Autism Spectrum Disorder Classification via Local and Global Feature Representation of Facial Image

Md. Nadim Mahamood[1†], Md. Zasim Uddin[2†*], Md. Arif Shahriar[3], Fady Alnajjar[4*], Md Atiqur Rahman Ahad[5]

***Abstract*— Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects social communication and interaction. Early diagnosis of ASD can mitigate the severity and help with ideal treatment direction. Computer vision-based methods with traditional machine learning and deep learning are employed in the literature for automatic diagnosis. Recently, deep learning with a facial image-based ASD classification has gained interest due to its ease of collection and non-invasiveness. We observed that the existing approaches utilized either local or global features of facial images to diagnose ASD. However, its important to consider both local and global features to obtain fine-grained details and larger contextual information for accurate detection and classification. This paper proposes a sequencer-based patch-wise Local Feature Extractor along with a Global Feature Extractor. Finally, the features from these modules are aggregated to obtain the final feature for the classification of ASD. Experiments on a publicly available Autism Facial Image Dataset demonstrate that our proposed framework achieves state-of-the-art performance. We achieved accuracy, precision, recall, and F1-score of 94.7%, 94.0%, 95.3%, and 94.6%, respectively.**

***Index Terms*— Autism Spectrum Disorder, ASD, Vision-transformer, classification, LSTM, Local feature extraction, and Global feature extraction.**

I. Introduction

Autism spectrum disorders (ASD) are a diverse group of neuropsychiatric conditions. They are characterized by some degree of difficulty with impairments in social communication, personal interaction, academic functioning, and restricted and repetitive behaviors [1]. Notably, people with ASD may behave, communicate, and learn in ways different from most others. The Autism and Developmental Disabilities Monitoring Network of the Centers for Disease Control and Prevention estimated that about one in 44 children had been identified with ASD in the United States [2], while the prevalence rate of ASD is one out of 100 worldwide [3]. Due to the complexity of the disorder, its challenging to examine the exact cause of ASD [4]. However, a combination of genetic, environmental, and neurological factors [5] may contribute to the development of ASD.

To improve cognitive, social, and language development outcomes, individuals with ASD must receive an early diagnosis and intervention [6]. Diagnosis can be performed both manually and automatically. Manual diagnosis by clinicians entails a combination of standardized tools and behavioral observations that can aid in identifying ASD. The Childhood Autism Rating Scale [7], Autism Spectrum Disorder-Observation for Children [8], and other manual diagnoses are available. This process, however, can be time-consuming and relies heavily on the clinician's expertise and experience, which can lead to inconsistency in diagnosis. Furthermore, in some areas, the availability of trained clinicians may be limited, resulting in delayed diagnosis and treatment.

On the other hand, the researcher employed different computer vision with traditional machine learning (ML) and deep learning (DL)-based techniques to automatically diagnose ASD [9]. For example, features are extracted from magnetic resonance imaging (MRI) [10], eye gaze data [9], body behavior [11], and facial image [12]. Compared with other modalities, facial images that offer details on various facets of facial morphology (including symmetry, shape, and size) are used in ASD classification [13]. It can capture a wealth of information on facial morphology and can be collected relatively easily and non-invasively without subjects' cooperation. This paper will consider detecting and classifying ASD using facial images.

DL with facial image-based detection and classification technique has recently gained increasing attention, and different models have been developed [14]–[16]. The approaches in [14] looked into different Convolution Neural Network (CNN)-based models for facial analysis to detect and classify ASD. They found some existing pre-trained models using extremely large-scale image datasets and fine-tuning the autism facial image datasets to classify ASD and Typically Developed (TD). For example, the MobileNet [15] obtained 94.6% accuracy in classifying ASD and TD on the Autism Facial Image Dataset (AFID) dataset. Besides these, Cao et al. [16] utilized the Vision Transformer model to predict ASD using facial images with the patch-based method. We can observe that all existing methods used either patch-based local or global features. However, its important to consider both local and global features to obtain fine-grained details and larger contextual information for accurate detection and

[1]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh. E-mail: cse1605030brur@gmail.com
[2]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh. E-mail: zasim@brur.ac.bd
[3]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh. E-mail: reyadhasan605@gmail.com
[4]College of Information Technology, United Arab Emirates University, UAE. E-mail: fady.alnajjar@uaeu.ac.ae
[5]Dept. of Computer Science and Digital Technologies, University of East London, UK. E-mail: m.ahad@uel.ac.uk
† Equal contribution.
* Corresponding author.

classification.

In this paper, we propose a deep learning-based model consisting of patch-based local and global feature extractors that extract local and global features of facial images to classify an individual with ASD and TD. The main contributions are summarized as follows:

- An LSTM-based sequencer block is utilized to extract patch-based local features, while CNN is employed to extract global contextual features. The extracted features from these modules are aggregated to produce robust discriminant features for ASD and TD classification.
- The proposed framework has been evaluated on a publicly available Autism Facial Image Dataset. The experimental results exhibit that it can achieve state-of-the-art performance.

II. Related Work

Traditional Machine Learning (ML)-based Approaches: Ganesan et al. [17] utilized VGG-16 for feature extraction, while support vector machine was employed to classify ASD and TD using the extracted feature. They demonstrated that it achieved an accuracy of 90.0% on a publicly available AFID dataset [18]. Again, Del et al. [19] analyzed facial expressions produced by ASD and TD children using traditional machine learning-based methods. A conditional local neural field was used to recognize and track facial landmarks. They conducted the experiment on a small dataset, including five for each ASD and TD group of children, and found that the lower part of the face is more important in distinguishing between ASD and TD.

Deep Learning (DL)-based Approaches: Besides the traditional ML-based approach, the researcher explored deep learning (DL)-based approaches to detect and classify ASD from facial images. For example, Akter et al. [20] proposed a system that employed 17 classifiers, seven of which are deep and previously learned transfer learning-based models, and the rest of which are ML-based models. They evaluated their models on the publicly available AFID dataset. Among the various methods they experimented with, MobileNet-V1 achieved the highest test accuracy, which was 92.1%. Similarly, Alsaade and Alzahrani [21] also explored pre-trained Xception [22], VGG-19, and NASNETMobile models for the ASD classification. They conclude that the Xception achieved the best accuracy at 91.0% on the same dataset. In addition, Lu and Perkowski et al. [23] explored VGG-16 [24] on their own collected dataset East Asian dataset, and they demonstrated that it achieved an accuracy of 95.0%.

Vision transformer-based patch-wise feature extraction is becoming popular in computer vision due to its ability to extract local fine-grained details [25]. In autism research, Cao et al. [16] proposed ViTASD, which considers the Vision Transformer (ViT) [26] as the backbone and added a Gaussian layer for feature extraction and achieved an accuracy of 93.2% and 94.5% on the AFID dataset, by employing the pre-trained model of extreme large-scale dataset ImageNet [27] and AffectNet [28] datasets, respectively.

We can observe that existing studies utilized either patch-based or pre-trained CNN-based models for feature extraction. Therefore, we employed patch-based local and global feature extraction in this paper to classify ASD and TD.

III. Proposed Method

The proposed framework comprises two modules: (i) Local Feature Extractor (**LFE**), which extracts the pattern of the feature locally based on the patch of the image using Bi-directional long short-term memory (BiLSTM), and (ii) Global Feature Extractor (**GFE**) extracts the global features using a conventional CNN-based model. The detailed architecture of the proposed framework is shown in Fig. 1.

A. Local Feature Extractor

In the proposed framework, the LFE is used to determine the linear correlation in the order of the patches. Motivated by the Vision Transformer [26], the input image is divided into several distinct, non-overlapping patches, allowing the network to concentrate on the image's smaller details while enhancing performance and lowering processing costs. Tatsunam et al. [29] proposed a sequencer method where the sequencer architecture used long short-term memory (LSTM) rather than self-attention for sequence modeling. Our proposed LFE module is based on that sequencer method.

Similar to [29], we employed BiLSTM2D and multi-layer perception to build a sequencer block for the LFE module. Here, the BiLSTM2D layer mixes up the spatial information more economically for high-resolution images than the transformer layer [26] and multi-layer perception for channel-fusion, which typically consists of one or more fully connected layers, where each neuron in one layer is connected to every neuron in the adjacent layer [30]. An illustration of a sequencer block is shown in Fig. 2.

The BiLSTM2D comprises two BiLSTM, namely, horizontal and vertical BiLSTM. Each BiLSTM includes two standard LSTMs, one of which processes data in the forward direction while the other in reverse order, to produce a comprehensive representation of the input sequence. The advantage of the BiLSTM is that it can record data from previous and future sequence contexts, enhancing the network's capacity to represent intricate connections and patterns in the input data. The vertical BiLSTM can take the number of tokens in the vertical direction, while the horizontal BiLSTM for the horizontal direction. For input image I, and I_h and $I_v \in \mathbb{R}^{H \times W \times C}$ are the set of sequences respectively for the horizontal and vertical BiLSTM. B_{ver} and B_{hor} $\in \mathbb{R}^{H \times W \times 2D}$ correspond to the output of horizontal and vertical BiLSTM and are concatenated and processed pointwisely in a fully connected layer to produce the final output. Where H and W are the numbers of sequences in the vertical and horizontal direction, respectively, and C and D are the number of channels and hidden dimensions. The mathematical expression of BiLSTM2D is given by:

$$B_{ver} = BiLSTM_{ver}(I_v), \tag{1}$$

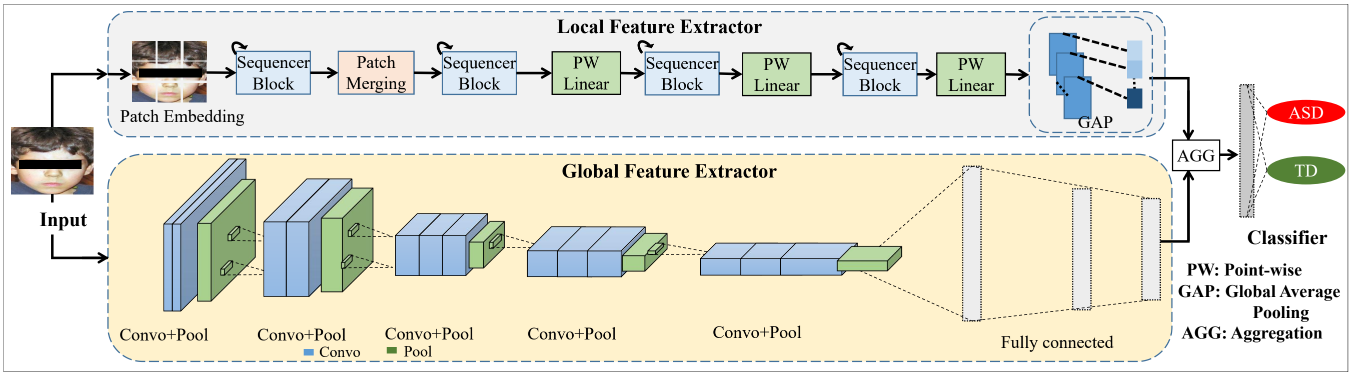

Fig. 1. Proposed framework of the sequencer-based local patch-wise feature extractor, along with the CNN-based global features extractor. Both features are aggregated for the classification of individuals having ASD and TD.

$$B_{hor} = BiLSTM_{hor}(I_h), \tag{2}$$

$$\bar{B} = concat(B_{ver}, B_{hor}), \tag{3}$$

$$Output = FC(\bar{B}), \tag{4}$$

where, $FC(\cdot)$ is fully connected layers with weight $W \in \mathbb{R}^{CX4D}$.

The number of sequencer blocks may generate different versions of the LFE: sequencer-18, sequencer-24, and sequencer-36, where the number of sequencer blocks is 18, 24, and 36, respectively. The output of the last sequencer block is processed by the global average pooling, which decreases the dimensionality of feature maps by computing the average value of each feature map and producing a single scalar value for each channel. For more information about sequencer blocks, please refer to [29].

B. Global Feature Extractor

The GFE refers to the high-level, abstract representations of the image that capture the overall context or information of the entire image. Although other pre-trained CNN-based models for image classification exist in the literature, such as ResNet in [31], MobileNet in [20], and Xception in [21], we chose the modified architecture of VGG-16 [24] as the baseline for the GFE due to its high accuracy. Unlike LSTM, which operates linearly, convolution transfers relationships around its neighbors, which may impact prediction.

In a filter mask, the convolution provides a relationship with its neighbor and moves throughout all spatial points. Then, pooling assists in downsampling the input feature maps' spatial dimensions while retaining key information. It uses 3x3 filters and max pooling to extract high-level features. The architecture consists of 13 convolutional layers and five max-pooling layers. We update the VGG-16 architecture by replacing the last layer with a dimension equal to the final output dimension of the local extractor for pointwise addition.

Finally, the output of the LFE is aggregated with the features extracted from GFE. We have experimented with different combinations to aggregate the features, such as pointwise addition, multiplication, and concatenation. However, pointwise addition provides a better result than other combinations. Later, we employed a fully connected and classifier layer for classification.

IV. Experiments

A. Dataset and Evaluation Metrics

AFID [18] is the only publicly available dataset including facial images for autism research. The images were collected from various websites and Facebook pages with equal distribution of ASD and TD. The dataset contains about 89.0% White ethnic group of children, while the remaining data are from others, with ages ranging from 2 to 14. The gender distribution in the autistic class (male vs. female) was roughly 3:1. The detailed training, validation, and testing protocol is shown in Table I.

TABLE I

About Autism Facial Image Dataset (AFID) [18].

	Type	Train	Validation	Test	Total
Image	ASD	1,268	50	150	1468
	TD	1,268	50	150	1468
Percentage [%]	ASD	43.2	1.7	5.1	50
	TD	43.2	1.7	5.1	50

Several evaluation criteria, including accuracy, precision, recall, and F1-score, are employed to validate the proposed framework.

B. Implementation Details

We conducted all experiments on a single NVIDIA GeForce RTX 2080 Ti GPU running on a Linux operating system. We used Python 3.9.2 [32] and PyTorch 1.10.0 [33] along with timm [34] for all our implementations. We also employed the pre-trained weight using a large-scale ImageNet dataset [1] to fine-tune the model. It has 1000 classes and contains 1,281,167 training images and 50,000 validation images. For regularization and data augmentation, random

[1]https://www.image-net.org/download.php

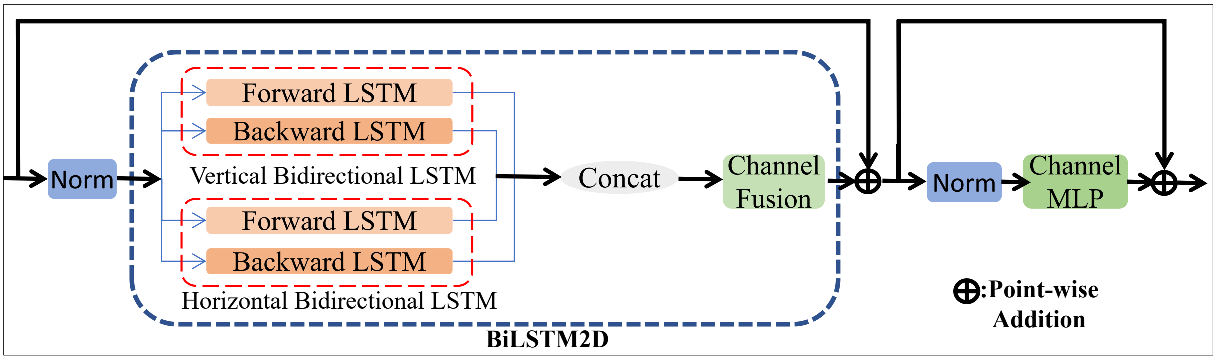

Fig. 2. Core components of the sequencer block. In this block, the BiLSTM2D uses two BiLSTM, which run over horizontal and vertical directions respectively, to extract eatures.

erasing [35], mixup [36], drop-path, smoothing [37], cut-mix [38], label smoothing and different kinds of horizontal and vertical flipping were considered.

Furthermore, we used Cross-Entropy as a loss function and the AdamW as an optimizer with a learning rate ranging from $1\text{x}e^{-3}$ to $1\text{x}e^{-5}$, the momentum of 0.90-0.99, and weight decay of $1\text{x}e^{-5}$. We also used a Droupout rate of 0.2. We trained our model for 310 epochs, where ten were used for cool-down and warm-up.

C. Comparison with State-of-the-Art Approaches

We compare the result of our proposed framework with the latest classification method considering the DL-based along with ViT-based approaches on the publicly available AFID image dataset, including MobileNet in [39], [20], and [14], ViTASD in [16], Xception in [21] and [40] VGG-16 in [41] and ResNet in [31]. The experimental results are shown in Table II, and the corresponding confusion matrix in Fig. 3. It can be observed that the proposed framework achieves the best accuracy compared to state-of-the-art approaches.

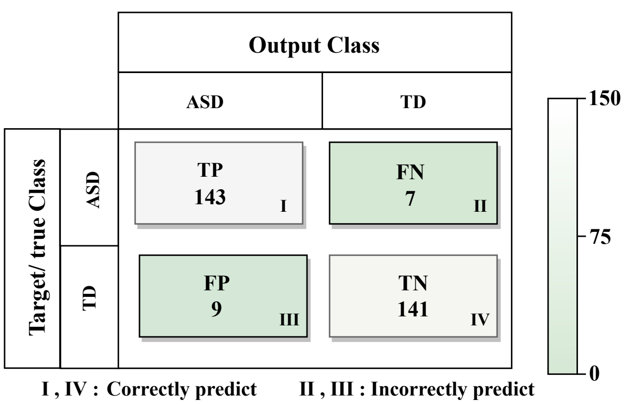

Fig. 3. Confusion matrices for the proposed framework on AFID dataset.

We observed that the proposed framework outperformed previous state-of-the-art approaches. For example, the accuracy based on MobileNet is 92.1% [20] and 94.6% [14], while our proposed framework achieved 94.7%; that is, the proposed framework improves accuracy at 2.6% and 0.1%, respectively. Furthermore, our proposed framework also outperforms the existing Vision Transformer (ViT)-based approach by a great margin. For example, ViTASD [16] achieved 93.2%accuracy with their pre-trained model on the ImageNet dataset, while our proposed framework achieved 94.7% using the pre-trained model of the same.

D. Ablation Study

This paper's proposed framework includes two key modules: LFE and GFE. Furthermore, patch-based LFE employs the sequencer LSTM model [29] consisting of three sequencer versions based on the numbers of sequencer blocks (i.e., sequencer-18, sequencer-24, and sequencer-36). Therefore, we design different ablation studies to analyze the contribution of each module. All experiments are conducted with the pre-trained model of the ImageNet dataset. The experimental results are shown in Table III. We can see that sequencer-24 achieved better accuracy than sequencer-18. However, accuracy slightly deteriorates from that of sequencer-36. We think that a large number of sequencer blocks overfit due to the small sample of the AFID dataset. Therefore, we choose sequencer-24 architecture in the LFE module. Regarding the GFE alone, we evaluated the accuracy of the VGG-16 model separately, achieving an accuracy of 86.3%. We conclude that aggregating the extracted features from LFE and GFE improves the overall accuracy, as demonstrated in Table II and Fig. 3.

V. Conclusion and Future challenge

We demonstrate that deep learning-based algorithms can analyze and interpret facial features to identify individuals with autism spectrum disorder (ASD). We proposed a method for the detection and classification of ASD using local and global feature representation of facial images. It sheds light on the possibility of automated detection and diagnosis of ASD. A vision-transformer (ViT)-based technique sequencer was exploited as Local Feature Extractor along with the modified VGG-16 model for Global Feature Extractor. Additionally, this paper illustrates that combining local and global facial features can increase the accuracy of ASD classification. Despite the proposed methods' great accuracy, certain things could still be improved. For instance, the proposed method

TABLE II

COMPARISON OF THE STATE-OF-THE-ART CLASSIFICATION METHODS ON AUTISM FACIAL IMAGE DATASET (AFID). '-' INDICATE THE INFORMATION IS NOT AVAILABLE ON THE RESPECTIVE METHOD.

Author	Method	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]
Alsaade et al. [21]	Xception	91.0	-	88.4	-
Elshoky et al. [41]	VGG-16	89.0	-	-	-
Mujeeb et al. [40]	Xception	90.0	92.0	88.5	-
Akter et al. [20]	MobileNet	92.1	-	-	-
Jahanara et al. [31]	ResNet50	84.0	-	-	-
Hosseini et al. [14]	MobileNet	94.6	-	-	-
Cao et al. [16]	ViTASD	93.1	-	-	-
This study	**Proposed**	**94.7**	**94.0**	**95.3**	**94.6**

TABLE III

STUDY OF THE EFFECTIVENESS OF THE NUMBER OF THE SEQUENCER BLOCKS OF LOCAL FEATURE EXTRACTOR (LFE) ALONG WITH GLOBAL FEATURE EXTRACTOR (GFE).

Model	Pre-trained	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]
Sequencer-18	ImageNet	92.3	91.5	93.3	90.0
Sequencer-24	ImageNet	93.0	92.1	94.0	90.0
Sequencer-36	ImageNet	92.3	89.4	96.0	90.0
VGG-16	ImageNet	88.3	91.0	88.4	89.6

considered two different kinds of feature extractors. It may be added other features to analyze the accuracy. Furthermore, the proposed framework was validated on a single publicly available dataset. It would be another challenge and future direction to collect a real facial dataset of individuals with autism to validate our proposed framework. Furthermore, the model's accuracy could be improved by fine-tuning the pre-trained model using the AffectNet dataset, which comprises the largest facial image dataset. Training on this dataset can capture more details about facial features like the distance between different key points, nose, eye, ear, and cheeks than those trained on other datasets.

ACKNOWLEDGMENT

This work was partially supported by the ICT division, Government of the People's Republic of Bangladesh, No: (1280101-120008431-3631108).

REFERENCES

[1] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders, text revision (dsm-iv-tr). *(No Title)*, 2000.

[2] Matthew J Maenner, Kelly A Shaw, Amanda V Bakian, Deborah A Bilder, Maureen S Durkin, Amy Esler, Sarah M Furnier, Libby Hallas, Jennifer Hall-Lande, Allison Hudson, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years?autism and developmental disabilities monitoring network, 11 sites, united states, 2018. *MMWR Surveillance Summaries*, 70(11):1, 2021.

[3] Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. *Autism Research*, 15(5):778–790, 2022.

[4] Marc Fakhoury. Autistic spectrum disorders: A review of clinical features, theories and diagnosis. *International Journal of Developmental Neuroscience*, 43:70–77, 2015.

[5] Yong-hui Jiang, Ryan KC Yuen, Xin Jin, Mingbang Wang, Nong Chen, Xueli Wu, Jia Ju, Junpu Mei, Yujian Shi, Mingze He, et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *The American Journal of Human Genetics*, 93(2):249–263, 2013.

[6] Andrew Pickles, Ann Le Couteur, Kathy Leadbitter, Erica Salomone, Rachel Cole-Fletcher, Hannah Tobin, Isobel Gammer, Jessica Lowry, George Vamvakas, Sarah Byford, et al. Parent-mediated social communication therapy for young children with autism (pact): long-term follow-up of a randomised controlled trial. *The Lancet*, 388(10059):2501–2509, 2016.

[7] Eric Schopler, Robert J Reichler, Robert F DeVellis, and Kenneth Daly. Toward objective classification of childhood autism: Childhood autism rating scale (cars). *Journal of autism and developmental disorders*, 1980.

[8] Daniene Neal, Johnny L Matson, and Megan A Hattier. A comparison of diagnostic criteria on the autism spectrum disorder observation for children (asd-oc). *Developmental Neurorehabilitation*, 15(5):329–335, 2012.

[9] Ibrahim Abdulrab Ahmed, Ebrahim Mohammed Senan, Taha H Rassem, Mohammed AH Ali, Hamzeh Salameh Ahmad Shatnawi, Salwa Mutahar Alwazer, and Mohammed Alshahrani. Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. *Electronics*, 11(4):530, 2022.

[10] Md Shale Ahammed, Sijie Niu, Md Rishad Ahmed, Jiwen Dong, Xizhan Gao, and Yuehui Chen. Darkasdnet: Classification of asd on functional mri using deep neural network. *Frontiers in Neuroinformatics*, 15:635657, 2021.

[11] Shuaibing Liang, Aznul Qalid Md Sabri, Fady Alnajjar, and Chu Kiong Loo. Autism spectrum self-stimulatory behaviors classification using explainable temporal coherency deep features and svm classifier. *IEEE Access*, 9:34264–34275, 2021.

[12] Beibin Li, Sachin Mehta, Deepali Aneja, Claire Foster, Pamela Ventola, Frederick Shic, and Linda Shapiro. A facial affect analysis system for autism spectrum disorder. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4549–4553. IEEE, 2019.

[13] Marco Leo, Pierluigi Carcagnì, Cosimo Distante, Paolo Spagnolo, Pier Luigi Mazzeo, Anna Chiara Rosato, Serena Petrocchi, Chiara Pellegrino, Annalisa Levante, Filomena De Lumè, et al. Computational assessment of facial expression production in asd children. *Sensors*, 18(11):3993, 2018.

[14] Mohammad-Parsa Hosseini, Madison Beary, Alex Hadsell, Ryan Messersmith, and Hamid Soltanian-Zadeh. Deep learning for autism diagnosis and facial analysis in children. *Frontiers in Computational Neuroscience*, page 119, 2022.

[15] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[16] Xu Cao, Wenqian Ye, Elena Sizikova, Xue Bai, Megan Coffee, Hongwu Zeng, and Jianguo Cao. Vitasd: Robust vision transformer baselines for autism spectrum disorder facial diagnosis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] Srividhya Ganesan, J Senthil, et al. Prediction of autism spectrum disorder by facial recognition using machine learning. *Webology*, 18(Special Issue on Information Retrieval and Web Search):406–417, 2021.

[18] Piosenka G. The ASD children dataset. `https://www.kaggle.com/datasets/imrankhan77/autistic-children-facial-data-set`, 2021.

[19] Marco Del Coco, Marco Leo, Pierluigi Carcagni, Paolo Spagnolo, Pier Luigi Mazzeo, Massimo Bernava, Flavia Marino, Giovanni Pioggia, and Cosimo Distante. A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1401–1407, 2017.

[20] Tania Akter, Mohammad Hanif Ali, Md Imran Khan, Md Shahriare Satu, Md Jamal Uddin, Salem A Alyami, Sarwar Ali, AKM Azad, and Mohammad Ali Moni. Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain Sciences*, 11(6):734, 2021.

[21] Fawaz Waselallah Alsaade and Mohammed Saeed Alzahrani. Classification and detection of autism spectrum disorder based on deep learning algorithms. *Computational Intelligence and Neuroscience*, 2022, 2022.

[22] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[23] Angelina Lu and Marek Perkowski. Deep learning approach for screening autism spectrum disorder in children with facial images and analysis of ethnoracial factors in model development and application. *Brain Sciences*, 11(11):1446, 2021.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In International Conference on Learning Representations*, May 2015.

[25] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10323–10333, 2023.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *In: Proceedings of the 9th International Conference on Learning Representations, 2021.*

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[28] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[29] Yuki Tatsunami and Masato Taki. Sequencer: Deep lstm for image classification. In *Conference on Neural Information Processing Systems, NeurIPS 2022*, 2022.

[30] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[31] Shaik Jahanara and Shobana Padmanabhan. Detecting autism from facial image. 2021.

[32] Guido Van Rossum, Fred L Drake, et al. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Ross Wightman et al. Pytorch image models, 2019.

[35] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[36] H Zhang, M Cisse, Y Dauphin, and D Lopez-Paz. mixup: Beyond empirical risk management. In *6th Int. Conf. Learning Representations (ICLR)*, pages 1–13, 2018.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[39] Zeyad AT Ahmed, Theyazn HH Aldhyani, Mukti E Jadhav, Mohammed Y Alzahrani, Mohammad Eid Alzahrani, Maha M Althobaiti, Fawaz Alassery, Ahmed Alshaflut, Nouf Matar Alzahrani, and Ali Mansour Al-Madani. Facial features detection system to identify children with autism spectrum disorder: deep learning models. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[40] KK Mujeeb Rahman and M Monica Subashini. Identification of autism in children using static facial features and deep neural networks. *Brain Sciences*, 12(1):94, 2022.

[41] Basma Ramdan Gamal Elshoky, Eman MG Younis, Abdelmgeid Amin Ali, and Osman Ali Sadek Ibrahim. Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images. *ETRI Journal*, 44(4):613–623, 2022.