ASD-EVNet: An Ensemble Vision Network based on Facial Expression for Autism Spectrum Disorder Recognition

Assil Jaby
Bahcesehir University
assiljaby@gmail.com

Md Baharul Islam
American University of Malta
e-mail address2

Md Atiqur Rahman Ahad
University of East London
e-mail address3

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects individuals' social interaction, communication, and behavior. Early diagnosis and intervention are critical for the well-being and development of children with ASD. Available methods for diagnosing ASD are unpredictable (or with limited accuracy) or require significant time and resources. We aim to enhance the precision of ASD diagnosis by utilizing facial expressions, a readily accessible and limited time-consuming approach. This paper presents ASD Ensemble Vision Network (ASD-EVNet) for recognizing ASD based on facial expressions. The model utilizes three Vision Transformer (ViT) architectures, pre-trained on imageNet-21K and fine-tuned on the ASD dataset. We also develop an extensive collection of facial expression-based ASD dataset for children (FADC). The ensemble learning model was then created by combining the predictions of the three ViT models and feeding it to a classifier. Our experiments demonstrate that the proposed ensemble learning model outperforms and achieves state-of-the-art results in detecting ASD based on facial expressions.

1 Introduction

The computer-assisted diagnosis uses computer-based techniques to analyze data and provide diagnostic information. It is becoming increasingly popular because of providing a rapid and accurate diagnostic process. Following its emergence, it has shown its efficacy in diagnosing Autism Spectrum Disorder (ASD).

However, deep learning techniques for pediatric ASD analysis have made only slight progress due to ASD being a diverse neuro-developmental condition with complicated cognitive traits. Consequently, gathering ASD patients' data is considerably arduous. One in every 68 children has ASD [1]. Given the widespread presence of ASD cases in children, the need for more sophisticated early diagnosis techniques is significant.

Several studies focused on screening protocols (e.g., ADOS, ADI-R, ASQ, STAT), traditional machine learning techniques, and CNNs for ASD diagnosis [2–8]. However, screening approaches are not deterministic and can only assist general practitioners in identifying prospective ASD cases with compromised accuracies. Recent studies introduced behavioral analysis to address these issues. They tracked the eye gaze of patients over 3 seconds after showing them a wide variety of pictures [9–12]. The idea is that ASD patients tend to show atypical attention to key visual information aspects. Nevertheless, these approaches fail to reach acceptable accuracy. To tackle it, neural network-based methods achieve remarkable accuracy. However, collecting neuroimaging data is time-consuming and requires high-end material and a clinical team [13, 14].

Facial expressions are crucial in social communication and convey emotions and intentions. Individuals with ASD often have difficulties interpreting and producing facial expressions, leading to difficulties in social interactions. Analysis of facial expressions could provide valuable information for the ASD diagnosis. Multiple CNN architectures for ASD children classification are proposed in [16] with 91% accuracy. This exhibits that neural networks can extract key features that separate ASD and neuro-typical children. Cao et al. [17] noted that despite vision Transformers (ViT) requiring a relatively large amount of data, it could outperform CNNs with the help of pre-trained weights. He shows that ViT large (ViT-L) pre-trained on ImageNet-21K can achieve an accuracy of 93.50% on ASD dataset [15]. However, their methods are computationally expensive, and improved performance is required. In this research, we develop a new ASD children dataset based on their facial expressions. Then, we propose ASD-EVNet, a novel ensemble vision network for facial expression-based ASD recognition. Our method provides more accurate and reliable predictions by combining multiple transformer-based models. The summary of our contributions is given below:

- We introduce an ASD-EVNet – a novel Vision Transformer-based ensemble network for ASD recognition. It exploits three ViT architectures: a modified ViT, a modified Swin Transformer, and a lightweight MobileViT.
- We develop a facial expression-based ASD dataset consisting of 8000 images. To the best of our knowledge, this is the largest dataset of its kind.
- An extensive ablation study has been conducted to validate our proposed method for ASD recognition that achieved state-of-the-art performance.

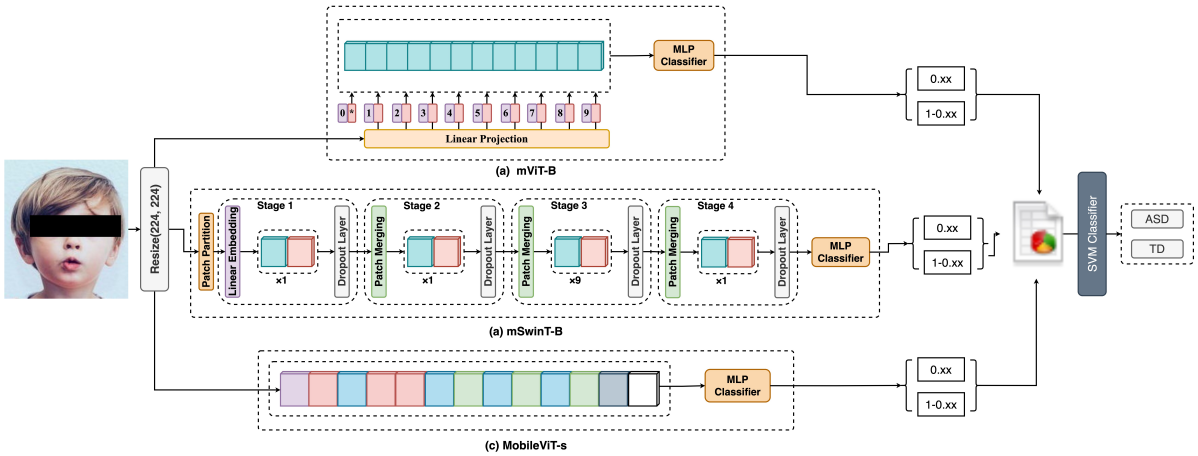


Figure 1: An overview of the proposed ASD-EVNet method. The input image is taken from the ASD dataset [15].

2 Methodology

Our proposed ASD-EVNet network utilizes three Vision Transformer models to recognize ASD based on their facial expressions, as shown in Figure 1. We modify vanilla ViT-B as mViT-B, a SwinTransformer-B as mSwinT-B, and a lightweight MobileVit-s model. These models underwent training and refinement using two facial expression datasets. Instead of relying on traditional voting techniques, the predictions generated by each ViT model are integrated using the Support Vector Machine (SVM).

2.1 Ensemble Vision Network

In a ViT architecture, an image is divided into a set of non-overlapping patches, and each patch is treated as a sequence of tokens. These sequences of tokens are then processed by the Transformer layers, allowing the network to learn complex relationships between the tokens and to make predictions based on the entire image. The final prediction is made based on a pooling operation that aggregates the outputs of the Transformer layers. Then, a one-layer MLP classifier is utilized as the final step in the prediction process after the ViT encoders have processed the input picture. The combination of sequence processing and pooling operations allows ViT models to capture local and global information on the image, making them well-suited for computer vision tasks. The Swin Transformer architecture is similar to the classic ViT architecture that uses a Transformer-based approach to process image data. However, the Swin Transformer introduces the "Swin" operation, which allows the network to attend to a broader range of spatial contexts, resulting in improved performance compared to the classic ViT architecture. It is achieved by using a series of shift-wise operations that allow the network to consider multiple scales of image features. The Swin Transformer and Vision Transformer use a non-linear Gaussian Error Linear Unit (GELU) as the activation function. The lightweight MobileViT is designed for highly efficient and fast while maintaining competitive accuracy. It is optimized for resource-constrained and embedded devices by reducing parameters compared to the classic ViT architecture. Additionally, it leverages to train on smaller datasets.

2.2 Model Selection

Compared to Convolutional Neural Networks (CNNs), ViT models have demonstrated improved performance on computer vision tasks because of their ability to handle image data more globally. In facial expression analysis, understanding the overall pattern of expressions is crucial. Multiple ViT models in the ensemble learning architecture provide the ability to capture a range of features and patterns, allowing for more robust and accurate predictions. Our network's models were selected due to their outstanding performance in image recognition tasks and their ability to capture intricate patterns in facial expressions. Thus, the combination of ViT models allows for a comprehensive analysis of facial expressions, capturing a wide range of patterns that could indicate ASD.

We exploit Support Vector Machine (SVM) in our ensemble learning architectures due to several key advantages. Firstly, it is specifically designed for binary classification problems, making them well-suited for recognizing ASD based on facial expressions. Additionally, it uses a hyperplane to separate the data into two classes, which provides a clear boundary between the classes and makes the decision boundary more robust. The SVM is relatively robust to outliers, significantly working with potentially noisy or unreliable data in a medical diagnosis setting. Furthermore, it can produce non-linear decision boundaries by transforming the input data into a higher-dimensional space, allowing for

the capture of more complex relationships between the variables. These factors make SVM models an attractive choice for the final decision-making step in the proposed ensemble learning architecture. To validate the choice of SVM, we compared its performance to the Random Forest (RF) classifier.

Dropout Layers: In addition to the attention and projection dropout layers, we added an extra dropout layer at the end of each ViT and Swin Transformer block to further enhance the ASD-EVNet ensemble learning architecture. Thus, it improves the model's generalization ability and reduces overfitting problems. Combining these layers could further regularize the model and improve its robustness. The probability of the extra dropout layers was set to 0.1.

Loss Function: We explored the binary cross-entropy (BCE) as a loss function for the binary classification task. It measures the difference between predicted and ground truth labels (y) for each class. The BCE is calculated as,

$$BCE = -(y \log(p) + (1 - y) \log(1 - p)) \quad (1)$$

where, p is the probability of it being the target.

3 Datasets and Experiment

Datasets: The ASD dataset consists of 3,014 facial images of a heterogeneous group of children, divided equally between children with ASD and typically developing (TD) children [15]. Both the test and validation sets consist of 200 instances each. Driven by the lack of available large facial expression-based datasets, we developed a large dataset called 'Face-based ASD Dataset for Children' (FADC)[1]. It has 7921 images, having 3976 and 3945 images of ASD and TD, respectively (with 6337 and 1584 cases for train and test set, respectively) (Figure 2). The TD instances were sourced from 3 datasets [18–20]. The ASD images were collected from ASD patients' videos from social media [21].

Figure 2: Sample instances from our FADC dataset.

Experiment and Evaluation: To find the optimal hyperparameters for our models, we used the Tree-structured Parzen Estimator (TPE) algorithm with optuna [22]. For mViT-B, we experimentally noticed the best learning rate is 0.007 with AdamW optimizer in 67 epochs. For mSwinT-B, the learning rate is 0.002, the same AdamW optimizer in 60 epochs. For MobileViT-s, we used 0.01 learning rate with RMSProp optimizer for 118 epochs. We used an SVC estimator with $C = 1.0$, and RBF kernel. We used a NVIDIA GeForce RTX 2080 GPU to train the ensemble model. We computed the Accuracy and AUROC (Area Under Receiver Operating Characteristic) curve. AUROC evaluates a model's performance across all possible thresholds and is particularly useful for binary classification tasks.

[1]This dataset will be available for researchers. Due to the double-blind review, we did not share the downloadable link here.

4 Results and Discussion

Computational Performance: The computational performance of each module has been shown in Table 1. The mViT-B and mSwinT-B had a longer training time per epoch, averaging 22.12 seconds and 22.83 seconds, respectively. On the other hand, MobileViT-s had a training time of 18.052 seconds per epoch. During testing, the mViT-B, mSwinT-B, and mobileViT-s took 1.15 seconds for 200 image testing (6ms/image), 1.20 seconds to test on 200 images (6ms/image), and 0.784 seconds to test 200 images (4ms/image) respectively (refer to Table 1). Overall, the testing times for all three models were relatively fast, with mobileViT-s being the fastest and the mViT-B being slightly slower.

Table 1: Computational performance of each model.

Models	Training (s)	Testing (s)	Prediction (s)
mViT-B	22.12	1.15	0.006
mSwinT-B	22.83	1.20	0.006
MobileViT-s	18.05	0.78	0.004

Ablation Study: We conduct an ablation study to analyze the contribution of different components to the model's performance (Table 2). Here, each model achieved an accuracy ranging from 92.0% to 94.0% and 97.28% to 98.38% on ASD and FADC datasets, respectively, where the mSwinT-B model outperformed the other models (Table 3). Next, we evaluated the performance of the different combinations of models. We trained each combination of models and computed performances. Our results demonstrated that the combination of mViT-B and mSwinT-B architectures achieved the highest accuracy of 94.50% and 99.08%.

Our mViT-B model has 86 million parameters and achieved impressive accuracy and AUROC. We found that the mSwinT-B backbone with 86 million parameters outperformed the mViT-B. It implies that the mSwinT-B model is highly effective and can perform superior to other backbones. MobileViT-s has impressive performance too, considering its compact size. The advantage of SVM classifier over majority voting is that

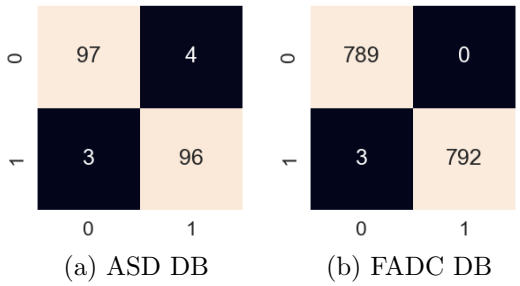
Table 2: Ablation study on ASD-EVNet (results in %). M-Vote stands for majority voting.

							ASD Dataset		FADC Dataset	
mViT-B	**mSwinT-B**	**MobileViT-s**	**M-Vote**	**Ours**	**Classifier**	**Param.**	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**
✓	✗	✓	✗	✓	SVM	91M	94.00	97.05	98.11	99.09
✗	✓	✓	✗	✓	SVM	91M	94.50	97.41	98.58	99.27
✓	✓	✗	✗	✓	SVM	172M	94.50	97.63	99.08	99.68
✓	✓	✓	✓	✗	-	179M	95.00	98.09	99.39	99.78
✓	✓	✓	✗	✓	RF	179M	95.00	98.22	99.53	99.84
✔	✔	✔	✗	✔	**SVM**	179M	**96.50**	**99.04**	**99.81**	**99.91**

Table 3: Results on ASD-EVNet and each component.

		ASD Dataset		FADC Dataset	
Models	**Param**	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**
mViT-B	86M	93.50	96.59	97.54	99.16
mSwinT-B	86M	94.00	97.10	98.38	99.55
MobileViT-s	5M	92.00	96.64	97.28	99.09
ASD-EVNet	179M	**96.50**	**99.04**	**99.81**	**99.91**

it considers each model's confidence as its prediction. It assigns weights to each model's prediction based on its confidence, whereas the majority voting assigns equal weight to each model's prediction. Hence, SVM can effectively leverage the strengths of each model and produce robust prediction than Random Forest (RF), as demonstrated in Table 2.



(a) ASD DB (b) FADC DB

Figure 3: Confusion matrices of test data.

Qualitative Performance: Figure 3a shows only 4 false positives and 3 false negatives. The confusion matrix on our dataset showed no false positives, indicating that our model correctly predicted all positive instances (Figure 3b). We observed that the only incorrect predictions made by our model were in instances where hands, objects, or other obstructions hid part of the face. This suggests that our model may need further improvement to better handle such cases.

Quantitative Performance: For the ASD Dataset, ASD-EVNet ensemble architecture outperformed all the individual models with an accuracy of 96.50% and an AUROC of 99.04% (Table 4). On FADC dataset, we achieved 99.81% accuracy and 99.91% AUROC score, indicating its strong predictive power in distinguishing ASD cases from TD ones. Our model outperformed all other recent methods in both accuracy and AUROC, demonstrating its effectiveness in diagnosing ASD in children based on facial expressions.

Table 4: Comparison of ASD-EVNet with all recent state-of-the-art methods (results in %).

			ASD Dataset		FADC Dataset	
Models	**Backbones**	**Param**	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**
Elshoky et al. [23]	ResNet50	24M	90.50	93.54	94.78	97.69
Hosseini et al. [24]	MobileNetV3	4M	91.00	94.32	95.27	97.98
Cao et al. [17]	ResNet152	60M	91.50	94.58	95.14	97.91
Alsaade and Alzahrani [16]	VGG16	139M	91.50	94.71	61.89	71.88
Mujeeb and Subashin [25]	EfficientNet	21M	91.50	94.89	94.72	96.93
Ahmed et al. [26]	Xception	21M	92.00	95.34	96.47	98.22
Cao et al. [17]	ViT-S	27M	91.00	94.06	-	-
Cao et al. [17]	ViT-M	86M	92.50	95.31	-	-
Cao et al. [17]	ViT-L	307M	93.50	96.62	-	-
ASD-EVNet (Ours)	mViT-B mSwinT-B MobileViT-s	179M	**96.50**	**99.04**	**99.81**	**99.91**

5 Conclusion

This paper presented a novel and effective ensemble architecture, namely ASD-EVNet, for ASD classification from facial images. It consists of three ViT models: mViT-B, mSwinT-B, and MobileViT-s. We compared our model with state-of-the-art models and achieved superior performances in terms of accuracy and Area Under the Receiver Operating Characteristic curve. Apart from the proposed model, we also introduced the largest ASD dataset of this kind. We exploited an added dropout layer along with attention and projection layers at the end of each ViT and Swin Transformer. We found that it is beneficial in improving the performance of the models. It also reduced the overfitting problem and improved the model's generalization capability. The proposed architecture achieved excellent performance on two of the largest datasets on ASD classification of children. Though it is extremely challenging to develop a very large dataset having facial expressions of ASD children, we need to develop a gigantic dataset to explore our method and improve thereby for any constraints in methods.

References

[1] P. O. Towle and P. A. Patrick, "Autism spectrum disorder screening instruments for very young children: a systematic review," *Autism research and treatment*, vol. 2016, 2016.

[2] D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 8, pp. 927–937, 2016.

[3] R. de Belen, T. Bednarz, A. Sowmya, and D. Favero, "Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019," *Translational psychiatry*, vol. 10, no. 1, 2020.

[4] M. Duda, R. Ma, N. Haber, and D. Wall, "Use of machine learning for behavioral distinction of autism and ADHD," *Translational psychiatry*, vol. 6, no. 2, pp. e732–e732, 2016.

[5] F. Hauck and N. Kliewer, "Machine learning for autism diagnostics: applying support vector classification," in *Int'l Conf. Heal. Informatics Med. Syst*, 2017, pp. 120–123.

[6] F. Thabtah, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informatics for Health and Social Care*, vol. 44, no. 3, pp. 278–297, 2019.

[7] B. van den Bekerom, "Using machine learning for detection of autism spectrum disorder," in *Proc. 20th Student Conf. IT*, 2017, pp. 1–7.

[8] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," 2012.

[9] A. Nebout, W. Wei, Z. Liu, L. Huang, and O. Le Meur, "Predicting saliency maps for ASD people," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019, pp. 629–632.

[10] W. Wei, Z. Liu, L. Huang, A. Nebout, and O. Le Meur, "Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019, pp. 621–624.

[11] J. S. Oliveira, F. O. Franco, M. C. Revers, A. F. Silva, J. Portolese, H. Brentani, A. Machado-Lima, and F. L. Nunes, "Computer-aided autism diagnosis based on visual attention models using eye tracking," *Scientific reports*, vol. 11, no. 1, p. 10131, 2021.

[12] S. Liaqat, C. Wu, P. R. Duggirala, S.-c. S. Cheung, C.-N. Chuah, S. Ozonoff, and G. Young, "Predicting ASD diagnosis in children with synthetic and image-based eye gaze data," *Signal Processing: Image Communication*, vol. 94, p. 116198, 2021.

[13] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, and G. Van Wingen, "Classifying autism spectrum disorder using the temporal statistics of resting-state functional MRI data with 3D convolutional neural networks," *Frontiers in psychiatry*, vol. 11, p. 440, 2020.

[14] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong *et al.*, "Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review," *Computers in Biology and Medicine*, vol. 139, p. 104949, 2021.

[15] P. G., "The ASD children dataset." https://drive.google.com/drive/folders/1XQU0pluL0m3TIlXqntano12d68peMb8A, 2021.

[16] F. W. Alsaade and M. S. Alzahrani, "Classification and detection of autism spectrum disorder based on deep learning algorithms," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[17] X. Cao, W. Ye, E. Sizikova, X. Bai, M. Coffee, H. Zeng, and J. Cao, "ViTASD: Robust Vision Transformer Baselines for Autism Spectrum Disorder Facial Diagnosis," *arXiv preprint arXiv:2210.16943*, 2022.

[18] IRON486, "Children vs adults classification dataset." https://www.kaggle.com/datasets/die9origephit/children-vs-adults resource=download, 2022.

[19] . Ali, "Child binary classification dataset." https://www.kaggle.com/datasets/alyusama/child-binary-class, 2020.

[20] E. Mostafa, "Egyptian kids faces dataset." https://www.kaggle.com/datasets/mostafaebrahiem/egyptian-kids-faces, 2022.

[21] S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 755–761.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[23] B. R. G. Elshoky, E. M. Younis, A. A. Ali, and O. A. S. Ibrahim, "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images," *ETRI Journal*, vol. 44, no. 4, pp. 613–623, 2022.

[24] M.-P. Hosseini, M. Beary, A. Hadsell, R. Messersmith, and H. Soltanian-Zadeh, "Deep learning for autism diagnosis and facial analysis in children," *Frontiers in Computational Neuroscience*, p. 119, 2022.

[25] K. Mujeeb Rahman and M. M. Subashini, "Identification of autism in children using static facial features and deep neural networks," *Brain Sciences*, vol. 12, no. 1, p. 94, 2022.

[26] Z. A. Ahmed, T. H. Aldhyani, M. E. Jadhav, M. Y. Alzahrani, M. E. Alzahrani, M. M. Althobaiti, F. Alassery, A. Alshaflut, N. M. Alzahrani, and A. M. Al-Madani, "Facial features detection system to identify children with autism spectrum disorder: deep learning models," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.