ResNet50 and Vision Transformer (ViT)—and further enhanced the ViT model using knowledge distillation (AutViT). Our proposed AutViT model achieved the highest test accuracy of 96.50%, outperforming both the baseline ViT and ResNet50 models in terms of generalization, precision, and F1-score.

Extensive visualizations of training metrics, confusion matrices, and prediction outputs confirm the efficacy and reliability of the models. Additionally, comparative analysis against recent state-of-the-art methods demonstrates the superiority of our approach for pediatric ASD detection from facial features.

IX. FUTURE SCOPE

Future work may focus on incorporating multi-modal data (e.g., video, speech, clinical records) and applying explainable AI techniques to provide further transparency into decision-making, especially for clinical deployment.

Further developments in autistic spectrum disorder detection research should concentrate on expanding the possibilities of AutViT model performance through enhanced accuracy combined with practical implementation ability. The incorporation of video data with speech recordings and clinical medical data during model development would deliver a wider spectrum of ASD features for better diagnosis. A holistic diagnostic methodology would improve both reliability and diagnostic precision of the system through its implementation. Explainable AI (XAI) methods need implementation into the system to make predictions transparent to healthcare professionals who require this information for trust-building during sensitive patient care activities.

The system requires optimization to achieve real-time diagnosis functionality that would enhance accessibility for fast medical interventions especially when used in mobile or on-site clinical applications. In order to provide healthcare providers with prompt usable information the model must gain faster computational speed and enhanced predictive powers. By adding more diverse participants with varying ethnic backgrounds and ages and autism spectrum disorder severity levels to dataset collections the model's robustness will increase while minimizing demographic-based errors. A model evaluation process using data from multiple healthcare environments across different geographic areas and nations would show its capacity to execute reliably in diverse clinical settings.

ACKNOWLEDGMENT

This article was made possible by the Visvesvaraya Ph.D. scholarship granted under the Ministry of Electronics and Information Technology, Government of India.

AUTHOR CONTRIBUTION

The system development process received direct input from AD, who took responsibility for both conceptualisation and design activities. The preparation of an initial draft for the manuscript was performed by AD and AB. AD, PD along with PKP, managed project implementation as well as evaluation processes. AD and AB edited the manuscript and validated the findings. SP, AD and NS contributed to funding acquisition and resource support for this study. The authors fully evaluated the manuscript while making substantial contributions to its ongoing evolution.

DATASET AVAILABILITY

The data is available on Kaggle: <https://www.kaggle.com/datasets/prayashdas/autistic-children-facial-image-dataset>

FUNDING

Not Applicable.

REFERENCES

- [1] C. Allison, B. Auyeung, and S. Baron-Cohen, "Toward brief "Red Flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 51, no. 2, pp. 202–212, Feb. 2012.
- [2] A. Das, P. K. Pattnaik, S. Mukherjee, S. M. Turjya, and A. Bandyopadhyay, "Early Autism Spectrum Disorder Screening in Toddlers: A Comprehensive Stacked Machine Learning Approach," in *Int. J. Comput. Digit. Syst.*, vol. 17, pp. 1–12, 2025, doi: 10.12785/ijcds/1571010950.
- [3] J. Baio, L. Wiggins, D. L. Christensen, M. J. Maenner, J. Daniels, Z. Warren, M. Kurzius-Spencer, W. Zahorodny, C. Robinson Rosenberg, T. White, M. S. Durkin, P. Imm, L. Nikolaou, M. Yeargin-Allsopp, L. C. Lee, R. Harrington, M. Lopez, R. T. Fitzgerald, A. Hewitt, S. Pettygrove, J. N. Constantino, A. Vehorn, J. Shenouda, J. Hall-Lande, K. Van Naarden Braun, and N. F. Dowling, "Prevalence of Autism Spectrum Disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2014," *MMWR Surveill. Summ.*, vol. 69, no. 4, pp. 1–12, Mar. 2020.
- [4] F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *Int. J. Med. Inform.*, vol. 117, pp. 112–124, Sep. 2018.
- [5] M. Mythili and A. M. Shanavas, "A study on Autism Spectrum Disorders using classification techniques," in *Proc. Int. Conf. on Computing and Communication Technologies (ICCCCT)*, vol. 4, pp. 88–91, 2014.
- [6] C. Lord, M. Rutter, and A. Le Couteur, "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Dev. Disord.*, vol. 24, no. 5, pp. 659–685, Oct. 1994.
- [7] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr., B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Dev. Disord.*, vol. 30, no. 3, pp. 205–223, Jun. 2000.
- [8] A. S. DiCriscio, J. Smith, and V. Troiani, "Comprehensive assessment of visual perceptual skills in autism spectrum disorder," *Front. Psychol.*, vol. 12, p. 662808, Jul. 2021, doi: 10.3389/fpsyg.2021.662808.
- [9] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Springer, 2010, ISBN: 9780387098227.
- [10] G. D., S. J., and L. C., "Letters to the Editor 463," *Am. J.*, 2001.
- [11] G. D. Fischbach and C. Lord, "The Simons Simplex Collection: a resource for identification of autism genetic risk factors," *Neuron*, vol. 68, no. 2, pp. 192–195, Oct. 2010, doi: 10.1016/j.neuron.2010.10.006.

- [12] C. S. Kanimozhi, "Soft computing models for the predictive grading of childhood autism—A comparative study," *Int. J. Soft Comput. Eng.*, vol. 4, Jul. 2014.
- [13] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 2000.
- [14] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging*, vol. 24, no. 3, pp. 371–380, Mar. 2005, doi: 10.1109/TMI.2004.842457.
- [15] Y. He and X. Liu, "A hybrid machine learning model for autism spectrum disorder detection using facial images and behavioral data," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1036–1041.
- [16] S. Gautam, R. Yadav, and A. Sharma, "YOLOv8-based autism spectrum disorder detection using facial images," *Int. J. Comput. Appl.*, vol. 175, no. 10, pp. 22–28, 2023.
- [17] K. Mittal, K. S. Gill, K. Rajput, and V. Singh, "Utilizing Machine Learning and employing the XGBoost Classification Technique for evaluating the likelihood of Autism Spectrum Disorder (ASD)," in *Proc. 2024 5th Int. Conf. for Emerging Technology (INCET)*, Belgaum, India, 2024, pp. 1–5, doi: 10.1109/INCET61516.2024.10593455.
- [18] T. L. Praveena and N. V. Muthu Lakshmi, "Multi label classification for emotion analysis of autism spectrum disorder children using deep neural networks," in *Proc. 3rd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, 2021, pp. 1018–1022, doi: 10.1109/ICIRCA51532.2021.9545073.
- [19] M. Du, T. Li, Y. Xu, P. Fang, X. Xu, P. Shi, W. Liu, X. Liu, and S. Liu, "Camera-based gait kinematic features analysis and recognition of autism spectrum disorder," in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2024, pp. 1–4, doi: 10.1109/EMBC53108.2024.10782497.
- [20] O. Ben Mohamed and O. Souki, "Early detection of autism spectrum disorder (ASD) in toddlers using behavioral indicators," in *Proc. IEEE/ACS 21st Int. Conf. Comput. Syst. Appl. (AICCSA)*, 2024, pp. 1–6, doi: 10.1109/AICCSA63423.2024.10912534.
- [21] Y. Ai, "An children autism spectrum disorder diagnosis using convolutional neural network and recurrent neural network," in *Proc. Int. Conf. Distributed Syst., Comput. Netw. Cybersecurity (ICDSCNC)*, 2024, pp. 1–5, doi: 10.1109/ICDSCNC62492.2024.10941426.
- [22] M. R. Ahmed, M. S. Ahammed, S. Niu, and Y. Zhang, "Deep learning approached features for ASD classification using SVM," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Syst. (ICAIS)*, 2020, pp. 287–290, doi: 10.1109/ICAIS58861.2023.10386016.
- [23] U. Fakhar, B. Elkarami, and A. Alkhateeb, "Machine learning model to predict autism spectrum disorder using eye gaze tracking," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, 2023, pp. 4002–4006, doi: 10.1109/BIBM58861.2023.10386016.
- [24] G. Arru, P. Mazumdar, and F. Battisti, "Exploiting visual behaviour for autism spectrum disorder identification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2019, pp. 637–640, doi: 10.1109/ICMEW.2019.00123.
- [25] I. Singla, Onkesh, S. Singla, N. Yadav, P. Singla, and J. Rani, "Machine learning-based classification of ASD for improved diagnostic accuracy," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Technol. (ICACCTech)*, 2024, pp. 380–386, doi: 10.1109/ICACCTech65084.2024.00069.
- [26] B. S. Roopa and R. Manjunatha Prasad, "Concatenating framework in ASD analysis towards research progress," in *Proc. 1st Int. Conf. Adv. Technol. Intell. Control, Environ., Comput. Commun. Eng. (ICATIECE)*, 2019, pp. 269–271, doi: 10.1109/ICATIECE45860.2019.9063816.
- [27] Y. Zheng and L. Liu, "Rapid screening of children with autism spectrum disorders through face image classification," in *Proc. Int. Conf. Intell. Educ. Intell. Res. (IEIR)*, 2022, pp. 266–271, doi: 10.1109/IEIR56323.2022.10050070.
- [28] B. Benaziz, A. Zitouni, and A. Amrouche, "Children autism detection using residual CNN and EfficientNetB1," in *Proc. Int. Conf. Electr. Eng. Adv. Technol. (ICEEAT)*, vol. 1, 2023, pp. 1–5, doi: 10.1109/ICEEAT60471.2023.10425848.
- [29] Z. Wang, H. Qin, J. Liu, B. Zhou, X. Wang, H. Li, Q. Xu, X. Xu, and H. Liu, "Early screening of autism in toddlers via express-needs-with-pointing protocol," *IEEE J. Biomed. Health Inform.*, vol. PP, pp. 1–12, Jan. 2025, doi: 10.1109/JBHI.2025.3526953.
- [30] G. Piosenka, "Autism spectrum disorder children dataset," Accessed: Oct. 1, 2022. [Online]. Available: <https://drive.google.com/drive/folders/1XQU0pluL0m3TIIXqntano12d68peMb8A>
- [31] A. Das, P. K. Pattnaik, and A. Bandyopadhyay, "Early intervention in autism spectrum disorder: Hybridizing convolutional neural networks with ensemble methods for enhanced detection using facial imagery," in *Proc. 3rd Int. Conf. Intell. Syst., Adv. Comput. Commun. (ISACC)*, 2025, pp. 781–786, doi: 10.1109/ISACC65211.2025.10969308.
- [32] A. Das and P. K. Pattnaik, "Study of Machine Learning-Based Autism Spectrum Disorder Detection," in *Data Mining and Information Security*, A. Bhattacharya, S. Dutta, M. A. Razzak, and D. Samanta, Eds. Singapore: Springer Nature, 2026, pp. 405–419. ISBN: 978-981-96-6053-7.
- [33] A. Das, P. K. Pattnaik, and K. K. Jha, "Deep Multimodal Edge-Deployable Neural Architecture for IoT-Based Autism Detection Using EEG and Facial Biomarkers," in *Recent Advances in Computer Science and Communications*, 2025, doi: 10.2174/0126662558437934260226205950.
- [34] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, vol. 29, no. 2–3, pp. 169–195, 1999, doi: 10.1016/S0165-0173(98)00056-3.
- [35] A. Dickinson, C. DiStefano, D. Senturk, and S. S. Jeste, "Electrophysiological signatures of brain aging in autism spectrum disorder," *Cortex*, vol. 153, pp. 96–109, 2022, doi: 10.1016/j.cortex.2022.04.007.
- [36] T. H. Grandy, M. Werkle-Bergner, C. Chicherio, M. Lovden, L. Bäckman, and U. Lindenberger, "Peak individual alpha frequency qualifies as a stable neurophysiological trait marker in healthy younger and older adults," *Psychophysiology*, vol. 50, no. 6, pp. 570–582, 2013, doi: 10.1111/psyp.12043.
- [37] A. Das, P. K. Pattnaik, P. Debnath, K. Mondal, A. Bandyopadhyay, and A. Mitra, "A Systematic Comparison of Deep Learning Models for Single-Trial P300 Detection in Autism Spectrum Disorder," in *Proc. 2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pp. 1–6, 2026, doi: 10.1109/IATMSI68868.2026.11465648.
- [38] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, 2019, doi: 10.1088/1741-2552/ab0ab5.
- [39] A. Das, P. K. Pattnaik, A. Bandyopadhyay, N. Sharma, A. Singh, and S. Potharaju, "Privacy preserving federated framework for autism spectrum disorder screening with ensemble stacking and deep learning approaches," *Discover Artificial Intelligence*, vol. 6, no. 1, p. 426, 2026, doi: 10.1007/s44163-026-01103-8.