DeepASD: Facial Image Analysis for Autism Spectrum Diagnosis via Explainable Artificial Intelligence

Hyebin Kang
Medical AI Research Team
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
khb1029@naver.com

Minuk Yang
Medical AI Research Team
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
yhu0409@naver.com
(0000-0002-2141-7484)

Geun-Hyeon Kim
Medical AI Research Team
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
kgh5408@nate.com

Tae-soo Lee
Department of Biomedical Engineering
Chungbuk National University
Chungcheongbuk-do, Rep. of Korea
tslee@chungbuk.ac.kr

Seung Park
Department of Biomedical Engineering
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
spark.cbnuh@gmail.com

***Abstract*— Early and accurate diagnosis of Autism spectrum disorder (ASD) is crucial, but current diagnoses are subjective, time-consuming, and expensive. Recent studies used deep learning for facial images to diagnose ASD. However, the criteria are still unclear. To address these issues, we applied an explainable artificial intelligence technique to four convolutional neural networks (MobileNet, Xception, EfficientNet, and an ensemble model). We utilized gradient-weighted class activation mapping to suggest ASD diagnostic criteria based on facial morphology features. We achieved a high AUROC of 0.89 with the ensemble models. Our study provides objective and easy-to-understand diagnostic methods for early diagnosis of ASD.**

***Keywords*— Autism spectrum disorder, Diagnosis, Deep learning, Explainable artificial intelligence, Gradient-weighted class activation mapping, Convolutional neural network**

I. INTRODUCTION

Autism spectrum disorder (ASD) is a disorder with developmental differences characterized by impairments in social interaction, communication disorders, and repeated behaviors [1]. According to the WHO, 1 in 100 children has autism [2] but some studies reported figures substantially higher [3]. ASD drastically reduces the quality of life, therefore, early and accurate diagnosis is vital for ASD management. However, the process can be extended through observation [4]. The diagnosis of ASD has not had a unique single tool, and the cost of testing depends on the circumstances, as clinical findings may add additional medical examinations [5]. In the hospitals typically using the ADOS-2 [6], CARS2 [7], and DSM-5 [8] to checkup ASD. These tools are based on the subjective judgment of a specialist [6-8], especially for ADOS-2 which is expensive and time-consuming [6]. Therefore, objective and easy-to-understand diagnostic methods are essential for early diagnosis [9].

Children with ASD are visibly different from typically developing (TD) children, as they are affected not only by their brains but also by their appearance [10]. Recently, studies have been published diagnosing ASD with deep learning models through facial images [9-14]. These studies used various convolutional neural network (CNN) models to classify images: ResNet [15], MobileNet [16], Xception [17], VGG [18], Inception [19], EfficientNet [20], and NasNet [21]. Although these studies have shown the possibility of diagnosing ASD children by their faces but did not provide criteria for the diagnosis.

We introduced the deep learning algorithm and an explainable artificial intelligence (XAI) technique to overcome mentioned previous issues: 1) Long time-consuming ASD checkups in the hospital 2) Diagnostic criteria for ASD unclear in a deep learning. XAI technique can explain the reason for the decision-making process of a deep learning that was unknown due to the black box characteristic. In this study, we utilized four CNN models MobileNet [16], Xception [17], EfficientNet [20], and ensemble to distinguish ASD from TD children. Furthermore, to suggest the ASD diagnostic criteria that facial morphology features, we applied gradient-weighted class activation mapping (Grad-CAM) [22], which is used widely on image classification tasks. We conducted experiments using Autism-Image-Data [23] from the Kaggle platform, which is a publicly available dataset containing 2,940 face images.

II. RELATED WORK

With the development of deep learning technology, various methods for diagnosing ASD using deep learning are being studied. Several studies have used facial images with deep-learning models to diagnose ASD. And most of these researchers used the Autism-Image-Data [23] as the only publicly available dataset for this purpose on Kaggle.

Alsaade, Fawaz Waselallah, and Mohammed Saeed Alzahrani. [10] aims to develop an accurate model that can automatically identify and diagnose ASD based on behavioral and clinical data. The authors used several deep learning algorithms, including CNNs and recurrent neural networks (RNNs), to analyze various features related to ASD, such as social communication skills, repetitive behavior, and sensory sensitivity. This study suggests that deep learning algorithms can effectively analyze behavioral and clinical data to assist in the diagnosis and treatment of ASD.

A. Lu and M. Perkowski [11] investigate the impact of ethnological factors on the model's development and application. The authors collected facial images of children with ASD and TD datasets from different racial and ethnic groups. The result showed that the proposed approach achieved high accuracy in detecting ASD in children using facial images. They suggest that further studies should be conducted on specific ethnic and racial groups to improve the performance of the model.

M. S. Alam, M. M. Rashid, R. Roy, A. R. Faizabadi, K. D. Gupta, and M. M. Ahsan [12] proposed an approach that utilizes a pre-trained CNN architecture, to extract features from the facial images. They used three different deep-learning models (VGG16, InceptionV3, and ResNet50) to classify the images. And they performed feature visualization to identify the specific facial regions that were most important for classification. These findings suggest that deep learning models can be an effective tool for diagnosing ASD using facial images and use in clinical settings.

Hosseini, Mohammad-Parsa, et al. [14] present a deep learning model for diagnosing ASD and analyzing facial features in children. The study analyzes the importance of different facial regions in ASD diagnosis and identifies specific facial features associated with ASD. The authors conclude that the proposed deep learning model has the potential to be a valuable tool for ASD diagnosis and understanding the relationship between facial features and the disorder.

Previous studies have achieved high performance in diagnosing ASD by using deep learning. However, due to the black box characteristic of deep learning, it is difficult to determine the specific reasons behind the results. Therefore, in this study, we aim to diagnose ASD using facial images with deep learning methods. Additionally, we will apply the XAI technique to visualize which facial features are affected by classification.

III. MATERIALS AND METHOD

A. Data description and data preprocessing

Autistic children commonly have a long face, short middle face, flat nose, big eyes, big mouth, and facial asymmetry as shown in Fig 1.

This database consists of a total of 2940 (train: 2540, test: 300, validation: 100). The range of children's ages in this dataset is from 2 to 14, and is mainly distributed between the ages of 2 to 8.

In the preprocessing, we resized all images to 224x224x3 size because all images have different sizes. TD and ASD were labeled as 0 and 1, respectively. We used to train and validation sets augmented as follows: rotation up to 30 degrees, horizontal and vertical shifts up to 20%, shearing up to 20%, zooming up to 20%, horizontal flipping, rotation up to 30 degrees, horizontal and vertical shifts up to 20%, shearing up to 20%, zooming up to 20%, and horizontal flipping.

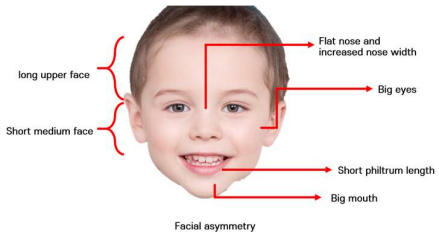

Figure 1. Common facial features of ASD children. This image is a non-copyright people used to show commonalities with ASD.

B. Explanation of autism diagnostic models

To classify ASD and TD children on the image dataset, we introduced a 2-dimensional CNN that is used for image classification, object detection, and computer vision tasks. In this study, we conducted experiments to evaluate the autism diagnosis performance using three representative CNN models that are MobileNet, Xception, and EfficientNet. All models get a 224x224x3 RGB image as input and output an autism probability.

1) MobileNet

MobileNet [16] is developed for efficient deployment on mobile devices with limited computational resources. In this model, a depthwise separable convolution [17] applies a separate convolutional filter to each input channel, and a pointwise convolution applies a 1x1 convolutional filter to combine the output channels of the depthwise convolution. This algorithm reduces the number of parameters and computational costs and still maintains the good performance of CNNs. Although various variants of MobileNet have been developed, we utilized a basic MobileNet model whose structure is illustrated in Fig 2.

Figure 2. MobileNet architecture. All convolution layers and depthwise convolution layers are followed by batch normalization and ReLU activation function. Conv: Convolution layer; DwConv: Depthwise convolution layer; S number: The number of strides.

2) Xception

Xception [17] is a variation of the Inception architecture that replaces the standard convolutional layers with depthwise separable convolutions, similar to those used in MobileNet. Depthwise separable convolution layers are used in entirely this model, and bottleneck structures with skip connection are applied for efficient use of parameters and computation. As shown in Fig 3., the Xception model has three compartments: Entry, Middle, and Exit flow. The entry flow reduces the spatial dimensions of the input image, the middle flow extracts image features, and the exit flow performs the final classification.

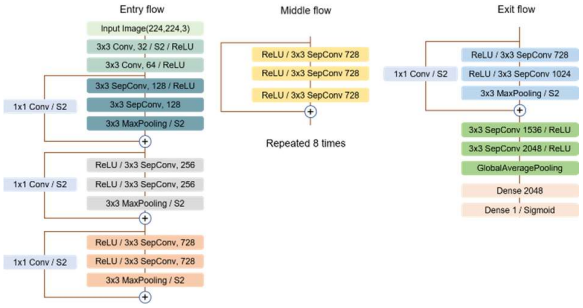

Figure 3. Xception architecture. All convolution layers are followed by batch normalization. Conv: Convolution layer; SepConv: Separable convolution layer.

3) EfficientNet

EfficientNet [20] uses a mobile inverted bottleneck convolution to standard convolution (MBConv) which balances depth, width, and resolution dimensions for efficient model construction. MBConv Block is a residual block that uses an inverted structure for efficiency in image classification and also adds squeeze-and-excitation optimization (Fig 4.).

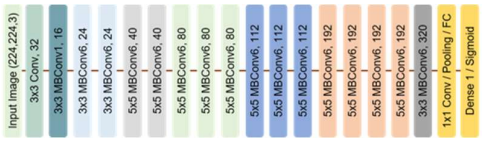

Figure 4. EfficientNet architecture. MBConv: Mobile inverted bottleneck convolution

4) Ensemble model

The ensemble model was implemented by combining three models: MobileNet, Xception, and EfficientNet. Because the heatmap outputted by Grad-CAM was different from the model, an ensemble model was created to observe the integrated decision-making process of the three models. As shown in Fig 5., the ensemble model concatenated the last layers of each model. Zero padding was added to MobileNet to match the shapes with the other models. After concatenating the three models, an additional convolution layer was created for inputting into Grad-CAM.

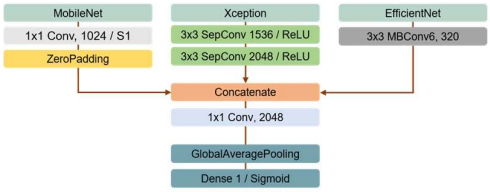

Figure 5. An ensemble model architecture combined by MobileNet, Xception, and EfficientNet.

C. *Grad-CAM for visualizing the decision-making process of CNNs*

The Grad-CAM is a gradient-based analysis method that visualizes feature maps and mainly produces heatmaps. The heatmap by Grad-CAM represents the contribution and importance of input pixels. By taking the weights of the last layer of the CNN and representing them as a heatmap, we can visualize the area of CNN model has focused on. An illustration is shown in Fig 6. for the Grad-CAM [22]. The weight w_c^k represents the contribution of the kth feature map to the target class c. We compute the gradient of y^cconcerning feature maps A of a convolution layer, i.e.$\frac{\partial y^c}{\partial A_{ij}^k}$.

The ReLU function receives the weighted sum of all feature maps to generate the class activation heatmap. w_c^k are precisely α_c^k where CAM can be applied to any CNN-based architecture.

$$W_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (2)$$

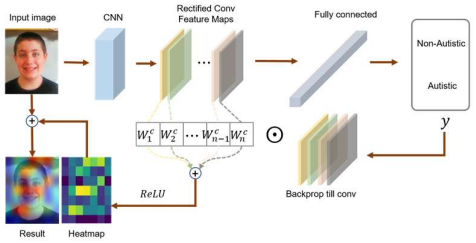

Figure 6. Gradient-weighted class activation mapping flow chart. Forward propagate the image through the model to obtain the raw class then, back propagated to the rectified convolutional feature map of interest, where we can compute the coarse Grad-CAM localization.

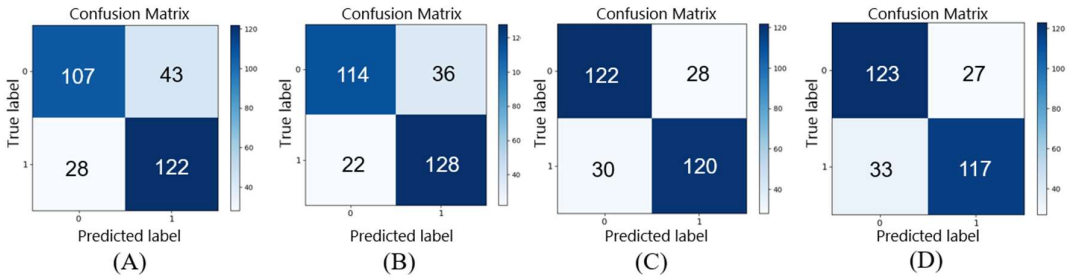

Figure 7. Confusion matrix for the (A) MobileNet, (B) Xception, (C) EfficientNet, (D) Ensemble model.

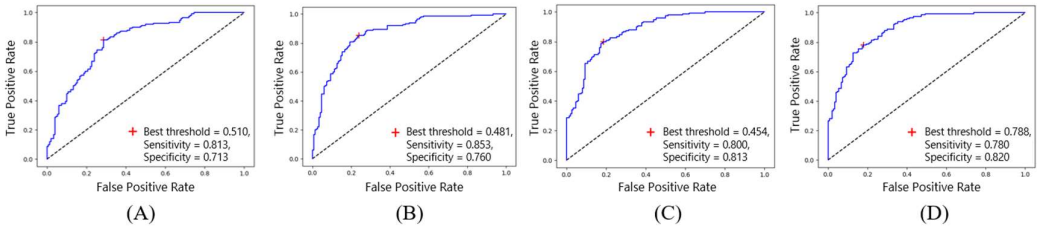

Figure 8. Receiver operating characteristic curve for the (A) MobileNet, (B) Xception, (C) EfficientNet, (D) Ensemble model.

D. Performance metrics

To evaluate the performance of the models, we used sensitivity, specificity, F1-score, and area under the receiver operating characteristic (AUROC) as evaluation metrics. Specificity is an index indicating the degree to which the actual negative class (normal persons) can be accurately classified, and sensitivity is an index indicating the degree to which the actual positive class (abnormal persons) can be accurately classified. Then calculated true negative (TN), true positive (TP), false negative (FN), and false positive (FP) were to compute sensitivity, specificity, and F1-score following these formulas.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{F1 score} = 2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}} \tag{6}$$

Sensitivity represents the proportion of people with the disease who are correctly diagnosed as having the disease. And sensitivity is the same recall. Specificity represents the proportion of healthy people who are correctly identified as healthy. F1-score is a combined metric that incorporates both recall and precision.

ROC is a way to visualize and evaluate the performance of binary classification models and AUROC represents the area under the ROC curve. The value of AUROC can be used to determine whether the model's performance is good or bad.

IV. Result and Discussion

In this study, we have performed binary classification experiments using four models on a dataset of ASD image database. All models were compiled with hyperparameters set as follows: epochs of 100, batch size of 32, a learning rate of 0.001, and Adam optimizer. Additionally, to compare ASD facial features with the decision-making process of deep learning, the heatmap localized by Grad-CAM was visualized. We listed the diagnostic performance of each model in Table 1. AUROCs of all models are higher than 0.80. While MobileNet demonstrated the lowest performance which achieved a sensitivity of 0.8133, specificity of 0.7133, F1-score of 0.7746, and AUROC of 0.8083, the ensemble model showed the highest performance which achieved a sensitivity of 0.7800, specificity of 0.8200, F1-score of 0.7959, and AUROC of 0.8882. In the confusion matrix shown in Fig 7., the sensitivity of the ensemble model is the highest, and the specificity of the Xception model is the highest. As shown in

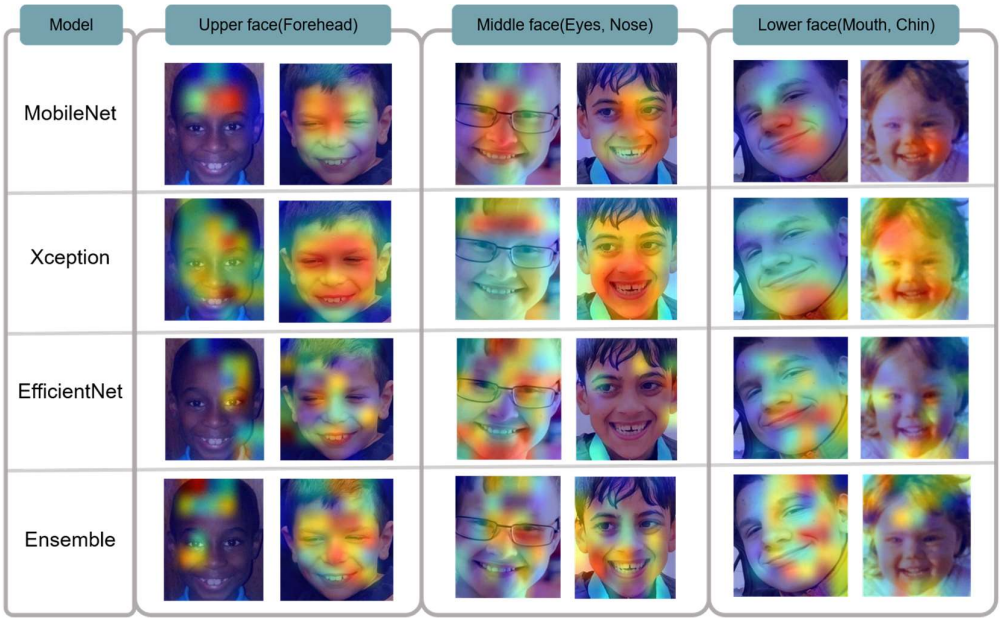

Figure 9. The output results of Grad-CAM are divided into three sections: Upper Face, Middle Face, and Lower Face, according to the model. The area highlighted in red represents the region that has the greatest impact on the result. However, depending on the model, it may not be displayed accurately in the criteria.

Fig 8., ROC curves show visually that the AUROC of the ensemble model is the largest. Furthermore, Fig 9. demonstrates the Grad-CAM results of each model. While MobileNet focused on the area where the ASD features are visible, Xception tends to focus on the entire face. EfficientNet heatmap shows ASD facial features in various details. Lastly, the ensemble model showed a combination of models that had the most effect on the image.

Table 1. **Performances of four models on test dataset.**

Models	**Performance score**			
	Sensitivity	***Specificity***	***F1-score***	***AUROC***
MobileNet	0.8133	0.7133	0.7746	0.8083
Xception	0.8533	0.7600	0.8153	0.8668
EfficientNet	0.8000	0.8133	0.8054	0.8778
Ensemble	0.7800	0.8200	0.7959	0.8882

This study suggests the possibility of diagnosing ASD with facial images, and this technique can recommend early diagnosis in a hospital to parents of children with suspected ASD. Therefore, early diagnosis and timely treatment can reduce treatment costs and help ASD children live normal life.

However, there are limitations to this study. First, since images don't reflect dynamic changes over time, there is a limitation to diagnosing the ASD method with a face image at a specific period. Therefore, for precision diagnosis, we need to analyze videos or several images taken from a newborn. Second, since each model focuses on different areas when models diagnose ASD (Fig 10.), it is necessary to analyze with various models rather than using a single model.

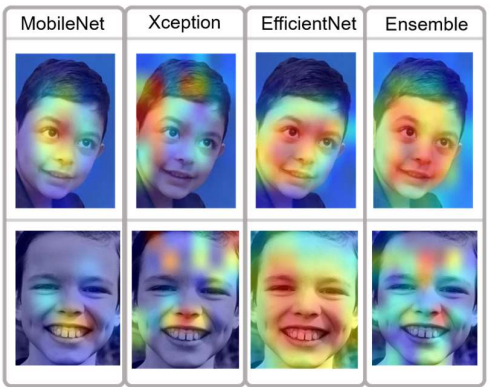

Figure 10. The difference in Grad-CAM output by the model, even though it's the same image.

V. Conclusion

An early and accurate diagnosis of ASD is important for managing the disorder, but the diagnosis process can be time-consuming and subjective. The use of deep learning algorithms has shown promise in diagnosing ASD through facial images. However, the lack of diagnostic criteria and the "black box" characteristic of the algorithms has been a challenge. This study aimed to overcome these issues by utilizing XAI techniques and four CNN models to distinguish ASD from typically developing children. The models used were MobileNet, Xception, EfficientNet, and an ensemble model. To suggest ASD diagnostic criteria based on facial morphology, the study applied Grad-CAM. The study utilized the Autism-Image-Data dataset from Kaggle, which

contains 2,940 face images. The results showed that the MobileNet, Xception, and EfficientNet models had AUROCs of 0.81, 0.87, and 0.88, respectively, while the ensemble model achieved an accuracy of 0.89. This study also identified facial features that could be used as diagnostic criteria for ASD. This study shows the potential of a deep learning and XAI in improving the diagnosis of ASD, which could lead to earlier and more accurate management of the disorder.

ACKNOWLEDGEMENTS

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1074070021).

REFERENCES

[1] "National Institute of Mental Health." https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd#:~%7B%7D:text=Autism%20spectrum%20disorder%20(ASD)%20is,first%20two%20years%20of%20life (accessed.

[2] J. Zeidan *et al.*, "Global prevalence of autism: A systematic review update," *Autism Research,* vol. 15, no. 5, pp. 778-790, 2022.

[3] "World Health Organization." https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (accessed.

[4] H. Wang, L. Chi, C. Su, and Z. Zhao, "ASDFace: Face-based Autism Diagnosis via Heterogeneous Domain Adaptation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 4999-5003.

[5] "Asan Medical Center, health info." https://www.amc.seoul.kr/asan/healthinfo/disease/diseaseDetail.do?contentId=31896 (accessed.

[6] V. Hus and C. Lord, "The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores," *Journal of autism and developmental disorders,* vol. 44, pp. 1996-2012, 2014.

[7] L. Jurek *et al.*, "Response (minimum clinically relevant change) in ASD symptoms after an intervention according to CARS-2: consensus from an expert elicitation procedure," *European child & adolescent psychiatry,* pp. 1-10, 2021.

[8] "autism speaks." https://www.autismspeaks.org/autism-diagnosis-criteria-dsm-5 (accessed.

[9] 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.

[10] 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.

[11] A. Lu and M. 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,* vol. 11, no. 11, p. 1446, 2021.

[12] M. S. Alam, M. M. Rashid, R. Roy, A. R. Faizabadi, K. D. Gupta, and M. M. Ahsan, "Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach," *Bioengineering,* vol. 9, no. 11, p. 710, 2022.

[13] Z. A. Ahmed *et al.*, "Facial features detection system to identify children with autism spectrum disorder: deep learning models," *Computational and Mathematical Methods in Medicine,* vol. 2022, 2022.

[14] 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.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[16] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861,* 2017.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[19] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019: PMLR, pp. 6105-6114.

[21] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697-8710.

[22] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?," *arXiv preprint arXiv:1611.07450,* 2016.

[23] "Kaggle." https://www.kaggle.com/datasets/cihan063/autism-image-data (accessed.