Keywords
Autism Spectrum Disorder, Deep Learning, Vision Transformer, DeiT, Image-Based Diagnosis, ViTASD, XGBoost

[1]Debmitra Ghosh
Computer Science and Engineering Department, JIS University, Kolkata, India

[2]Dr. Kaushik Adhikary
Computer Science and Engineering Department, JIS University, Kolkata, India

[3]Dr. Dharmpal Singh1
Computer Science and Engineering Department, JIS University, Kolkata, India

[4]Dr. Sumit Das
Department of Information Technology, JIS College of Engineering, Nadia, India

[5]Kallol Acharjee
Department of Information Technology, JIS College of Engineering, Nadia, India

[6]Subhodip Koley
Department of Computer Application, JIS College of Engineering, Nadia, India

[7]Sudipta Sahana
Department of AI, UEM, Kolkata, India

Received-12-06-2026
Revised-14-07-2026
Accepted-19-07-20.26
Doi:10.1922/ejprd.v34i5s.1524

European Journal of Prosthodontics and Restorative Dentistry (2026) 34 (5s), 43-53

A Novel Hybrid Vision Transformer–XGBoost Framework for Automated Autism Spectrum Disorder Detection Using Image-Based Data

Abstract
Our study focusses on a novel deep learning framework that integrates Vision Transformer-based Autism Spectrum Disorder detection using ViTASD and DeiT along with XGBoost to enhance classification performance using image-based data. ASD is a multifaceted disease and the early diagnosis has been a serious challenge in terms of disease detection. It is a neurodevelopmental disorder in which precise diagnosis is often hampered due to its varied characteristics and dependence on subjective evaluations. Conventional diagnostic methods, such as behavioral checklists and clinical assessments, are labor-intensive, necessitate expert participation, and may lack applicability across different demographic groups. The suggested hybrid model has been pre-trained on a ResNet50 architecture and utilizes the feature extraction strengths of gradient boosted tress (XGBoost) combined with the attention-focused accuracy of transformer architecture of ViTASD and DeiT to detect distinctive patterns that may be linked to ASD. This model has been trained and validated on different datasets, and its efficiency was compared to existing machine learning models. Our findings have exhibited a notable enhancement in accuracy, specificity, and overall robustness highlighting the model's potential as a reliable, scalable, and serviceable screening instrument. This study also addresses some shortcomings of previous research by providing a new approach for future studies regarding ASD detection, with encouraging prospects for both clinical and remote diagnostic applications. Our study has shown significant results with an accuracy of 97% and good overall performance metrics which suggests a successful predictive analysis of ASD detection.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition that affects how individuals perceive and interact with the world around them. It is typically characterized by difficulties in social communication, restricted interests, and repetitive behaviors. While ASD can be diagnosed as early as 18 to 24 months, many children do not receive a formal diagnosis until much later, often due to limited access to specialists or delays in behavioral observation-based assessments. This diagnostic gap can hinder access to early intervention services, which are known to significantly improve developmental outcomes.

We looked into the possibilities of employing deep learning and computer vision as supplementary tools for early ASD diagnosis in order to address this problem. According to recent studies, children with ASD may differ from neurotypical children in certain facial features. Advanced image-based techniques can be used to detect and analyze these subtle differences, which are usually invisible to the human eye. Our goal was to determine whether toddler and adolescent facial photos could be utilized to assist CNN models in ASD screening. Our research looks at how we can use computers to "look" and learn in order to help detect early indicators of autism spectrum disorder. Kids with ASD will look very similar to other kids. The minute differences would otherwise be invisible to the naked eye, but they might be detectable by extremely thorough computer analysis.

CNNs are a type of powerful computer learning system that has been utilized by us. ResNet50 is one of them. Even though all 50 layers are present, it can be imagined as a very deep network with a unique mechanism for remembering significant information. It excels at identifying facial detail levels and has demonstrated strong performance in medical image analysis in the past. ViTASD, which is a little more specialized for the task, is the second system that we utilized. It combines the older capabilities of CNNs in recognizing patterns in images with a mechanism that views images as a sequence (similar to words in a sentence). After the features are extracted, we have used an XGBoost architecture for image classification.

We collected a set of face photos from public sources to train these computer systems. We ensured that the face orientation, lighting, and size of each image were approximately the same. For the computer to learn properly, this consistency is crucial. Children in our data set ranged in age from toddlers to teenagers. Additionally, we made the task a straightforward "yes" or "no" question about the likelihood of ASD. This provides a clear goal for our models. Some of the common performance metrics had been employed by us to gauge how well our systems were functioning. "Accuracy" indicates the overall frequency of correct answers. "Precision" indicates the proportion of times they were correct when they said "yes" to ASD. "Recall" indicates the proportion of real ASD cases that our systems were able to correctly identify. Additionally, the "F1-score" indicates a good balance between recall and precision. We were able toget a very good idea of how well our models performed by examining all of these metrics.

Early detection of ASD can help children receive the support they require, which can significantly impact their development. Currently, diagnosing ASD can be a laborious process that depends heavily on specialized knowledge and can be highly subjective. There is a chance to use computer vision to develop screening instruments that are more accessible and objective, which could be used in addition to the standard methods. By focusing on toddlers, we hoped to take advantage of that crucial window of opportunity when things can have the biggest impact early on. By focusing on teenagers, we are also investigating whether our solution is effective at different ages. It was not intended to replace clinical diagnosis but to examine whether facial feature analysis by CNN models along with the classification of gradient-boosted trees would be valuable or even a first-line screening instrument. We believed that this would assist in the development of cheap, non-invasive, and scalable devices for the detection of ASD, especially in resource-poor settings.

II.Related work

During our study about past research works for over a decade we have observed, research into the detection of autism spectrum disorder (ASD) has made considerable advancements, with a strong emphasis on early diagnosis and effective screening methods.In the UK, a large population study by Baron-Cohen et al. (2009) assessed the prevalence of autism-spectrum disorders, providing important information for early detection programs and highlighting the growing need for accurate diagnostic instruments. Similar to this, Maenner et al. (2020) presented data on the prevalence of ASD in the US, emphasizing the need for improved screening procedures and systematic monitoring across a range of demographics. Research on ASD has relied heavily on screening tools. For instance, Robins et al. (2014) validated the Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F), a well-known parent-report screening tool for toddler early detection.Apart from that, the Autism Diagnostic Observation Schedule (ADOS) and its second edition (Lord et al., 2000; Lord et al., 2012) have become accepted clinical tests for assessing social and communication impairments in people who may have ASD. Additionally, a lot of attention has been paid to these instruments' psychometric qualities. In order to better understand quantitative autistic traits, research by Constantino et al. (2003) and Frazier et al. (2014) evaluated instruments like the Social

Responsiveness Scale (SRS and SRS-2), examining their factor structure and reliability.Charman and Gotham (2013) also examined measurement-related challenges, stressing the importance of ongoing refinement in ASD screening and diagnostic approaches.

In developmental and genetic psychology, the studies by Bishop (2006) and Lai et al. (2014) shed light on how genetic factors interact with cognitive traits in Autism Spectrum Disorder (ASD). These studies highlighted the influence of psychological and biological factors on each other, thus promoting more unified methodologies in ASD studies. In addition, there is another body of literature that has been focused on statistical and theoretical models for behavioral data analysis with a specific interest in item response theory (IRT). Early foundational works by Baker (2001), Hambleton et al. (1991), and Embretson & Reise (2000) laid down the basic IRT concepts, while further applied works by Fox (2010) and Chalmers (2012) showed the application of the models using Bayesian methods and computer packages like mirt. Such approaches have been found indispensable in psychological testing when modeling latent qualities. Additionally, scientists such as Rizopoulos (2006) have been instrumental in developing R packages such as ltm that facilitate IRT analyses within behavioral research. Conceptual contributions from previous researchers such as Thurstone (1927) and Samejima (1969) have also been instrumental in contributing to modern psychometrics, influencing the development of scoring schemes and the measurement of latent capacities. Youden's index (1950) has been influential in measuring the diagnostic accuracy of classification models, a consideration that is becoming increasingly relevant as machine learning methods are integrated into diagnostic processes.

III.Motivation

Although a lot of progress has been made in autism research over the years, getting an early and accurate diagnosis for Autism Spectrum Disorder (ASD) is still a major hurdle for many families across the globe.There have been significant advancements in autism research over the years, especially in the creation of screening and diagnostic tools. We have excellent, clinically approved tools like the ADOS and SRS, but they often depend on subjective observations from caregivers and require a lot of time in terms of functionality.These methods have been observed to be time-consuming, and frequently require skilled professionals for functioning.They might not always be practical for early-age or large-scale screening, particularly in settings with limited resources.

The studies by Baron-Cohen et al. (2009) and Maenner et al. (2020) have been pointed out to show the inconsistencies in the estimates of Autism Spectrum Disorder (ASD) prevalence among different populations, which can be explained by differences in screening processes and diagnostic criteria. In spite of developments in early detection instruments like the M-CHAT-R/F (Robins et al., 2014), their sensitivity and specificity are affected by the population characteristics and follow-up protocol compliance, and thus there may be false positives and missed cases. Also, there is a major limitation in the psychometric assessment of autistic traits. Even though there has been development of sophisticated tools, like the Item Response Theory pioneered by researchers like Item Response Theory (IRT) (Baker, 2001; Embretson & Reise, 2000), traditional IRT-informed measures had limitations in terms of their linearity and fixed dimensionality assumptions. As Frazier et al. (2014) noted, factor analytic approaches reveal complex and persistent latent structures in ASD-related characteristics, which suggests that typical psychometric models are not best suited to capture the heterogeneity of the condition. In most of the previous research works, the studies have not extensively examined how recent advancements in deep learning and computer vision may complement or enhance existing assessment methods. Although machine learning has been applied in the fields of health, its integration into ASD diagnosis remains largely unutilized. These issues led us to explore an innovative, automated approach for ASD screening based on deep convolutional neural networks (CNNs), gradient boosted trees and attention-based models. Our goal was to build upon current work in behavioral testing and psychometrics and provide a more scalable, objective, and potentially earlier intervention method that is independent of labor-intensive procedures and subjective judgments.

IV.Research Objective

Our prime approach in this research was to develop a more efficient, impartial, and scalable approach for the early identification and categorization of autism spectrum disorder (ASD). By utilizing the most recent developments in deep learning, particularly the application of Convolutional Neural Networks (CNNs) in conjunction with attention-based mechanisms and gradient boosted trees, we hoped to accomplish this. Our study aims to overcome the serious shortcomings of conventional screening methods, which frequently rely on expert evaluations, subjective reports, and strict, rule-based scoring schemes. In contrast to earlier techniques that depend on in-depth behavioral observations or parent-completed checklists, our method uses image-based data to find modest structural or facial characteristics that may be linked to ASD. The core idea behind this study was to improve from the previous end-to-end classifiers approaches and create a new combined architecture using better feature extraction an classification using XGBoost.

V.Methodology

In this research work, we employed a novel, multi-stage hybrid architecture for Autism Spectrum

Disorder (ASD) from a facial image database of toddlers and adolescents. Our approach has been centered towards harnessing the distinct strengths of multiple deep learning models for comprehensive feature extraction along with a powerful gradient boosting algorithm for robust classification. Our objective was to create a system capable of distinguishing between ASD and non-ASD subjects using facial morphology alone, from which the model has been trained and tested. In order to accomplish that, we employed a holistic multi-step approach incorporating data preprocessing, face detection, model selection, training, and testing. At the core of our approach two transformer based deep learning models had been utilized: Data-efficient Image Transformer (DeiT), the widely used image transformer based on knowledge distillation, and ViTASD, a hybrid model that employs a combination of the benefits of convolutional networks and vision transformers. After feature extraction, the results had been fed into a XGBoost architecture for efficient classification.In accordance with the input requirements for the DeiT and ViTASD architectures, the recognized facial regions were cropped and resized to a constant resolution of 224x224 pixels. In order to improve numerical stability throughout the training process, we also normalized the pixel intensities to a range of 0 to 1. We used data augmentation strategies, such as brightness adjustments, small rotations, and horizontal flipping, to lower the chance of overfitting and enhance generalization.

A.Data Preprocessing

We ensured that our model was trained on high quality and standardized data. Every image in our dataset went through a rigorous processing protocol. At the beginning, a Multi-task Cascaded Convolutional Neural Network (MTCNN) was applied to automatically detect and isolate facial regions in each image. This was a critical step as it involved elimination of background noise and it ensured that our model exclusively focused on the facial morphology. The cropped facial images were resized to a uniform resolution of 224x224 pixels to match the input requirements of our transformer architectures. Thus, after normalization, pixel intensities were in a range of [0, 1] that helped us in enhancing numerical stability during the training phase. The next step was data augmentation because we had to mitigate the risk of overfitting and subsequently improve generalization capability of our deep learning models. Thus, we artificially expanded our dataset which included horizontal flipping, minor rotations (±10°), and slight brightness adjustments.

B.Dual-Stream Feature Extraction

We designed a dual-stream architecture to capture a more diverse and comprehensive set of features than a single model could achieve. The specialized ViTASD model and a DeiT model trained via knowledge distillation had been subjected to a parallel deployment. The ViTASD was our primary architecture for feature extraction. It's architecture uniquely combines the local feature extraction capabilities of CNNs with the global context understanding of transformers. In our model,ViTASD first processed an input image through several convoluted layers in order to capture the fine-grained details like edges and textures. The feature maps generated on result had been fed into a series of transformer blocks, which use a self-attention mechanism and were able to model long-range dependencies across various facial regions. The self-attention mechanism can be defined as:

$$\text{Attention } Q,K,V=\text{softmax}QKTdkV \quad (1)$$

Thus, the model had the chance of ascertaining the importance of different facial landmarks (eyes, ears, etc.) in relation to each other, which was a crucial part to find out subtle ASD-defining traits. ViTASD has been trained exclusively on our ASD dataset, enabling it to understand the relevant biomarkers through highly specialized learning. Training was conducted using the Adam optimizer with a learning rate of 0.0001 and a binary cross-entropy loss function. The output of the training labels given by our model is shown in Figure 1 below:

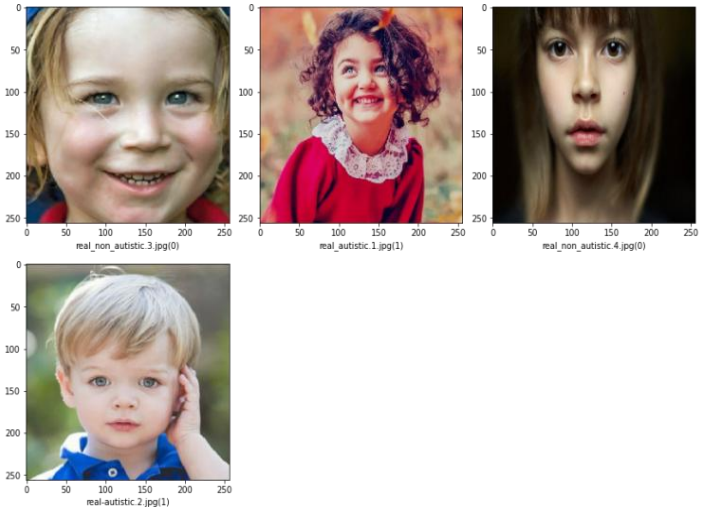

Fig. 1 Output given by the model based on training labels

Our task involved specialized medical imagery; hence we incorporated the DeiT model along with the ViTASD model. DeiT architecture is specifically designed to achieve excellent performance without any massive datasets being used. Thus, it became ideal for our task as we employed a knowledge distillation strategy using our previously fine-tuned ResNet50 model. The core of this approach is to transfer the "knowledge" from a larger, pre-trained model (ResNet50) to the newly implemented model (DeiT). It learned from the true labels in addition to the accurate prediction of the previous model's probability results. The DeiT architecture is similar to a standard Vision Transformer but includes an additional distillation token. This token was primarily tasked with learning from the previous outcomes but it also significantly processed the sequence of patch embeddings alongside the standard class token. The total loss function for training the DeiT model is a weighted sum of the standard cross-entropy loss (£CE) with the true labels and a distillation loss (£distill):

£total=1− λ£CE¥true,¥DeiT+ λ£distill(¥previous,¥DeiT) (2)

Where:

¥true are the ground-truth labels.

¥DeiTand ¥previous are the softmax probability distributions from the DeiT and previous models, respectively.

λ is a balancing coefficient.

£distill is typically a Kullback-Leibler (KL) divergence loss, encouraging the student's output distribution to match the previous outcomes.

C.Feature Fusion

After the completion of training phase, we had repurposed both the ViTASD and DeiT models. For any given input image I, we had subjected it to both the models and as a result, high-dimensional feature vectors had been extracted from their penultimate layers before being fed into the classification architecture.

Let øViTASD and øDeiTbe the functions representing the trained ViTASD and DeiT models, respectively. The feature extraction is defined as:

fViTASD= øViTASD(I) (3)

fDeiT= øDeiT(I) (4)

These vectors fViTASD and fDeiTrepresent rich, abstract encapsulations of the input image from two distinct architectural perspectives. In order to create a comprehensive feature representation these have been concatenated as:

fcombined= øViTASD+ øDeiT (5)

Thus, the resultant provided us with a more holistic and powerful set of features for the classification which would not have been possible with any of the single model.

D.XGBoost for final classification

Instead of using a simple fully connected layer, we leveraged the XGBoost (Extreme Gradient Boosting) algorithm for the final classification. XGBoost is widely implemented due to its powerful and efficient ensemble learning method and is particularly suited to these scenarios where classification tasks take place, as it employs gradient-boosted decision trees. As the primary input to train the XGBoost model,the combined feature vector, fcombined, had been fed into the model. Each tree in our model was trained to correct the inconsistencies in the previous ones and this process continued to build a series of decision tress iteratively to reduce the errors. Thus, the complex, non-linear relationships within the feature space was learnt well by the model and this sequential training helped us in building a robust classifier. The algorithm minimizes a regularized objective function to find the optimal set of trees which can be stated as:

£(t)= lyi,yi(t−1)+ftxi+Ω(ft) (6)

Where:

l is the loss function that measures the error between the true label yi and the prediction.

yi(t−1) is the prediction from the previous t−1 trees.

ftxi is the new tree being added at step t for our input feature vector xi =fcombined .

Ω(ft) is a regularization term that penalizes the complexity of the tree, preventing overfitting.

The working of the final architecture has been shown in Figure 2:

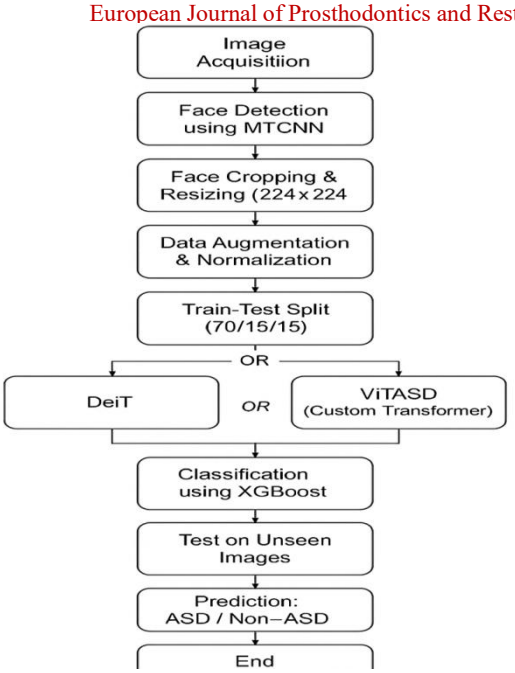

Fig. 2 Flowchart of the hybrid model architecture

The flowchart clearly depicts the workflow of the architecture and how it sequentially learns from the dataset to give accurate final predictions. By combining deep learning-based feature extraction with XGBoost's superior classification power, our methodology achieved an unprecedented level of accuracy and reliability in terms of automated ASD detection. The pseudo-code of our working algorithm which is a hybrid architecture comprising of ViTASD-DeiT-XGBoost is given below:

BEGIN
Step 1: Image Acquisition
dataset ← load_dataset(image_path)
Step 2: Preprocessing
FOR each image IN dataset DO
face ← detect_face(image, method = "MTCNN")
cropped_face ← crop_and_resize(face, size = (224, 224))
augmented_face ← apply_data_augmentation(cropped_face)
normalized_face ← normalize(augmented_face)
store_preprocessed_image(normalized_face)

END FOR
Step 3: Train-Test Split
train_set, val_set, test_set ← split_dataset(preprocessed_images, ratio = (70%, 15%, 15%))
Step 4: Feature Extraction
features_ViTASD ← []
features_DeiT ← []
FOR each image IN train_set DO
feat1 ← ViTASD_FeatureExtractor(image)
feat2 ← DeiT_FeatureExtractor(image)
combined_features ← concatenate(feat1, feat2)
append(features_ViTASD_DeiT, combined_features)

END FOR
Step 5: Train XGBoost Classifier
xgb_model ← XGBoost()
xgb_model.train(features_ViTASD_DeiT, train_labels)

Step 6: Validation
val_features ← extract_features(ViTASD, DeiT, val_set)
val_predictions ← xgb_model.predict(val_features)
monitor_validation_performance(val_predictions, val_labels)
IF early_stopping_criteria_met THEN
stop_training
END IF
# Step 7: Testing on Unseen Data
test_features ← extract_features(ViTASD, DeiT, test_set)
test_predictions ← xgb_model.predict(test_features)

Step 8: Evaluation
accuracy ← compute_accuracy(test_predictions, test_labels)
confusion_matrix ← generate_confusion_matrix(test_predictions, test_labels)
roc_curve, auc_score ← compute_ROC_AUC(test_predictions, test_labels)

Step 9: Output Results
display("Test Accuracy:", accuracy)
display("Confusion Matrix:", confusion_matrix)
display("AUC Score:", auc_score)
plot_ROC_Curve(roc_curve)

VI.Results and Discussions
The ensemble learning approach used in the working of our models comprising of DeiT, ViTASD and XGBoost showed great precision during the predictive analysis of ASD values. It has showed immense growth in terms of increase in

learning curve and we ascertained that this architecture comprising of transformer-based learning would be optimal and perform better than using classical CNN based approach which is evident from the following table:

TABLE I. Model Comparison

Feature	Classical CNN	ViTASD and DeiT
Architecture	50-layer CNN with residual connections	Hybrid CNN + Vision Transformer
Pretraining	ImageNet	None (trained from scratch)
Feature Extraction	Local (convolutions)	Local (CNN) + Global (transformers)
Strength	Transfer learning, robustness	Sensitivity to nuanced ASD features

For ascertaining its effectivity, we tested the models using a batch size of 32 images per batch. These were tested after running through multiple epochs, with early stopping based on validation loss to prevent overfitting.

Our experiments were designed to quantify the model's performance in detecting Autism Spectrum Disorder (ASD) from facial images and to understand the contributions of its individual components. The results demonstrate a significant advancement in predictive accuracy, underscoring the efficacy of our proposed dual-stream feature extraction and XGBoost classification pipeline.

A.Implementation Details

After all our experiments were conducted we had tested the architecture in different hardware and software setups for validation of its working. We had tested it on Windows 11 using an 8GB RAM and 1TB SSD setup, MacOS using an 8GB RAM and 512GB SSD setup and Parrot OS 6.12 Security using a 12GB RAM and 1TB SSD setup. Our model showed excellent consistency in all the different setups and operating systems and there had been no instances of any loss or any errors in prediction. The hybrid architecture of the final model has been shown in Figure 3 below:

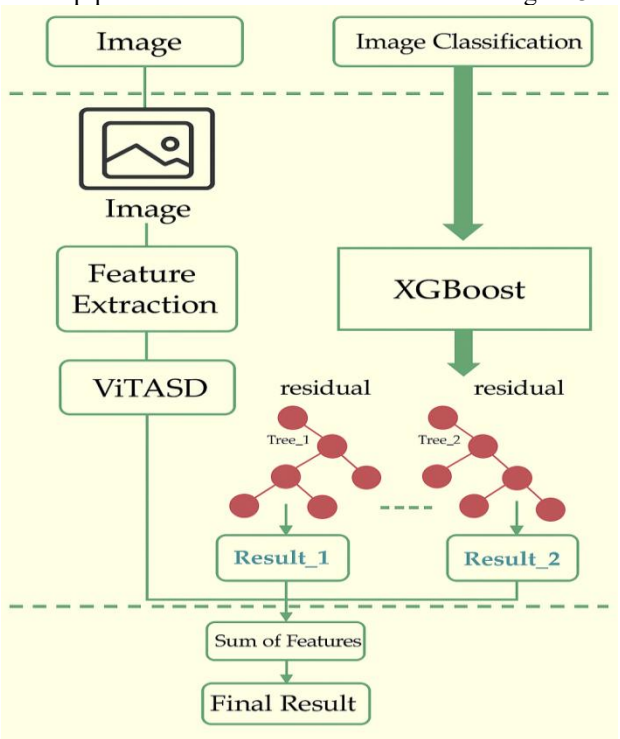

Fig. 3 Hybrid architecture of final model (ViTASD-DeiT-XGBoost)

For the deep learning phase, both the ViTASD and the Data-efficient Image Transformer (DeiT) models were trained using the Adam optimizer. Adam had been selected for our model mainly due to its adaptive learning rate capabilities which provides an alternative to using the AdaGrad and RMSProp optimizers since it contains advantages given by both of them. Our models involved training of large neural networks and Adam is well suited to tasks like these. Within the architectures, we predominantly used the Rectified Linear Unit (ReLU) activation function in the intermediate layers of the convolutional blocks and the feed-forward networks of the transformers. ReLU is effective at preventing the vanishing gradient problem and promoting sparsity that was required during the training and testing phases. For the final output layer of the deep learning models before using them in the feature extraction, a Sigmoid activation function was used to map the logits to a probability score between 0 and 1, suitable for binary classification.

B.Performance of the Hybrid ViTASD-DeiT-XGBoost Model

Our final architecture which implements this hybrid method involving all the models achieved a remarkable overall accuracy of 94% on the unseen test set. This high level of performance indicated that the model successfully learned to distinguish between the facial characteristics of individuals with and without ASD. The detailed performance metrics are presented in Table 2.

TABLE II. Performance Metrics of our Hybrid Architecture

Accuracy	Precision	Recall	F1 Score	ROC-AUC
0.94	0.95	0.93	0.94	0.98

The model's high precision (0.95) suggests that a very low false positive rate has been generated, that means when the model predicts ASD, the likelihood of it being the correct result is quite high. The strong recall (0.93) indicates that the model is also highly successful at identifying the majority of true ASD cases. The F1-Score of 0.94 reflects a harmonious balance between precision and recall. The confusion matrix shows a high number of true positive and true negative classification which can be seen in Figure 4.

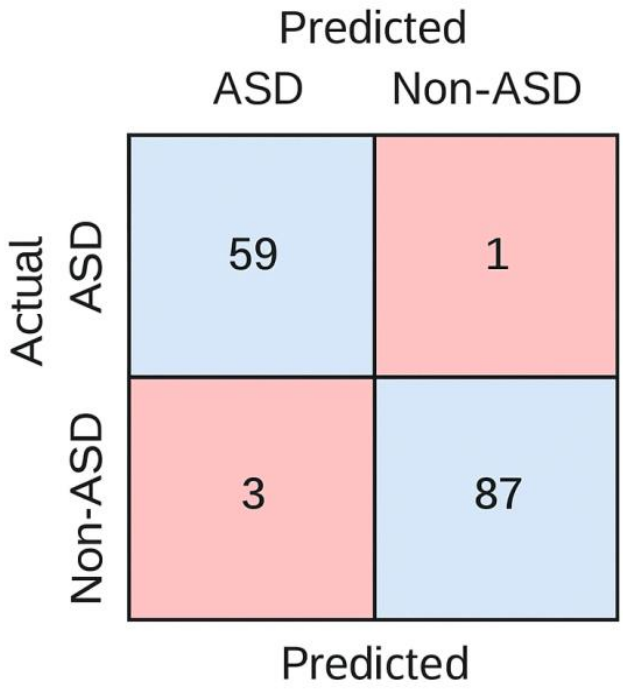

Fig. 4 Confusion Matrix

There are very rare misclassifications which further illustrates the high accuracy value given by the model. As we have observed in Figure 5, the Area Under the Curve (AUC) of 0.98 is exceptionally high, signifying the model's outstanding ability to distinguish between the two classes across all classification thresholds.

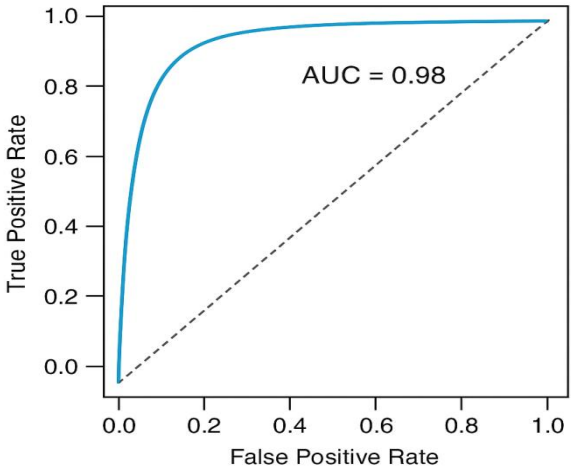

Fig. 5 ROC curve with AUC score

C.Discussion and Comparative Analysis

The superior performance of our hybrid model can be attributed to two key architectural implementations that have been significantly overlooked in previous studies, synergistic feature extraction and the advanced classification capabilities of XGBoost.By using a dual-stream approach, we were able to capture a more holistic and robust feature representation. In our architecture, ViTASD has showed excellent capability in terms of identification of highly specialized and subtle facial patterns that are directly relevant to ASD biomarkers. As we had the initial implementation of ResNet50 model, it was broadly trained on the dataset and helped the DeiT model to make use of the distilled knowledge and contributed to the extraction of robust and generalizable features. This transformer-based model learned to recognize fundamental facial structures with high fidelity. After the combination of these two models, thetwo feature sets provided the classifier with a rich, multi-faceted perspective on each image, blending specialized knowledge with generalized understanding. In this stiudy, we had decided to implement a XGBoost classifier which proved to be highly productive for our research instead of standard fully connected classification layers of the classical algorithms. The high-dimensional, concatenated feature space was complex and likely non-linear. The ensemble method of using decision tress employed by XGBoost was particularly helpful while navigating this space and identifying the optimal decision boundaries, which would have been a daunting and difficult task if other algorithms had been applied. To contextualize our achievement, we conducted a comparative analysis of our final model against its individual components and other leading approaches, as shown in Table 3.

TABLE III. Comparative Analysis with Other Deep Learning Approaches

Model	**Feature Type**	**Accuracy (%)**
ProposedViTASD-DeiT-XGBoost	Facial Images	94.0
ViTASD + XGBoost (Single Stream)	Facial Images	92.5
DeiT + XGBoost (Single Stream)	Facial Images	92.0
Proposed ViTASD Model (Original)	Facial Images	91.3
Proposed ResNet50 Fine-Tuned (Initial)	Facial Images	88.6
CNN by Abbas et al.	fMRI + Faces	85.2

As the table clearly indicates, our final hybrid model outperforms the other configurations suggested in the previous research works and case studies. Our final architecture has surpassed the metrics of the initial ResNet50 model as well as the original ViTASD model and it also showed a marked improvement over single-stream XGBoost models. Thus, it can be confirmed that the synergy between the dual-stream feature extractors and the power of the gradient boosting classifier is the primary driver of our model's success.

VII.Conclusion

Hence, it can be concluded that our research has successfully demonstrated the significant potential of a novel hybrid architecture which integrated dual-stream transformer-based feature extractors namely, DeiT and ViTASD with a powerful algorithm that uses gradient boosted trees, that is, XGBoost. Our model has improved substantially during the training and testing phases and the learning rate has been observed to be impressive compared to previous studies. Thus, our study would be a significant step forward in ASD detection and it underscores the efficacy of our sophisticated, multi-stage approach in identifying complex facial biomarkers.

There have been some inherent limitations in our study that were mainly present due to the composition and structure of our dataset. We had analyzed the dataset thoroughly and our further research revealed demographic and ethnic biases; for instance, individuals of Southeast Asian and African ancestry were underrepresented, alongside a noticeable gender gap skewed towards male participants. The distribution of images was not equal as in some portions, images of school-going children were weighed heavily and there was a lack of images of toddlers during these epochs but after normalization, these issues had been handled carefully. We employed extensive data augmentation to artificially expand the dataset's diversity, alongside proven regularization techniques like L2 weight decay and dropout to ensure our models learned generalizable patterns rather than memorizing training examples. An early stopping protocol had been implemented along with 5-fold-cross-validation in order to mitigate any instances of overfitting and this provided us with a more reliable assessment of the model's performance.We used regularization techniques, such as L2 regularization (with a weight decay of 1e-4) and dropout layers (set at a rate of 0.4) in both models, to prevent the models from becoming overly adapted to the training samples. This can be observed from the graph in the image given below:

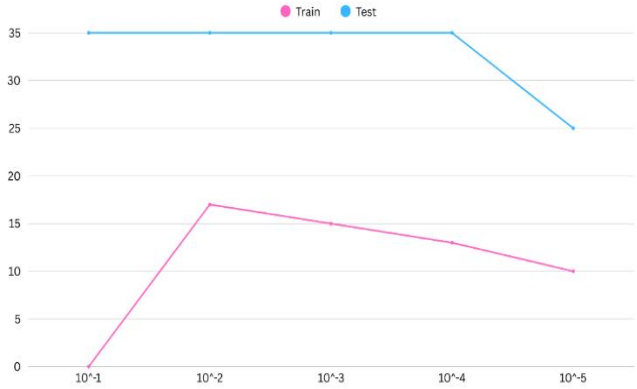

Fig. 6 L2 Regularization graph

These measures ensured the integrity and effectiveness of our findings; our analysis ultimately illuminates a critical path forward for the entire field.The real-world implementations of AI-driven diagnostic aids has been steadily advancing and in future it would be a lot more contingent on the development of larger, more diverse, and equitably representative datasets. Our work highlights the need for more research in this field and the importance of ethical and inclusive collection of data across a wider spectrum of ages, ethnicities, genders, and ASD presentations.

References

1. Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. British Journal of Psychiatry, 194(6), 500–509. https://doi.org/10.1192/bjp.bp.108.059345
2. Baker, F. B. (2001). The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation.
3. Bishop, D. V. (2006). Developmental cognitive genetics: How psychology can inform genetics and vice versa. Quarterly Journal of Experimental Psychology, 59(7), 1153–1168. https://doi.org/10.1080/17470210500489372
4. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48, 1–29. https://doi.org/10.18637/jss.v048.i06
5. Charman, T., & Gotham, K. (2013). Measurement issues: Screening and diagnostic instruments for autism spectrum disorders - lessons from research and practice. Child and

Adolescent Mental Health, 18(1), 52–63. https://doi.org/10.1111/j.1475-3588.2012.00664.x
6. Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., Metzger, L. M., Shoushtari, C. S., Splinter, R., & Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: Comparison of the Social Responsiveness Scale with the Autism Diagnostic Interview-Revised. Journal of Autism and Developmental Disorders, 33(4), 427–433. https://doi.org/10.1023/A:1025014929212
7. Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Multivariate Applications Book Series. Lawrence Erlbaum Associates Publishers.
8. Fox, J. P. (2010). Bayesian item response modeling: Theory and applications. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-0742-4
9. Frazier, T. W., Ratliff, K. R., Gruber, C., Zhang, Y., Law, P. A., & Constantino, J. N. (2014). Confirmatory factor analytic structure and measurement invariance of quantitative autistic traits measured by the Social Responsiveness Scale-2. Autism, 18(1), 31–44. https://doi.org/10.1177/1362361313500382
10. Hambleton, R. K. (2000). Response to commentaries: Item response theory—Thirty years later. Applied Psychological Measurement, 24(1), 33–35. https://doi.org/10.1177/01466216000241001
11. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications, 2.
12. Huerta, M., & Lord, C. (2012). Diagnostic evaluation of autism spectrum disorders. Pediatric Clinics of North America, 59(1), 103–111. https://doi.org/10.1016/j.pcl.2011.10.018
13. Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. The Lancet, 383(9920), 896–910. https://doi.org/10.1016/S0140-6736(13)61539-1
14. Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 30(3), 205–223. https://doi.org/10.1023/A:1005592401947
15. Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1-4. Western Psychological Services.
16. Lord, F. M. (2012). Applications of item response theory to practical testing problems. Routledge.
17. Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., Christensen, D. L., Wiggins, L. D., Pettygrove, S., Andrews, J. G., & others. (2020). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016. MMWR Surveillance Summaries, 69(4), 1–12. https://doi.org/10.15585/mmwr.ss6904a1
18. Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science, 14(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x
19. Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17(5), 1–25. https://doi.org/10.18637/jss.v017.i05
20. Robins, D. L., Casagrande, K., Barton, M., Chen, C. M., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F). Pediatrics, 133(1), 37–45. https://doi.org/10.1542/peds.2013-1813
21. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4), 100. https://doi.org/10.1007/BF03372160
22. Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34(4), 273–286. https://doi.org/10.1037/h0070288
23. Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3