Autism Spectrum Disorder Detection using Deep Learning Techniques

Shailaja. K P[1], *Shilpa.* S[1], *Tharun.* M A[1*], and *Sachin.* K S[1]

[1]Department of Computer Applications, B.M.S. College of Engineering, Bengaluru, India

Abstract. In recent years autism spectrum disorder (ASD) has been rapidly increasing in people irrespective of age which effects social interactions, communication and behaviour in daily life. Early diagnosis is essential for effective intervention; however, existing diagnostic procedures rely heavily on behavioural assessments, which are subjective, time-consuming, and inaccessible in under-resourced settings. In this research deep learning–based diagnostic framework using facial imagery is proposed to classify ASD and non-ASD cases. The system combines two pretrained convolutional neural networks, VGG16 and InceptionV3, into a weighted soft-voting ensemble. To address dataset imbalance, SMOTE is applied during training, and threshold tuning is used to optimize classification boundaries based on F1-score. The proposed model achieved 97.11% accuracy and AUC of 0.99 on a curated facial image dataset. Additionally, Grad-CAM is used to enhance interpretability by highlighting salient facial regions. The pipeline is fully reproducible via Google Collab with GPU support. The results confirm the ensemble's effectiveness as a scalable, non-invasive screening tool for early ASD detection.

1 Introduction

Over the past few years ASD impacting numerous people across many countries in the world such as socialising, communication and behaviour. With a global prevalence now estimated at 1 in 36 children, early diagnosis is crucial for initiating timely and effective intervention strategies. However, conventional ASD diagnostic protocols remain reliant on observational assessments and caregiver reports, making them time-consuming, subjective, and inaccessible in low-resource environments.

To address these limitations, deep learning (DL) and computer vision methods—particularly those involving facial image analysis—have been explored as non-invasive and scalable screening tools. Several studies have validated the potential of convolutional neural networks (CNNs) in identifying facial markers associated with ASD phenotypes. In [1], Varsha and Maheswari achieved 96.66% accuracy using DenseNet121 on a Kaggle-based dataset. In [2], Sajeev et al. trained a custom CNN on over 2,900 facial images, achieving 91% accuracy. In [3], Jindal et al. applied an ensemble of VGG16, ResNet50, and MobileNet, reporting improved F1-scores. In [4], Arockia Dass et al. proposed hybrid CNN frameworks, although their work lacked interpretability.

Despite these advances, major challenges remain:

- Lack of interpretability in predictions (e.g., absence of visual explanations),
- Poor handling of class imbalance,
- Use of fixed decision thresholds without optimization,
- High computational complexity that limits clinical usability.

This study addresses these gaps by introducing an ensemble-based ASD classification framework that integrates VGG16 and InceptionV3 with a weighted soft-voting mechanism. To further enhance robustness and clinical applicability, the system includes SMOTE-based class balancing, threshold tuning for optimal F1-score, and Grad-CAM for interpretable visualization.

Methodology of this research as follows:

1. implementation of a dual-CNN ensemble (VGG16 + InceptionV3) weighted by validation accuracy.
2. Use of SMOTE to balance training data and improve minority class sensitivity.
3. Application of threshold tuning to improve classification boundary precision.
4. Use of Grad-CAM to highlight critical facial features influencing model output.
5. Cloud-based deployment using Google Colab with GPU acceleration for scalable, reproducible results.

1.1 Organization of the paper

Section II discusses recent literature in ASD detection. Section III contains complete proposed methodology architecture. Section IV presents the methodology and training configuration. Section V discusses the experimental results and comparative evaluation.

* Corresponding author: tharunma.mca23@bmsce.ac.in

Section VI concludes the study with key findings, limitations, and future scope.

2 Literature review

The increasing prevalence of Autism Spectrum Disorder (ASD) has accelerated research into automated diagnostic tools using deep learning (DL) and computer vision. Recent studies from 2021 to 2025 have primarily focused on facial image–based classification using convolutional neural networks (CNNs), ensemble architectures, and explainable AI (XAI) frameworks to enhance robustness, accuracy, and clinical trust.

Varsha and Maheswari [1] implemented three pretrained CNNs—VGG19, DenseNet121, and InceptionV3—on a publicly available ASD image dataset and achieved the highest accuracy (96.66%) using DenseNet121. However, their model lacked ensemble learning and interpretability mechanisms. Similarly, Sajeev et al. [2] proposed a custom CNN trained on 2,900 facial images and achieved 91% accuracy, though it lacked visual explanation outputs, reducing clinical transparency.

Jindal *et al.* [3] explored multiple CNNs, including VGG16, VGG19, and MobileNetV2, and demonstrated that ensemble learning enhances classification performance, particularly in F1-score and recall. Arockia Dass *et al.* [4] demonstrated VGG16 and DenseNet for binary ASD classification but did not integrate class balancing techniques or explainability features.

Other recent approaches focus on novel architectures. Paayas et al. [5] introduced a Hybrid Classical–Quantum Neural Network (HCQNN), reporting improved accuracy through quantum entanglement, although the model remains computationally intensive and less reproducible. Alam et al. [6] used transfer learning along with Grad-CAM for interpretability and achieved 95% accuracy. Rahman and Subashini [7] presented a deep network with class balancing and hyperparameter tuning, obtaining a 94% diagnostic accuracy without Grad-CAM.

Comparative CNN studies by Alsaade and Alzahrani [8] showed Xception outperformed NASNet and VGG19, achieving 91% accuracy. Zheng and Liu [9] performed rapid screening using VGG16 and ResNet18 on a Kaggle dataset but did not incorporate visualization. In contrast, Garg et al. [10] presented a Grad-CAM–based explainable CNN that exceeded 95% accuracy while supporting clinical interpretability.

Petricia *et al.* [12] introduced a VGG16–ResNet transfer learning model with moderate success, whereas Jain and Mangal [13] developed a webcam-based eye-tracking CNN system achieving 91% accuracy, though it lacked visualization and fine-grained interpretability.

Ding *et al.* [14] demonstrated a systematic review and meta-analysis of deep learning–based approaches for ASD detection across multiple modalities, including facial imagery and neuroimaging. Analyzing data from 11 studies covering 9,495 ASD cases, they reported a pooled sensitivity of 95%, specificity of 93%, and an AUC of 0.98. While these findings confirm the high potential of deep learning in ASD classification, the authors noted considerable heterogeneity across datasets and models. The present work addresses some of these gaps by standardizing preprocessing, incorporating SMOTE-based class balancing, and integrating Grad-CAM interpretability within an optimized ensemble framework.

Alkahtani *et al.* [15] employed transfer learning on facial landmark data using MobileNetV2 and hybrid VGG19 models, achieving 92% accuracy on the Kaggle ASD dataset. Their work underscores the utility of lightweight CNNs for ASD screening. Building on this, The proposed ensemble leverages SMOTE-based class balancing, decision threshold tuning, and Grad-CAM explainability strategies to enhance both prediction accuracy and interpretability.

While these studies demonstrate that CNNs are effective for ASD detection, they often overlook key aspects such as class imbalance mitigation, ensemble optimization, or model explainability. In response, this study introduces a novel ensemble-based framework combining VGG16 and InceptionV3 with SMOTE-based data balancing, threshold tuning for improved sensitivity, and Grad-CAM visualizations. The system is trained and evaluated in a cloud-based environment, ensuring scalability and reproducibility for real-world deployment.

3 Proposed methodology

In this section proposed methodology consists of modular deep learning pipeline designed for the Autism Spectrum Disorder (ASD) detection using facial imagery. It incorporates six sequential components: (1) input preprocessing, (2) SMOTE-based class balancing, (3) dual CNN-based feature extraction using VGG16 and InceptionV3, (4) weighted soft-voting ensemble fusion, (5) threshold tuning to refine classification sensitivity, and (6) Grad-CAM visualization to improve model interpretability. The complete architecture is cloud-executable and optimized for scalability using Google Colab with T4 GPU acceleration.

After that Data augmentation is applied to enhance generalization and resizing input images into $128 \times 128 \times 3$. Later normalization technique is applied to scale all the numerical features into [0,1] The preprocessed image is simultaneously passed through both pretrained CNNs, each producing softmax probability scores. These outputs are merged using a weighted soft-voting mechanism, with weights derived from validation accuracy.

For handling class imbalance Synthetic Minority Oversampling Technique (SMOTE) is performed after feature extraction for synthetically balance the minority (autistic) class. The final ensemble prediction is passed through a threshold tuning module that scans different probability cutoffs and selects the optimal threshold ($t^* = 0.48$) that maximizes the F1-score on the validation set.

Finally, for final convolutional layers of both models to generate heatmaps that highlight critical facial regions influencing classification Grad-CAM

(Gradient-weighted Class Activation Mapping) is applied. This enhances interpretability and enables human-in-the-loop review, particularly useful in clinical and educational screening scenarios.

The overall system workflow is illustrated in Figure 1, which shows the end-to-end path from raw input image to final ASD classification decision through ensemble fusion and visualization.

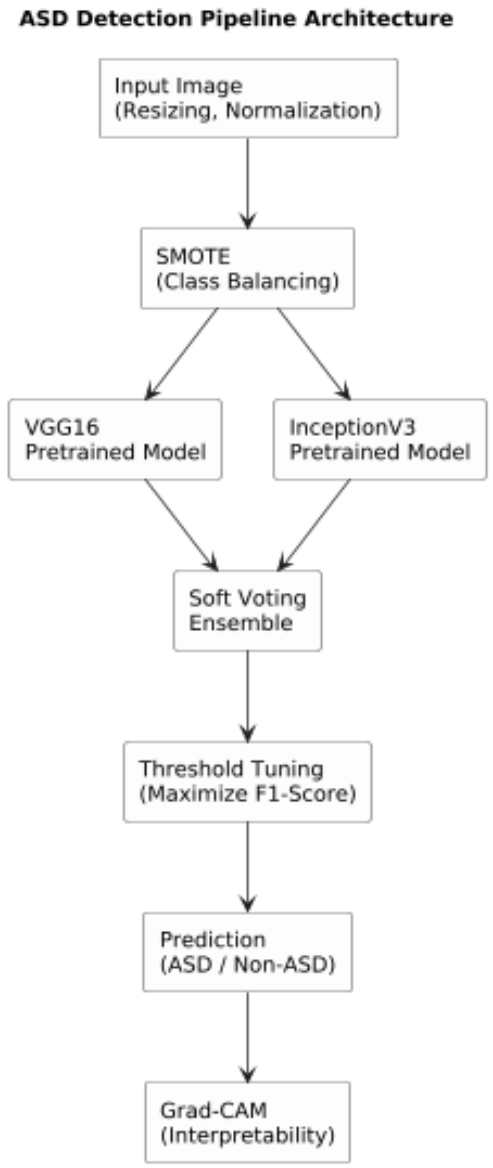

Fig. 1. Block diagram of the proposed ensemble-based ASD classification framework.

4 Methodology

This section presents the dataset preparation, preprocessing, CNN model configurations, ensemble fusion strategy, training parameters, and interpretability methods used in the proposed ASD detection system.

4.1 Dataset description

The dataset consists of 5,880 labeled facial images of children categorized as either Autistic or Non-Autistic. It was sourced from the publicly available Kaggle dataset titled “Autism Image Data” [11], uploaded by user cihan063. The dataset is organized into four subsets: training (2,540 images), validation (1,100), testing (900), and consolidated (1,340), each containing two class folders: Autistic and Non_Autistic. Only .jpg, .jpeg, or .png image formats were included.

To simulate real-world variability, the dataset was manually curated to include diverse lighting, facial orientations, and backgrounds. All images were resized to 128×128 pixels with 3 RGB channels for compatibility with CNN input layers. Class labels were encoded numerically: 0 for Non-Autistic and 1 for Autistic.

4.2 Data preprocessing

To maintain class distribution across splits, stratified sampling was applied. The following preprocessing steps were used:

- Resizing to 128×128×3 format
- Pixel normalization to the [0, 1] range
- Label encoding (0 = Non-Autistic, 1 = Autistic)
- Data augmentation: random rotation, horizontal flip, and zoom (applied during training)

4.3 Transfer learning models

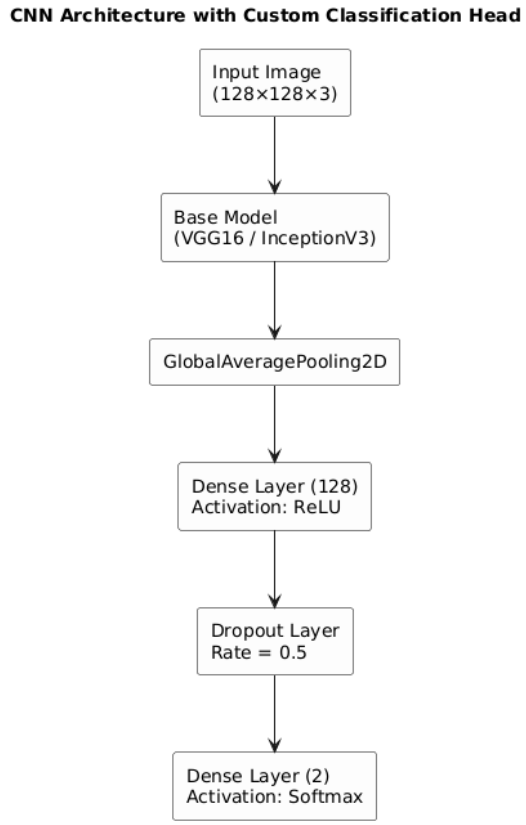

Fig. 2. Custom classification head used for both VGG16 and InceptionV3 models.

Two pretrained CNN architectures were utilized:

- VGG16: Initialized with ImageNet weights. Convolutional layers were frozen. The custom classification head included:

GlobalAveragePooling → Dense(256, ReLU) → Dropout(0.5) → Dense(2, Softmax).

- InceptionV3: Also initialized with ImageNet weights. Top layers were removed and replaced with a similar classification head.

The custom classification head used for both backbones is shown in Figure 2 and both models were trained using TensorFlow-Keras on Google Colab with T4 GPU acceleration.

4.4 SMOTE for class imbalance

In this section Synthetic Minority Oversampling Technique (SMOTE) is applied to handle class imbalance issue at feature level after CNN-based feature extraction. This enhanced the representation of the minority (autistic) class in training batches.

4.5 Ensemble fusion strategy

The outputs from VGG16 and InceptionV3 were combined using a weighted soft-voting ensemble. The final prediction probability. The final prediction score is calculated as:

$$P_{final} = w_1 P_{VGG16} + w_2 P_{inception\ v3} \quad (1)$$

As shown in Equation (1), the ensemble probability is computed as a weighted sum of the two model outputs.

Where:

- P_{VGG16} and $P_{Inception\ V3}$ are softmax probabilities
- w_1 and w_2 are model weights based on validation accuracy, normalized as:

$$w_i = \frac{a_i}{\sum a_i} \quad (2)$$

Equation (2) defines the normalized weights using the validation accuracies a_{VGG} and a_{inc}.

4.6 Threshold optimization

To improve classification confidence and optimize the decision boundary for imbalanced data, a grid search–based threshold tuning strategy was applied. The default softmax threshold of 0.5 was varied between 0.1 and 0.9 in increments of 0.01. At each step, the F1-score was calculated for evaluation metrics recall and precison. The threshold value was maximized the F1-score on the validation set was selected as the optimal threshold.

The F1-score was computed as:

$$F1 = \frac{2 * precision * Recall}{Precision + Recall} \quad (3)$$

As shown in Equation (3), F1-score denotes harmonic mean for recall and precision . Based on the validation results, the optimal threshold $t^* = 0.48$was selected. This dynamic threshold significantly improved model sensitivity without sacrificing overall accuracy or precision.

4.7 Model interpretability using Grad-CAM

To promote explainability and clinician trust, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to both VGG16 and InceptionV3. Grad-CAM highlights the most influential facial regions that contributed to the final classification decision by generating heatmaps from the final convolutional layers. These visualizations were overlaid on the original input images and consistently focused on medically relevant zones such as the eyes, nose, and facial symmetry. This interpretability feature enables human-in-the-loop validation and supports practical use in real-world screening settings.

4.8 Training configuration

All training experiments were conducted using Google Colab with T4 GPU acceleration to ensure reproducibility. In this every model was trained using adam optimizer to 15 epoches and evaluated using standard classification metrics. Table 1 summarizes the training setup.

Table 1. Training configuration.

Parameter	Values
Batch Size	32
Epochs	15
Optimizer	Adam (learning rate = 0.00005 for VGG16 & 0.0002 for InceptionV3)
Loss Function	Categorical Cross-Entropy
Metrics	AUC, Precison , F1 Score, Recall and accuracy
Platform	Google Colab (T4 GPU)

5 Experimented results and discusssion

In this section experimented results with performance analysis and comparative analysis of proposed ASD detection system. Performance metrics of individual models (VGG16 and InceptionV3) and their weighted soft-voting ensemble are reported. Metrics include accuracy, precision, recall, F1-score, and ROC-AUC, all evaluated on the test set. Additionally, interpretability is demonstrated using Grad-CAM visualizations.

5.1 Training and validation trends

From Figures 3 and 4 it is clearly illustrated that training and validation performance of the VGG16 and InceptionV3 models, respectively. Both models exhibit a consistent increase in training accuracy and a gradual decline in training loss. After that Validation accuracy followed by training curve with minor fluctuations, suggesting stable learning and minimal overfitting. InceptionV3 showed slightly more variance mid-training but achieved smooth convergence by the final epochs.

The model shows consistent accuracy and a steady decrease loss, indicating effective learning and generalization.

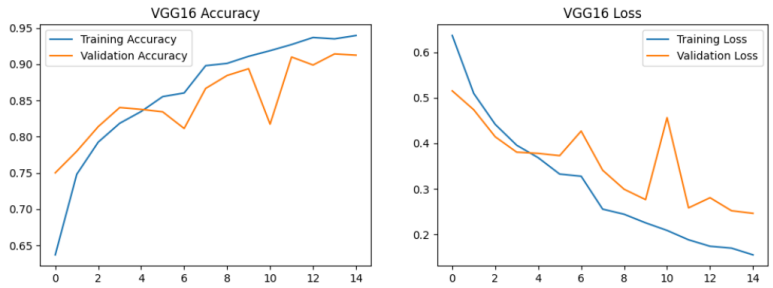

Fig. 3. Training and validation accuracy/loss curves for VGG16.

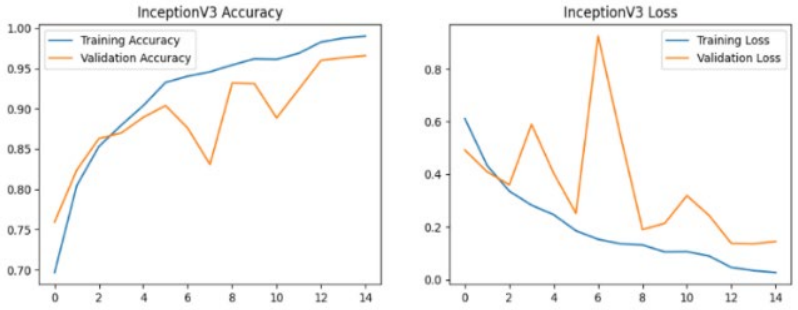

Fig. 4. Training and validation accuracy/loss curves for InceptionV3.

Despite minor fluctuations, the validation metrics stabilize toward the final epochs, confirming convergence and robustness.

5.2 Classification report

Table 2. Classification performance of individual and ensemble models.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	91.24%	91.00%	91.00%	91.00%
InceptionV3	96.34%	96.00%	96.00%	96.00%
Ensemble	97.11%	97.00%	97.00%	97.00%

The classification results on the test dataset are summarized in Table 2. The ensemble model outperforms both individual models across all evaluation metrics. It achieved the highest accuracy (97.11%) and F1-score (97.00%). The integration of SMOTE enhanced recall, and threshold tuning (set to 0.48) optimized the classification boundary.

5.3 ROC curve

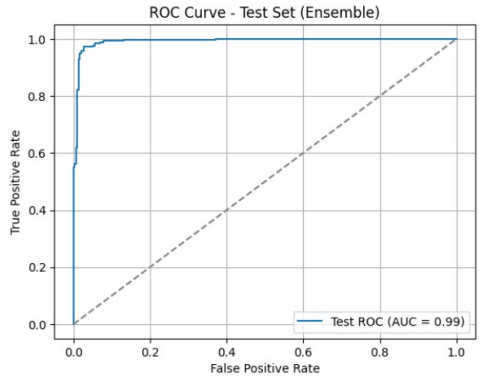

Fig. 5. ROC curve for ensembled model.

To further evaluate the discriminative power of the ensemble model, a (ROC) Receiver Operating Characteristic curve was plotted using the forcasted probabilities. From Figure 5, that AUC reached 0.99, displaying a near-perfect capability of the model to differentiate among ASD and non-ASD cases.

5.4 Overview of Confusion matrix

From Figure 6 it clearly illustrates that confusion matrix for the ensemble model. It correctly classified 298 ASD cases and 273 non-ASD cases. There were only 9 false negatives and 8 false positives, indicating high precision and recall. These values validate the model's clinical relevance for screening applications.

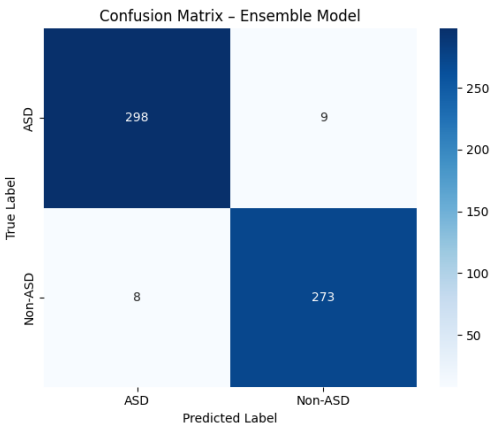

Fig. 6. Confusion matrix for the ensemble model.

5.5 Grad-CAM visualizations

Representative Grad-CAM maps for VGG16 are provided in Figure 7, and for InceptionV3 in Figure 8. These examples highlight attention around clinically relevant facial regions such as the eyes, gaze direction, and overall facial symmetry. This interpretability aids clinicians in validating predictions and supports trust in real-world applications.

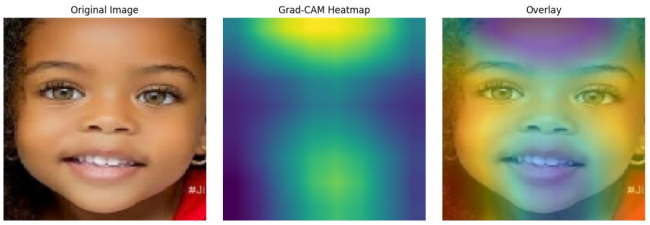

Fig. 7. Grad-CAM visualizations for ASD and non-ASD cases using VGG16

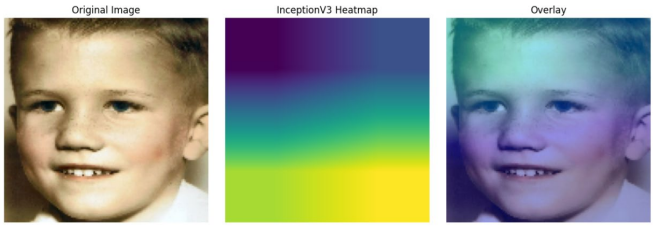

Fig. 8. Grad-CAM visualizations for ASD and non-ASD cases using InceptionV3.

Moreover, visual outputs proves that the models concentrates on medically related facial zones. This enhances transparency and trust for potential clinical applications.

5.6 Comparative analysis with existing work

Table 3. Comparative analysis with recent ASD detection models.

Study	Model	Accuracy	Explainability	Year
Varsha & Maheswari [1]	DenseNet121	96.66%	No	2024
Sajeev et al. [2]	Custom CNN	91.00%	No	2022
Jindal et al. [3]	VGG16 + ResNet + MobileNet (Ensemble)	96.42%	No	2023
Alam et al. [6]	VGG19 + Grad-CAM	95.00%	Yes (Grad-CAM)	2022
Jain & Mangal [13]	CNN + ResNet	91.00%	No	2024
Petricia et al. [12]	VGG16 + ResNet (Transfer Learning CNN)	91.67%	No	2025
This Work	VGG16 + InceptionV3 + Grad-CAM	97.11%	Yes (Grad-CAM)	2025

Table 3 compares the proposed method with recent ASD detection studies (2022–2025). While prior work reported strong results, most lacked visual interpretability or ensemble optimization. The proposed model combines high accuracy with explainability and reproducibility, offering significant advancement.

5.7 Computational environment

All models were implemented using TensorFlow and Keras on Google Colab with an T4 GPU. Additional experiments were also tested on a local system with an NVIDIA RTX 3080 GPU and AMD Ryzen 7 5800H processor with 16 GB RAM. Each model required approximately 3–4 minutes per epoch for convergence.

5.8 Limitations

While the proposed ensemble model achieved high classification accuracy and interpretability, several limitations remain. First, the dataset—although publicly available and diverse—is limited in demographic and geographic representation, which may reduce the model's generalizability. Second, the use of static facial images does not capture behavioral or temporal features critical for comprehensive ASD diagnosis. Third, Grad-CAM, while helpful for visualization, may not highlight all diagnostically relevant features with sufficient granularity. Lastly, the model was evaluated on a single internal dataset without external testing on independent sources. Future research will address these limitations by incorporating multi-modal data (e.g., video or speech), expanding to more diverse populations, and developing lightweight models suitable for deployment in resource-constrained environments.

6 Conclusion

This study presented a robust ensemble-based deep learning framework for automatic detection of ASD using facial image analysis. By integrating VGG16 and InceptionV3 through a weighted soft-voting strategy, along with SMOTE-based class balancing and threshold tuning, the model achieved a classification accuracy of 97.11% on the test set. The use of Grad-CAM visualizations provided critical interpretability, making the system more transparent and clinically applicable. Experimental results confirmed the superiority of the

ensemble over individual models, while also offering practical deployment potential through cloud-based implementation. Despite strong results, the study is limited by dataset diversity, lack of behavioral features, and reliance on static imagery. Future work will focus on expanding to multi-modal datasets, performing external validations, and optimizing the model for real-time, resource-constrained clinical environments.

References

1. V. Varsha, R. Maheswari, Autism Spectrum Disorder Classification Using Deep Learning Models on Facial Images, In Proceedings of the International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, India, (2022), 215-220
2. M. Sajeev, P. Rao, K. Rajesh, CNN-based model for autism detection using facial images. Int. J. Adv. Comput. Sci. Appl. **13**, 124-130 (2022).
3. R. Jindal, N. Sharma, S. Mehta, Ensemble Learning for ASD Classification Using CNN Architectures, In Proceedings of the International Conference on Emerging Trends in Intelligent Systems and Network Technologies (ETISNT), New Delhi, India, (2023), 140-146
4. J.M.A. Dass, A. Girisha, Autism Spectrum Disorder Detection Using Face Features Based on Deep Neural Network, 2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET), IEEE, Gautam Buddha Nagar, India, May 13 (2024), 1008-1012
5. S. Paayas, A. Mohan, P. Senthilkumar, Hybrid classical–quantum neural network for autism classification. Procedia Comput. Sci. **218**, 45–50 (2023)
6. M. S. Alam, M.M. Rashid, R. Roy, A.R. Faizabadi, K.D. Gupta, M.M. Ahsan, Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach. Bioeng. **9**, 710 (2022). https://doi.org/10.3390/bioengineering9110710
7. T. Rahman, R. Subashini, Facial feature–based ASD classification using deep learning techniques. Comput. Biol. Med. **150**, 106070 (2023).
8. F. Alsaade, S. Alzahrani, Comparative Analysis of CNN Backbones for Autism Detection. J. Artif. Intell. Health 5, 20-30 (2023).
9. Y. Zheng, C. Liu, Rapid Autism Screening Using Transfer Learning with Facial Imagery, In Proceedings of the International Conference on Computer Vision and Image Processing (CVIP), (2023).
10. S. Garg, M. Bansal, Explainable deep learning framework for autism detection with Grad-CAM. IEEE Access **10**, 89912-89923 (2022).
11. Cihan063, "Autism Image Data," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/cihan063/autism-image-data
12. V. Angel Petricia, J.M. Jacksy, R. Bhavadharani, Autism Spectrum Disorder Detection using Convolutional Neural Network, In Proceedings 2025 the International Conference on Visual Analytics and Data Visualization (ICVADV), IEEE, Tirunelveli, India, April 22 (2025), 743-748
13. R. Jain, D. Mangal, Early Detection of Neurodevelopmental Disorders: Quantifying Autism Behavioral Markers with Computer Vision and Artificial Intelligence, In 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT), IEEE, Bhubaneswar, India, February 06 (2024), 1-6
14. Y. Ding, H. Zhang, T. Qiu, Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis. BMC Psychiatry **24**, 739 (2024). https://doi.org/10.1186/s12888-024-06116-0
15. H. Alkahtani, T.H.H. Aldhyani, M.Y. Alzahrani, Deep learning algorithms to identify autism spectrum disorder in children-based facial landmarks. Appl. Sci. 13, 4855 (2023).