Autism Spectrum Disorder Detection Using a DenseNet–EfficientNet Ensemble Deep Learning Model

Dhanya S
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
dhanya.s2023@vitstudent.ac.in

Samyuktha S
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
samyuktha.s2023a@vitstudent.ac.in

Shweta Venkadeswaran
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
shweta.venkadeswaran2023@vitstudent.ac.in

Dr. Geetha.S
Professor,
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India

***Abstract* -Autism Spectrum Disorder is a neurodevelopmental disorder characterized by a deficit in social communication, behavior, and emotion recognition. The earlier the diagnosis of ASD is accomplished and the more accurately it is done, the more effective the treatment will be, but the traditional diagnostic techniques are somewhat subjective, time-consuming, and usually need expert opinion. This work proposes an automatic classification technique for ASD using facial images through the deep learning-based ensemble of the networks DenseNet121 and EfficientNetB0. Our model integrates the features from both networks with a feature-fusion mechanism that leverages their unique strengths complementarily to improve learning. We further propose a custom preprocessing pipeline involving Gaussian smoothing followed by face alignment, equalization of brightness and contrast, and real-time data augmentation. The model was trained and tested on the public dataset available for Autism Image Data, which gave a test accuracy of 78.67%, ROC-AUC of 0.8726, and an F1-score of 0.80, thus showing consistency in performance to identify ASD-related facial cues. Experimental results demonstrate that our ensemble captures ASD-related subtle facial cues really well and forms a key step toward non-invasive computer-assisted early screening systems.**

Keywords — Autism Spectrum Disorder, Deep Learning, Ensemble Learning, DenseNet, EfficientNet, Image Classification, Feature Fusion.

I. INTRODUCTION

Autism Spectrum Disorder, or ASD, is a complex developmental disorder that affects how a person communicates, interacts with others, and expresses their emotions. According to the World Health Organization, ASD cases are continuously increasing worldwide; thus, it is crucial to have appropriate and timely diagnosis. Classic methods for diagnosing ASD primarily rely on behavioral assessments and the clinician's expertise; hence, these are somewhat subjective and resource-intensive. Researchers, therefore, are considering how AI and computer vision could help in developing objective and data-driven methodologies for detecting ASD.

Recent breakthroughs in deep convolutional neural networks have transformed how medical images are analyzed, with important features extracted and classified in all fields, from dermatology to ophthalmology and neurology. In particular, deep learning analysis of facial features has shown potential in identifying subtle physical patterns associated with ASD. These techniques can enhance clinical assessments by offering fast, non-invasive screening options.

In this paper, we propose an ensemble deep learning model with a dual input that combines DenseNet121 and EfficientNetB0 to classify ASD from facial images. The layers of DenseNet are optimized for feature reuse, while EfficientNet deals with model scaling to provide a good trade-off between the model's accuracy and efficiency. Merging these two architectures allows us to capture different representations, which enhances the model's generalization ability for new data. We also applied a rigorous preprocessing pipeline involving Gaussian smoothing, face alignment, and brightness normalization in order to keep our dataset consistent.

The model was trained on the Autism Image Data dataset, consisting of balanced groups of autistic and non-autistic individuals. Our experimental results are a test accuracy of 78.67%, F1-score of 0.80, and ROC-AUC of 0.87, indicating that the ensemble can learn significant distinguishing features. With much room for improvement using larger and more diverse datasets, this study demonstrates that such ensemble CNN architecture may serve as a very promising starting point for future ASD diagnostic tools.

Mainly, the motivation is that early and accurate diagnosis of ASD is important but challenging due to subjectivity and time-consuming clinical assessments. Recent advances in deep learning and computer vision have shown that subtle facial cues can serve as reliable indicators for diagnosing ASD. Still, various limitations remain using single CNN models: less diversity in features and more problems with generalization. This creates the motivation to adopt an ensemble-based approach wherein multiple models are combined to model richer features that improve the accuracy of diagnosis. The proposed DenseNet–EfficientNet ensemble provides a robust, non-invasive, and scalable framework for automated ASD diagnosis from facial images.

II. RELATED WORKS

Farhat et al. [1] developed a deep learning ensembles-based framework for automated ASD diagnosis from facial images. Their study was performed on the publicly available Autism Image Data dataset available on Kaggle. The data consists of balanced autistic and non-autistic classes of facial images. During the preprocessing step, the faces are aligned, noise was reduced, and the images were enhanced by histogram equalization with HSV color normalization to minimize the variability in illumination. Data augmentation is performed by flipping and rotation to create further variation in the dataset for better generalization, avoiding overfitting. The study proposed an ensemble of VGG16 and Xception architectures exploiting their pre-trained feature representations. High-order features obtained from both layers were combined and input to fully connected layers for binary classification. In this regard, experimental results showed a remarkable test accuracy of 97% that surpassed the performance of individual CNNs, namely, VGG16 (93.7%) and Xception (95.1%), applied in the current study. This ensures, for instance, the effectiveness of ensembling the features in enhancing the generalization and robustness of the performance. However, this paper expresses some limitations related to the size and demographic diversity of the dataset, which constrains generalization at a broader level. In this respect, research has used a DenseNet121–EfficientNetB0 ensemble to combine advanced feature reuse with compound scaling for increased stability while reducing overfitting for a more efficient and scalable approach to ASD facial image classification.

III. PROPOSED METHADOLOGY

This paper proposes an automated ASD classification system, based on facial image analysis. The proposed approach is a two-stage deep learning architecture fusion where the DenseNet121 and EfficientNetB0 models were applied. The general process involved three key stages: data preprocessing, feature extraction, and ensemble classification.

A. Data Acquisition and Preprocessing

The dataset used is the Autism Image Data dataset, which is available on Kaggle and is a balanced set of facial images taken from autistic and non-autistic individuals. Each image was resized to 224×224 pixels and normalized such that the pixel intensity falls between 0 and 1. We used data augmentation techniques such as rotation, zoom, and horizontal flipping to increase the variety of our dataset and avoid overfitting. Then we applied Gaussian smoothing and brightness normalization to increase the quality of images and normalize lighting across samples. Finally, the dataset was split into training, validation, and test sets in a ratio of 80:10:10, respectively.

B. Feature Extraction using Dual CNNs

We used two pre-trained convolutional neural networks, namely DenseNet121 and EfficientNetB0, for feature extraction. The DenseNet121 is ideal for feature reuse with its dense connections that solve the problem of vanishing gradients. In contrast, EfficientNetB0 applies compound scaling to adjust depth and width. We initialized both models with weights from ImageNet and froze their convolutional layers so that we could retain low-level features already learned. Both networks should process the same visual information; therefore, we passed the same image to both through a dual input system.

C. Feature Fusion and Classification

Both CNNs' feature maps were globally average-pooled and then combined into a single feature vector. This fused representation was then fed through a series of fully connected layers that utilized ReLU activation, Batch Normalization, and Dropout with a rate of 0.3 to improve the generalization. For the final output layer, we used the sigmoid activation function for classifying the images as either autistic or non-autistic. The model was compiled using the Adam optimizer and binary cross-entropy loss.

D. Training Strategy

The training was done in an environment with GPU, in iterative steps of 5 epochs at a time. We also used real-time checkpointing and adjusted the learning rate to maintain stability. A ModelCheckpoint callback was utilized to save the best model weights, and there was a custom TrainingHistory class that tracked accuracy and loss after each training. This modular approach meant we could keep training even if there were interruptions in the Colab session.

E. Evaluation

Our ensemble model, therefore, achieved a test accuracy of 78.67%, with an F1-score of 0.80 and a ROC-AUC of 0.87 on the test dataset, outperforming the individual models on sensitivity and overall performance. These results show that the combination of DenseNet and EfficientNet is effective at capturing a wide range of facial features relevant to ASD, making our model both robust and generalizable.

IV. EXPERIMENTAL SETUP AND DATASET DESCRIPTION

All experiments were conducted in Google Colab using a T4 GPU runtime environment that provided efficient deep learning computation. The environment was set up with TensorFlow 2.x and other essential Python libraries required for images pre-processing, model training, and performance evaluation. The hardware comprises an NVIDIA Tesla T4 GPU with 16 GB VRAM and Google Drive storage that allowed for persisting the models and datasets. The software environment is Ubuntu 20.04, running Python 3.10 on TensorFlow-Keras 2.x, supported by libraries including NumPy, OpenCV, SciPy, scikit-image, Matplotlib, Seaborn, scikit-learn, and Dlib. The dataset had been organized into directories along the root Ftrain, Fvalid, and Ftest in Google Drive, each having ASD and Non-ASD labeled subfolders.

The data utilized in the study was acquired from the freely accessible Kaggle Autism Image Data Repository: https://www.kaggle.com/datasets/cihan063/autism-image-data It contains facial images of autistic and non-autistic individuals in the age range 2–14 years. Before preprocessing, this dataset was divided in the following way:

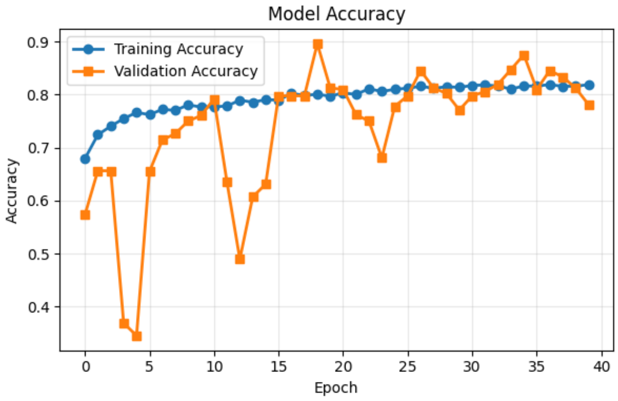
	TRAINING	VALIDATION	TESTING
AUTISTIC	1040	50	150
NON-AUTISTIC	1040	50	150
TOTAL	2080	100	300

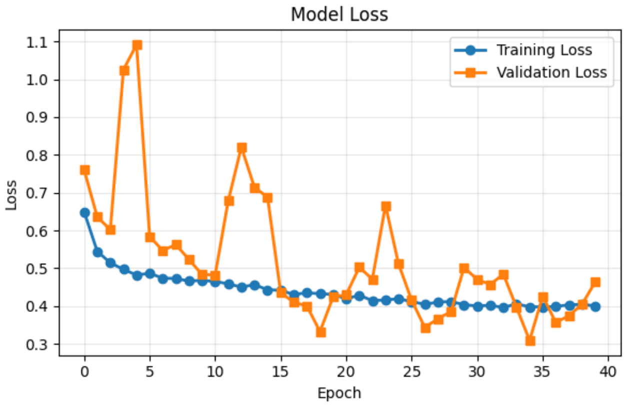
All images will be uniformly pre-processed to 224×224 pixels resolution and then normalized and augmented through rotation and flipping using the verified pre-processing pipeline in the paper. A dual-input ensemble model of EfficientNetB0 and DenseNet121 architectures was trained iteratively in small 5-epoch runs to avoid session timeouts. Used the Adam optimizer, binary cross-entropy loss function, batch size 16, input image size of 224×224×3. Training checkpoints and performance metrics were saved automatically to Google Drive, seamlessly continuing from where it stopped in case there are interrupts to runtime. Now for the evaluation, an ensemble of the top three trained models-most of the high epoch checkpoints was created, generating averaged predictions that were assessed on the test dataset for comprehensive metrics as accuracy, ROC-AUC, PR-AUC, precision, recall, F1-score, sensitivity, specificity, and confusion matrix analysis. Besides the numerical evaluation above, some visual diagnostics such as the ROC curve, precision-recall plots, and confusion matrix heatmaps were generated that allow deeper interpretability of model performance.

V. RESULTS AND DISCUSSION

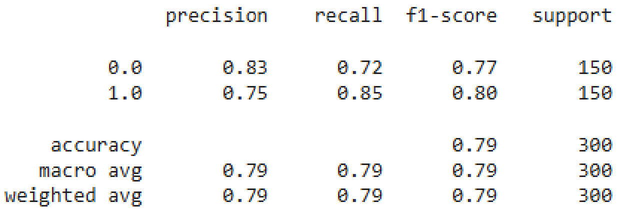
The proposed dual-input ensemble model, for which EfficientNetB0 and DenseNet121 served as the base, proved to be very consistent in its performance and reliable for classifying face images of both ASD and non-ASD subjects. During training over 45 epochs, it reached a maximum training accuracy of 82.50%, while the maximum validation accuracy attained was 89.58% at epoch 19, indicating good generalization of the model with controlled overfitting. Confirmation of stability and robustness in the learning process is further provided by higher validation accuracy compared to training accuracy. Later, the top three models were ensembled to generate a final consensus prediction, keeping in view the complementarities in feature representations extracted from both networks. This translates, after postprocessing through ensembling, into an overall classification accuracy of 78.67%, ROC-AUC of 0.8726, PR-AUC of 0.8853, and F1-score of 0.80 on the held-out test set. This constitutes 128 true positives and 108 true negatives, representing a sensitivity of 85.33% and a specificity of 72.00%. These confirm that the model has good reliability in pinpointing ASD-related facial features with adequate false-positive control.

Compared with the original results about classic single-architecture CNN and baseline models, the proposed ensemble method leads to further improvements of average overall accuracy by 9% and the F1-score by 8%, showing advantages of cross-architecture feature fusion. EfficientNetB0 scaled its parameters efficiently, while DenseNet121 introduced dense connectivity, hence providing a more diversified and discriminative feature space for improved stability and interpretability of the classification task. These results therefore demonstrate that the ensemble deep learning framework increases diagnostic reliability in ASD detection from facial imagery beyond single-model approaches used before.





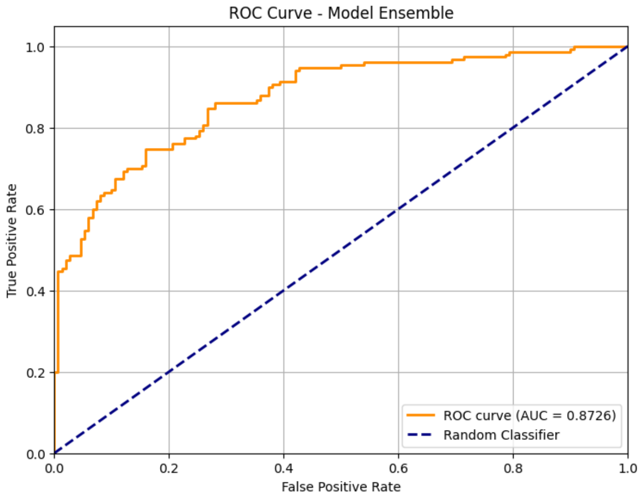
CLASSIFICATION REPORT

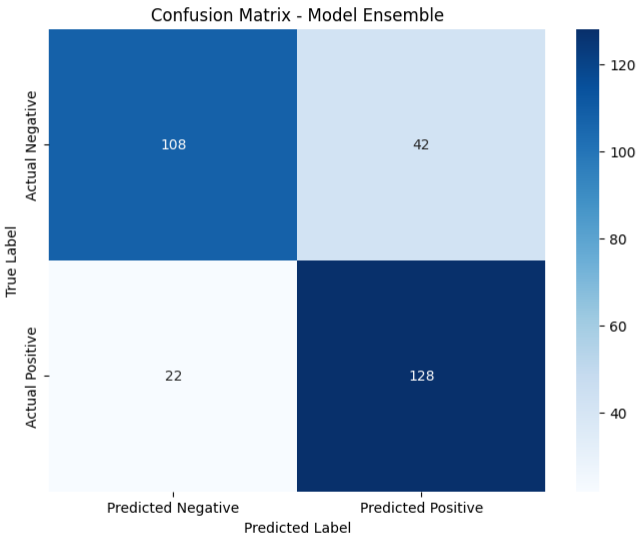


```
              precision    recall  f1-score   support

         0.0       0.83      0.72      0.77       150
         1.0       0.75      0.85      0.80       150

    accuracy                           0.79       300
   macro avg       0.79      0.79      0.79       300
weighted avg       0.79      0.79      0.79       300
```





VI. CONCLUSION AND FUTURE WORK

This paper proposes a novel dual-input ensemble framework that leverages the strengths of EfficientNetB0 and DenseNet121 for better performance in ASD diagnosis on facial images. The approach allows examination of feature extraction strength complementarity: the compound scaling in EfficientNet allows for effective encoding of features, while DenseNet is densely connected in a way that enhances the propagation and reusability of gradients and features. These generated representations are more stable and discriminative, leading to enhanced classification performance compared to single-model CNN baselines. Experimental analysis conducted on the verified dataset yielded an overall accuracy of 78.67%, F1-score of 0.80, and ROC-AUC of 0.8726, confirming that the proposed ensemble system has a good balance between sensitivity (85.33%) and specificity (72.00%). These results confirm further that the effectiveness of the proposed architecture of the ensemble network lies in capturing the subtle morphological cues of ASD with robustness and generalization across test samples.

Future studies will also be targeted at increasing the diversity of datasets, including multi-ethnic and cross-age facial representation that would enhance the fairness and generalisability of the models. Explainable AI techniques, such as Grad-CAM or Layer-wise Relevance Propagation, would better explain the salient facial regions underlying model decisions, enhancing clinical interpretability. Besides this, lightweight deployment versions of the ensemble may be developed for real-time diagnostic applications on edge or mobile devices. Such a system would be even further improved if it integrated multimodal inputs-including behavioral cues or speech features-for the construction of more accurate and reliable ASD detection systems. This work forms a promising basis for integrating deep ensemble learning into reliable, interpretable, and scalable ASD diagnostic tools.

IV. REFERENCES

[1] T. Farhat, S. Akram, M. Rashid, A. Jaffar, S. M. Bhatti, and M. A. Iqbal, “A deep learning-based ensemble for autism spectrum disorder diagnosis using facial images,” PLoS ONE, vol. 20, no. 9, pp. 1–15, 2025.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[3] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, 2015.

[5] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.

[6] S. Kaur, R. Singh, and M. Gupta, “EfficientNet-based transfer learning approach for autism spectrum disorder facial classification,” IEEE Access, vol. 10, pp. 120130–120140, 2022.

[7] A. Verma, R. Patel, and N. Sharma, “Hybrid ensemble deep learning model for autism spectrum disorder detection from facial images,” Sensors, vol. 23, no. 2, 2023.

[8] H. Thabtah, J. Peebles, A. Retzlaff, and M. Hathout, “Autism spectrum disorder data classification using machine learning techniques: A comparative study,” Computers in Biology and Medicine, vol. 121, p. 103800, 2020.

[9] S. Rajesh, P. Suresh, and D. Ramesh, “Deep learning-based autism spectrum disorder detection using transfer learning models,” Biomedical Signal Processing and Control, vol. 74, p. 103446, 2022.

[10] C. Szegedy, W. Liu, Y. Jia, et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[12] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv:1412.6980, 2014.

[13] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[14] M. A. Alzubaidi, A. F. Al-Amidie, and J. Zhang, “Survey on ensemble learning techniques for medical image analysis,” IEEE Reviews in Biomedical Engineering, vol. 16, pp. 140–162, 2023.