IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 766~779
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp766-779

Improving efficiency of autism detection based on facial image landmarks

Nguyen Trong Tung[1,2], Ngo Duc Vinh[3], Ha Manh Toan[4], Do Nang Toan[4]
[1]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, Dong A University, Da Nang City, Vietnam
[3]Faculty of Information Technology, Hanoi University of Industry, Hanoi, Vietnam
[4]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

Article Info

Article history:

Received Jun 11, 2025
Revised Nov 4, 2025
Accepted Jan 10, 2026

Keywords:

Autism detection
Deep learning
Facial augmentation
Facial child image
Facial landmark
Heatmap

ABSTRACT

Autism is a serious mental health problem with long-term effects on life. Therefore, early diagnosis is a topical issue for effective treatment. This study proposes a novel facial landmark transformation-based data augmentation method that allows for the generation of geometric transformations related to facial geometry. This method increases the generalizability and provides a perspective on the role of facial regions in autism detection. The proposed augmentation method ensures the generation of variants that are consistent with the facial image structure and the nature of the facial image. Next, conduct a comprehensive and comparative study with EfficientNet-B0, EfficientNet-B4, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2, DenseNet-121 and DenseNet-201. Also analyze the model's attention over the main regions of the face that are related to facial landmarks. The results clearly show that the models trained with the proposed method outperform the default augmentation method. Specifically, when averaging the measures across the tested models, the results are 0.905417 for accuracy, 0.962133 for area under the curve (AUC), 0.9198 for precision, 0.888333 for recall, and 0.903678 for F1-score. Furthermore, when analyzing the gradient-weighted class activation mapping (Grad-CAM) heatmaps, the high-value regions are clearly concentrated on the main areas of the face. Source code is published on GitLab platform.

Corresponding Author:

Do Nang Toan
Institute of Information Technology, Vietnam Academy of Science and Technology
Hanoi, Vietnam
Email: dntoan@ioit.ac.vn

1. INTRODUCTION

One of the important problems affecting neurodevelopment is autism spectrum disorder (ASD), which is related to abnormalities in the development of the brain and has a visual impact on physical characteristics and facial expressions [1]. Signs of autism can appear in many different situations in the life of the affected person. Autism is often expressed in many different characteristics, such as repetitive behaviors and a lack of social communication [2]. This phenomenon, when appearing for a long time, will have serious impacts on the lives of autistic people as well as those around them.

Sulkes [3] has shown that the risk of having another child with autism is about 3-10% for parents who already have a child with autism. In the United States, it is estimated that 1 in 59 children aged 8 and older has autism [4]. In addition, the World Health Organization has published alarming figures on the prevalence of autism. Specifically, according to a report [5], when conducting statistics on 16% of children

globally, 0.67% of them have autism. The Vietnamese government and community have been contributing several attempts to deal with autism. One case is the effort relating to world autism awareness day of the Ministry of Labor, Invalids and Social Affairs [6]. This activity helps the community understand and support autistic children. To detect autism in children, an effective approach is to use deep learning models. On the one hand, traditional diagnosis with behavioral assessment is time-consuming and requires careful observation by experts. Effective assessments will require experts with many years of experience and enough time to observe children, not to mention that such evaluations will be subjective. On the other hand, the use of deep learning approaches has been demonstrated in many studies and implementations for image recognition problems in the medical field. In practice, the application of deep learning methods will be a way to screen quickly, objectively, and easily expand.

Recognizing autism from facial images is a complex pattern recognition task involving subtle facial cues. That is because the facial expressions of autistic children are not always obvious. They can appear in small changes in eye gaze, facial expression, or the correlation between facial features. This requires artificial intelligence models to be able to exploit deep features instead of relying solely on surface visual cues. In addition, artificial intelligence models need to meet several important requirements in the medical context. The first is generalizability. Accordingly, the model must not only perform well on training data but also maintain acceptable performance when applied to new data from different sources. The second is interpretability. Specifically, predictions must be accompanied by clear explanations, such as heat maps depicting attention to help doctors understand which regions of the image the model focuses on. The third is data-efficiency. Since medical data, especially data on children with autism, are often scarce and difficult to collect, the model must learn well even with a limited number of training samples. This can be done through transfer learning, augmentation, or semi-supervised learning.

To extract richer features from limited data, other approaches besides traditional transfer learning can also help exploit the limited data more effectively in medical applications. One example is multi-layer fine-tuning, which helps convolutional neural network (CNN) models learn features that are more suited to subtle facial cues. Another example is multi-task learning, which allows the combination of related tasks, such as landmark localization or emotion prediction, to share a common representation. In addition, self-supervised learning leverages unlabeled data to learn facial structures before classification. These strategies promise to provide more effective features for autism recognition.

The topic of autism diagnosis has attracted the attention of many researchers in the field of artificial intelligence. In 2022, a study evaluating eye behaviors was performed in [7] for autism diagnosis. The studies were performed with the construction of various tasks, with path computation and recognition model design. Various techniques were tested such as ResNet18 and inception CNN as well as image transformation techniques with gray level co-occurrence matrix and local binary pattern (LBP). A study on autism detection on magnetic resonance imaging (MRI) data was presented by Heinsfeld *et al.* [8]. The authors presented an architecture consisting of two convolutional encoders and tested it multiple times using cross-validation. In 2023, Farooq *et al.* [9] published their work about autistic diagnosis with federated learning method. They combined support vector machine (SVM) and logistic regression (LR) to experiment on tabular data and reached 0.98 accuracy. An autistic classification study was published [10]. The authors used a framework for evaluating 8 machine learning algorithms. The results were performed on four autistic datasets, including toddlers, adolescents, children, and adults, and evaluated with various statistical evaluation measures. In 2022, Karri *et al.* [11] presented a work using facial images. Their work used DenseNet for identifying ASD and was tested on a face dataset on the Kaggle platform. They also built a simple web tool to support the medical facilities. In 2023, Ghazal *et al.* [12] presented research on designing a CNN that was inspired by AlexNet architecture. They used input as facial image data and tried to extract facial features effectively. The authors achieved 87.6% validation sensitivity, 87.6% validation specificity, and 87.7% validation accuracy.

In 2023, Li *et al.* [13] conducted a study using MobileNetv3-Large and MobileNet-V2 to diagnose autism based on facial child images. The authors designed a framework using transfer learning and integrating different classifiers. In results, their work achieved 87.67% accuracy evaluating MobileNet-V3-Large and 88.33% accuracy evaluating MobileNet-V2. In 2024, Ahmad *et al.* [14] presented a study to detect autism from facial images using many models as visual geometry group (VGG)16, VGG19, MobileNet-V2, AlexNet, ResNet-34, and ResNet50. They used approximately 2 hours for training and nearly 3 minutes for testing. They evaluated several resolutions of the input image and achieved the highest accuracy of 0.86 with 248×248. In 2024, Reddy and Andrew [15] conducted a deep learning study to classify autism. With a transfer learning approach, three pretrained models, including EfficientNetB0, VGG16, and VGG19, were experimented with. In the results, the authors reached the highest accuracy is 0.879 for the EfficientNetB0 model in their experiment. A comparative table summarizing existing methods, datasets, augmentation strategies, and performance metrics is presented in Table 1.

Table 1. The comparative table summarizing existing studies

Study	Method/technique	Datasets	Augmentation	Performance metrics
Ahmed *et al.* [7], 2022	LBP, grey level co-occurrence matrix, SVM, Google-Net, ResNet-18	Figshare data (eye-tracking scan paths)	Flipping, multi-angle rotation, displacement, and shearing	Most of the accuracy, precision, sensitivity, specificity, and AUC are more than 0.93
Heinsfeld *et al.* [8], 2018	Two stacked denoising autoencoders, multilayer perceptron	Autism brain imaging data exchange (ABIDE)	Not specified	Accuracy, sensitivity, and specificity are around 0.7
Farooq *et al.* [9], 2023	Federated learning, LR, SVM	Four datasets according to the quantitative checklist for autism in children	Not specified	Accuracy for children is around 0.98, and accuracy for adults is around 0.8
Hasan *et al*. [10], 2023	AdaBoost, random forest, decision tree, k-nearest neighbors, gaussian naive Bayes, LR, SVM, and linear discriminant analysis	Four datasets from Kaggle and UCI MLA	Not specified	Most of the accuracy, precision, recall, F1-score, and AUC are more than 0.96
Li *et al.* [13], 2023	MobileNetV2 and MobileNetv3-Large	A dataset from Kaggle	Not specified	Most of the accuracy, sensitivity, specificity, and AUC are more than 0.90
Ahmad *et al.* [14], 2024	ResNet34, ResNet50, AlexNet, MobileNetV2, VGG16, and VGG19	A dataset from Kaggle	Flipping	Best accuracy is 0.92 in the case of ResNet50
Reddy and Andrew [15], 2024	VGG16, VGG19 and, EfficientNetB0	A dataset from Kaggle	Rotating, horizontal flipping, zooming, and height and width shifting	Accuracies are around 0.87, and AUCs are around 0.93

Our research is about diagnosing autism in children with a deep learning approach. This work are interested in using children's facial image data by exploiting facial expression characteristics. It is known that one of the characteristics of autism is abnormal facial expressions in children, such as abnormal signs of facial asymmetry or abnormal facial development due to the influence of neurological development. Based on these observations, this study focuses on using and analyzing landmarks on children's faces in our research using deep learning models for diagnosing autism.

More specifically, to enhance the diversity of data as well as the performance of deep learning models, an augmentation technique using facial landmarks is proposed. Our augmentation differs from geometric transformations in semantics in this problem. While conventional geometric transformations, such as rotation, translation, scaling, or affine, only globally affect the entire face and preserve the overall morphological structure, our augmentation focuses on locally shifting landmarks on the face and warping the image. This approach allows for more sophisticated deformations that accommodate small changes in facial expression and geometric structure. As a result, the model can learn features more effectively than global geometric transformations in the problem of autism recognition from facial images. In previous studies, facial landmarks have also been applied to expression analysis or to support morphological modeling, such as in [16] for disentangling expression and identity, or [17] for diagnosing mandibular deformity. However, in terms of image augmentation, conventional image augmentation methods such as rotation, flip, and affine transformation hardly utilize landmarks, leading to a lack of connection with the biological structure of the face. This emphasizes the novelty of our study in exploiting landmark displacement as an augmentation technique, which both generates diverse data and preserves the semantics of facial structures related to expressions in autistic children.

In addition, to obtain an objective and comprehensive evaluation, comparative experiments are conducted between different well-known models. These experiments analyze the accuracy as well as evaluate the ability to deploy and expand. Another issue of concern is to evaluate the regions of interest of the models on the input image using the gradient-weighted class activation mapping (Grad-CAM) technique in relation to meaningful regions on the facial image. This way can exploit the relationship between the model's autism recognition and the feature locations on the facial image. This will be clear evidence of the role of facial expression features in autism recognition and will be an important basis for further research. In detail, our main contributions include:

i) Propose a novel facial image augmentation technique based on displacing facial landmarks to improve the performance of deep learning models.
ii) Comprehensively evaluate and clarify our hypothesis by conducting a comparative study with EfficientNet-B0, EfficientNet-B4, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2, DenseNet-121, and DenseNet-201.
iii) Analyze the interpretability of models by visualizing model attention with Grad-CAM.

2. METHOD

This section will present the specific contents proposed in our research. Basically, facial image data will be augmented with a focus on facial image augmentation techniques based on facial landmark displacement and facial image warping. Besides, the famous deep learning models are presented for comprehensive and comparative testing. To clarify their effectiveness, experiments with three different augmentation strategies are proposed, and the prediction results are analyzed in relation to facial regions using the Grad-CAM technique [18].

2.1. Autism spectrum disorder detection dataset

This study conducts experiments using a child face image dataset published on the Kaggle platform at https://www.kaggle.com. There are total of 2,936 child face images in this dataset and they are divided into two groups, including autistic and non-autistic. More specifically, the number of images of autistic children is 1,468 images and the number of images of non-autistic children is 1,468 images. This dataset was already split into three subsets as the train set, the validation set, and the test set. In detail, the train set includes 1,268 images of non-autistic children and 1,268 images of autistic children. The validation set includes 50 images of non-autistic children and 50 images of autistic children. The test set includes 150 images of non-autistic children and 150 images of autistic children.

2.2. Proposed facial image augmentation

For training deep learning models, data augmentation plays an important role in dealing with data scarcity. Default image augmentation methods often include operations such as flipping, rotating, and random cropping. These techniques play an important role in enhancing the generalization ability of deep learning models. However, for facial image data, default methods do not take advantage of the structural features of the face. This study proposes to augment facial image data based on manipulating facial landmarks and, on that basis, generate new facial image data, which is described in Algorithm 1. This method helps to create facial image variations based on the structure of facial image data. Thus, it creates new data samples that are consistent with the nature of facial images.

Algorithm 1. The landmark displacement augmentation
Input: face image I
Output: augmented image I'
Process:

```
F=detect_face_bbox(I)
L=detect_landmarks(F, I)
W=calculate_face_with(F)
L'=displace_landmarks(L, W, I, MAX_SHIFT_RATIO)
tris=delaunay_triangulation(L)
for each t in tris:
    p1=get_vertices(L, t)
    p2=get_vertices(L', t)
    T=compute_transform(p1, p2)
    warp_triangular_region(I, I', p1, p2, T)
```

The process is described in Figure 1. With the input being a face image, the first step is to detect facial landmarks. These are the points that play an important role in facial morphology such as eye corners and nose points. This work uses a set of 68 facial points supported in the Dlib library at https://dlib.net/. These landmarks are the basis for warping facial images. For the warping to be performed, the set of facial landmarks is triangulated using the Delaunay triangulation technique [19]. Thus, the warping will be performed by interpolating the pixel values in each sub-triangle of the resulting image based on the corresponding positions for the three vertices of the triangle. This experiments also have the option to pre-calculate the triangulation set to use for the images without having to recalculate each time.

The landmark displacement augmentation method is performed by detecting 68 facial landmarks using the Dlib library. These points are then randomly displaced within a limited range based on the face width and the MAX_SHIFT_RATIO scale value. The displacement ratio parameter is randomly generated with an upper bound of MAX_SHIFT_RATIO, and specifically in the experiment MAX_SHIFT_RATIO is set to 0.2. MAX_SHIFT_RATIO constrains the maximum displacement of landmark points on the face. Here the displacement of the landmark is calculated proportionally to the width of the face, to ensure that the deformation is always proportional to the size of the face, avoiding abnormal deformation of the face leading to unrealistic results. Next, the facial region in the original image is divided into small triangles using Delaunay triangulation. All triangles are iterated through and each triangle is computed by an affine

transformation from the original landmark location to the displaced landmark. Based on that, the resulting image is generated by warping each sub-region in the original image according to the calculated affine transformation. The parameters in our augmentation process are currently set based on intuition and experience. In the future, also consider expert evaluation. In this way, expert evaluation can validate the realism of augmented images for the autism assessment task.

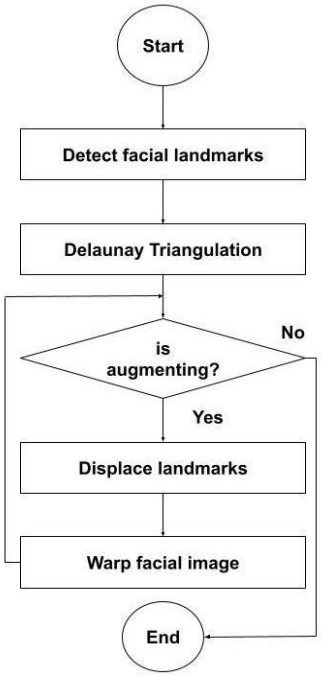


Figure 1. Proposed landmark displacement-based augmentation process

During the data augmentation iteration for each face image, the landmarks are transformed by randomly shifting them in a neighborhood of their original positions. The displacement range is set proportional to the horizontal length of the face and is small enough that the displacement will not disrupt the structure of the face in the image. Once the facial landmarks have been transformed, the new face image is warped by interpolating each region of the identified triangles. The resulting image is fed into the dataset for model training. Some example results were described in Figure 2.

After the transformation using landmarks displacement, the images are further diversified using the default augmentation method. The default augmentation method is designed to accommodate a wide range of input images, not just faces. First, the images are resized to a chosen standard size. This makes images of different sizes compatible with deep learning models. This case will bring them to 224×224. Next, the images are randomly horizontally flipped with a specified probability, in this case 50%. This transformation is a popular choice to help models learn reflection variations, and it is also suitable for faces because of its symmetry. Next, the images are randomly rotated clockwise or counterclockwise within a specified angle range, in this case 10 degrees. Next, randomly transform the pixel values in terms of brightness, contrast, saturation, and hue, making the dataset richer in terms of lighting conditions. Finally, the images are geometrically transformed affinely with small offsets. In the implementation, the default augmentation techniques were performed with the support of the Kornia library [20].

Thus, our augmentation method improves data efficiency, which is a major AI challenge, by generating more realistic facial variants from the original data. This is especially true when working with limited facial image datasets. Instead of relying on simple geometric transformations, this method directly exploits the facial structure, thereby learning more semantically relevant features. In addition, facial data often has potential biases in terms of ethnicity, age, or gender. Our landmark displacement method allows for algorithmically consistent generation of models with random displacements based on common facial landmarks. Due to these characteristics, our method is able to contribute to reducing the impact of potential biases in terms of ideas. As a result, the model is expected to generalize better and be fairer.

Figure 2. Some results of the proposed landmark displacement-based augmentation process

2.3. Deep learning models

EfficientNet [21] is a family of CNN architectures that supports the ability to allow the model to scale evenly across multiple dimensions such as resolution, width, and depth. The optimal use of EfficientNet can help the program still achieve high performance while being able to use fewer resources compared to some other CNN models. In general, EfficientNet can work effectively with complex image data and provides generalization capabilities. This architecture is customized with many different versions depending on the scale of the model. This study chose two versions for our experiment as EfficientNet-B0 and EfficientNet-B4.

ResNet [22] is a well-known deep network architecture in various deep learning problems with a residual connections mechanism that helps to minimize the phenomenon of gradient vanishing as the depth of the network increases. This mechanism is special in that it allows researchers to train very deep networks while maintaining stability. Thanks to that, the model is capable of generalizing many complex features in images. Specific versions of ResNet are often named according to the depth of the architecture. This study chooses 3 versions: ResNet-18, ResNet-50, and ResNet-101 for evaluating the proposed method.

MobileNet [23] is a CNN architecture designed for use in resource-constrained scenarios, such as mobile devices. It is built with depth wise separable convolutions to minimize the number of parameters and the computational burden while still providing significant performance gains in image classification problems. This study uses MobileNet-V2 for experiments.

Similar to ResNet, DenseNet is also a CNN architecture designed to improve the propagation of gradient signals. In DenseNet, each layer is connected to all previous layers. This not only helps in the propagation of gradients but also contributes to the reuse of intermediate features computed at different levels of abstraction. This study uses DenseNet-121 and DenseNet-201.

Compared to newer approaches such as vision transformers, swin transformers, or hybrid CNN-recurrent neural network (RNN) models, vision transformers have the advantage of modeling global relationships between facial regions, but often require large datasets for effective training, which is difficult to meet in the context of limited pediatric autism data. Swin transformer has the advantage of incorporating a hierarchical structure and local attention, but comes with a higher computational cost. In addition, hybrid CNN-RNN models are mainly suitable for video data when analyzing facial motions, while the current study focuses on still images. Therefore, this work uses well-known CNN models such as EfficientNet-B0, EfficientNet-B4, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2, DenseNet-121, and DenseNet-201, which have been proven effective in extracting local features from facial images. This is consistent with our proposed contribution regarding the utilization of facial landmarks.

To train models efficiently and to reduce overfitting, some strategies commonly used in deep learning model training are considered. One example is early stopping, where the training process stops when the evaluation metric on the validation set no longer improves. Others are dropout, or regularization, which help improve the generalization of models.

2.4. Visualizing model attention with Grad-CAM

Grad-CAM plays an important role in indicating the focus of CNN networks in image recognition applications. Convolutional layers discover and record spatial features computed from an input image. For a trained CNN model, the first convolutional layers will play the role of analyzing the basic features of the image while the last convolutional layers will model the semantic features. Thus, the last layers will provide information that can be visually mapped to the location of an object in the input image. In this study, Grad-CAM is used to interpret and validate the role of facial feature regions in autism classification.

Regions in facial images are intuitively understood as regions associated with facial landmarks. Given an input facial image, Grad-CAM will allow the generation of heatmaps with different trained models.

By analyzing the visual location of the heatmaps relative to regions in the facial image, it can be analyze whether the models pay attention to regions important for facial expression. This will also need to be considered in relation to the accuracy of the resulting models. This is evidence that suggests a correlation between important facial locations and autism classification.

2.5. Proposed training and evaluation workflow

Figure 3 presents the proposed workflow in our study, including both the training phase and the evaluation phase. The proposed workflow was designed to analyze the effect of facial landmarks on the way deep learning models classify images to diagnose autism. Experiments were set up for strategies in both the training phase and the evaluation phase.

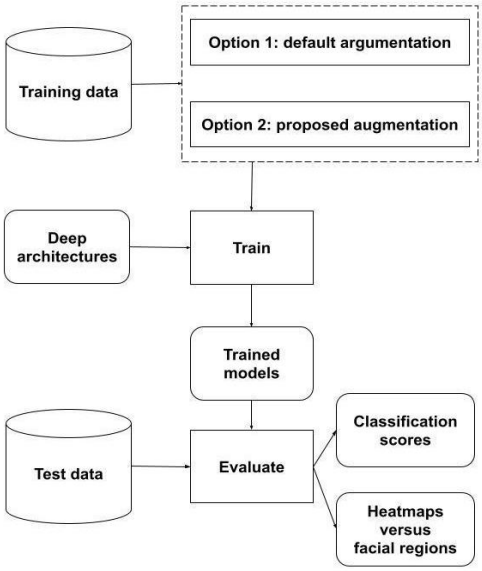


Figure 3. Proposed training and evaluation workflow

In the training phase, different training strategies are set up for all the deep learning models in the experiment. Specifically, there are 2 strategies: default augmentation and the proposed facial image augmentation method. In this way, the influence of facial landmarks, which represent facial expression features, on the training performance of deep learning models. Thus, in the evaluation phase, the expected result is that the proposed facial image augmentation strategy will yield the highest performance, and thereby also clarify the effectiveness of this technique.

The training strategies will be tested with various popular CNN models, namely EfficientNet-B0, EfficientNet-B4, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2, DenseNet-121, and DenseNet-201. The comprehensive and comparative evaluation with various models helps to confirm the correctness of the proposed hypothesis. Theoretically, the proposed facial image augmentation strategy will achieve the highest performance on most of the tested CNN models. This experiment also provides a perspective on how the models compare with each other in terms of the image data characteristics of the problem.

In experiments, images are normalized by transforming the pixel values before being fed into the deep learning model. This is important because it helps to stabilize the range of weights of deep learning models during training. First, the images are converted to 32-bit floating point format with the pixel value range of [0, 1]. Then, the pixel values are normalized based on the expected value and standard deviation value calculated on the ImageNet set. This helps to keep the distribution of pixel values consistent before feeding into the deep learning model.

In the evaluation phase, the trained CNN models will be evaluated on the test data. First, classification scores will be calculated and used as a criterion for comparison between testing strategies as well as between specific models to confirm the hypotheses. Second, heatmap results will be calculated for the trained models corresponding to the input images. These heatmap results will indicate the attention regions of each trained model for a specific input image. These attention regions will be discussed based on the comparison with the locations of important regions in the image, specifically the facial regions around landmarks-areas considered to represent the expressive features of human faces.

3. EXPERIMENT AND RESULTS

This section presents the details of the experiment as well as the analysis of the experimental results. First, present the experimental setup. Second, present the impact of the proposed method. Third, analyze the Grad-CAM heatmaps in relation to facial landmarks and discuss the related details.

3.1. Experimental setup

In the experiments, the models would be trained using the Adam algorithm [24]. By using a transfer learning approach, those models would be fine-tuned from pre-trained models provided by the PyTorch deep learning library at https://pytorch.org. To ensure the performance achieved for training the models, the efficient use of optimizers, learning rate schedulers, and batch sizes will be exploited in the experiments. This will effectively control the update rate of the model weights, the training time, and the generalization ability. In detail, two steps of changing the frozen state of the parameters are performed to optimize the update of parts of the model during the experiment. Furthermore, the cyclical learning rates mechanism is used [25] to help the model train faster while ensuring convergence by allowing the learning rates to grow. Besides, the batch size values are also chosen appropriately to optimize the memory of the GPU hardware. Source code is published on GitLab platform [26]. Experiments were performed on the Kaggle platform with an NVIDIA Tesla P100 GPU with 16 GB VRAM. Kaggle is a popular platform that is optimized for deep learning tasks. The classification scores used include accuracy, precision, recall, AUC, F1-score.

3.2. The impact of the proposed augmentation method

Figure 4 clearly shows that the models trained with the proposed facial image augmentation method data have significantly superior performance on all metrics. Specifically, the average accuracy value is 0.905417, the average AUC value is 0.962133, the average precision value is 0.9198, the average recall value is 0.888333, and the average F1-score value is 0.903678. Additionally, the metrics of the case using default augmentation are lower at all metric types. In detail, with average precision, this case achieves 0.87125, which is 0.034167 lower than the proposed technique. Similarly, with average AUC, this case achieves 0.950561 and is 0.011572 lower. With average precision, this case achieves 0.915678 and is 0.004122 lower. With average recall, this case achieves 0.819167 and is 0.069166 lower. With average F1-score, this case achieves 0.863209 and is 0.040469 lower. Therefore, they reflect that applying the proposed facial image augmentation method makes a significant difference.

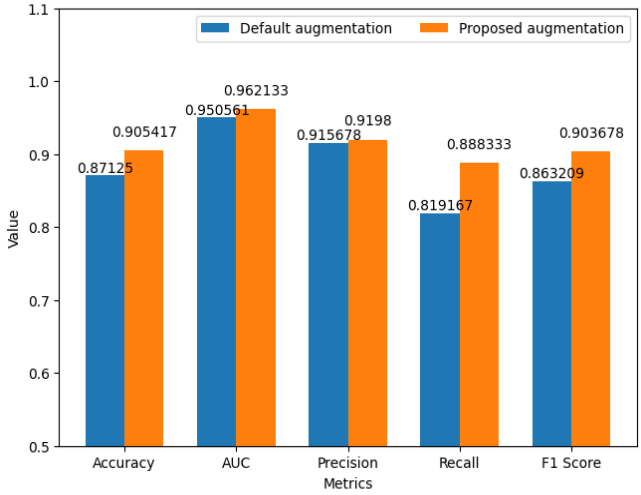


Figure 4. Average metrics for the default augmentation strategy and the proposed augmentation strategy

When the experimental results are arranged by the accuracy scale, the distribution of the locations of the models trained with the three augmentation strategies can be seen clearly. In Table 2, the scenarios trained using the proposed facial image augmentation method demonstrated high diagnostic accuracy. More specifically, among the 8 highest-accuracy training scenarios, 6 were models trained using the proposed facial image augmentation method, accounting for 75%. That means the scenarios where the model is trained with default augmentation only account for 25% of the 8 best cases in terms of accuracy.

More importantly, the cases that do not use the proposed technique not only account for a small proportion but also rank very low in the top 8 cases in terms of accuracy, namely 5th place with DenseNet-201

and 7th place with EfficientNet-B4. In detail, 7th place and 6th place have the same accuracy, but 7th place has a clearly smaller AUC. This clearly shows the disadvantage of the default augmentation compared to the proposed technique.

Table 2. The top 8 cases sorted in terms of accuracy

Strategy	Model	Accuracy	AUC	Precision	Recall	F1-score
Proposed augmentation	EfficientNet-B4	0.926667	0.967933	0.926667	0.926667	0.926667
Proposed augmentation	ResNet-18	0.916667	0.968689	0.92517	0.906667	0.915825
Proposed augmentation	EfficientNet-B0	0.916667	0.974089	0.931034	0.9	0.915254
Proposed augmentation	MobileNet-V2	0.913333	0.976022	0.930556	0.893333	0.911565
Default augmentation	DenseNet-201	0.9	0.973333	0.889610	0.913333	0.901316
Proposed augmentation	DenseNet-201	0.896667	0.971422	0.928058	0.86	0.892733
Default augmentation	EfficientNet-B4	0.896667	0.965111	0,928058	0.86	0.892734
Proposed augmentation	DenseNet-121	0.893333	0.952756	0.898649	0.886667	0.892617

Further evidence is shown in Table 3 with the 8 worst performing cases sorted by accuracy measure. Among them, the case using the proposed technique occupies only 2 positions, equivalent to 25% of the total 8 positions. These two positions belong to the ResNet-50 model with an accuracy of 0.893333 and ResNet-101 with an accuracy of 0.886667, both of which are asymptotically close to 0.9. Furthermore, both of these cases are in the upper half of the table, which means that they are in the high-accuracy region of the table. In other words, the cases that do not use the proposed technique occupy 75% of the table and are also mostly in the lower-scoring region. This is also a clear demonstration of our hypothesis in this paper.

Table 3. The bottom 8 cases sorted in terms of accuracy

Strategy	Model	Accuracy	AUC	Precision	Recall	F1-score
Proposed augmentation	ResNet-50	0.893333	0.941911	0.915493	0.866667	0.890411
Default augmentation	DenseNet-121	0.89	0.952	0.939850	0.833333	0.883392
Proposed augmentation	ResNet-101	0.886667	0.944244	0.902778	0.866667	0.884354
Default augmentation	MobileNet-V2	0.886667	0.957422	0.886667	0.886667	0.886667
Default augmentation	ResNet-18	0.863333	0.959244	0.950413	0.766667	0.848708
Default augmentation	EfficientNet-B0	0.85	0.948133	0.92	0.766667	0.836364
Default augmentation	ResNet-50	0.846667	0.931422	0.919355	0.76	0.832117
Default augmentation	ResNet-101	0.836667	0.917822	0.891473	0.766667	0.824373

Next, statistical tests using a paired t-test between pairs of models on the same test set are also included to validate improvements. The results show that the majority of models show clear and statistically significant differences, with 5 out of 8 models achieving p-value less than 0.05. Specifically, ResNet-18 results with a t-stat of 6.097847 and p-value less than 0.000001, indicating that there is a statistically significant difference between the paired groups. Similarly, EfficientNet-B0 has a t-stat of -3.274006, p-value of 0.001185, and EfficientNet-B4 has a t-stat of -2.960936, p-value of 0.003313, also demonstrating significant improvements. In addition, ResNet-101 with p-value of 0.004857 and DenseNet-201 with p-value of 0.002328 are also strong evidence against the null hypothesis. In contrast, MobileNet-V2 model with p-value of 0.212335 does not reach significance. These results confirm that there is a statistically reliable improvement in most of the test models. This also proves our research hypothesis for the proposed method.

Additionally, for the cases applying the data augmentation, the model calibration is evaluated using the expected calibration error (ECE) index. The results obtained for most models are relatively well-calibrated with quite small ECE values. The model with the highest ECE is ResNet-101 with ECE of 0.100295. The remaining models all have ECE less than 0.1. Accordingly, the ResNet-50 model achieved ECE of 0.089291, the DenseNet-201 model achieved ECE of 0.084470, the DenseNet-121 model achieved ECE of 0.081230, the EfficientNet-B0 model achieved ECE of 0.075313, the MobileNet-V2 model achieved ECE of 0.074853, the ResNet-18 model achieved ECE of 0.069943, the EfficientNet-B4 model achieved ECE of 0.064409. This also contributes to demonstrating the effectiveness of data augmentation when in the experiment, the architectures not only achieved high accuracy but also provided reliable probability estimates, which is important in the medical context.

In addition, some information about the computational time, memory usage, and training time is also provided. The experiment was performed with the Kaggle server, so the memory is allocated within the range allowed by the Kaggle server. We also provide more details about the computational time and training time for each epoch in Table 4.

Table 4. Average predicted times per sample and average training times per epoch

Model	Average predicted time per sample (seconds)	Average training time per epoch (seconds)
EfficientNet-B0	0.000149	70.790747
EfficientNet-B4	0.000610	116.379686
ResNet-18	0.000014	58.893420
ResNet-50	0.000110	88.798205
ResNet-101	0.000207	128.164967
MobileNet-V2	0.000105	69.011875
DenseNet-121	0.000588	76.951495
DenseNet-201	0.001018	111.679654

3.3. Analysis of Grad-CAM heatmaps in relation to facial landmarks

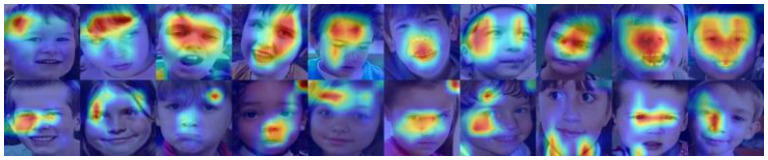
In this section, the relationship between the model's attention, represented by Grad-CAM heatmaps, and key regions in the facial image s analyzed. This will contribute to revealing the relationship between the quality of autism diagnosis from facial images and the image regions that the model focuses on. In other words, does the model's attention to key regions, which are related to facial landmarks, affect the final diagnostic performance? We start with the Grad-CAM heatmaps of the two best-performing cases, EfficientNet-B4 model and ResNet-18 model, both of which have been trained with proposed augmentation technique. The specific results are shown in Figure 5. For each model case, the results are plotted in two rows. The first row shows 10 best-performing cases, and second row shows the 10 worst-performing cases.

Figure 5 shows a striking demonstration of our prediction hypothesis, with some Grad-CAM heatmaps of the EfficientNet-B4 model in Figure 5(a) and the ResNet-18 model in Figure 5(b) using the proposed augmentation strategy. In the first row of two cases, it is easy to see that the models' attention is clearly focused on important facial regions in the correct prediction cases. Specifically, regions such as the eyes and nose are also important regions marked by landmarks commonly seen in facial expression analysis. This reflects that the models have effectively learned facial expression features to be able to recognize autism well. Theoretically, autism has atypical facial expressions, so the fact that Grad-CAM heatmaps are strongly associated with regions associated with facial landmarks suggests that the models capture meaningful features in facial images for autism diagnosis.

Oppositely, in the second row of two cases, the images of the wrong results show that the heatmaps are often not focused on important areas of the face, which are usually determined by landmarks. The results are often scattered across many areas of the face, even outside the face. This suggests that the models are capturing features that are irrelevant or have little to do with facial expression features that are important for autism. This could therefore be an explanation for why the results are poor or unstable.

In Figure 6, Grad-CAM heatmaps of the two worst-performing cases are presented, the ResNet-50 model in Figure 6(a) and the ResNet-101 model in Figure 6(b), both of which have been trained with default augmentation techniques. Similarly, for each model case, results are plotted in two rows. The first row shows the 10 best-performing cases, and the second row shows the 10 worst-performing cases.

Figure 6 also shows evidence for our prediction hypothesis. In each case, the first row with the best results consistently exhibits a stronger focus on important facial regions than the second row with the worst results. This also helps answer the question of the relationship between the autism diagnosis performance of the deep learning model and the model's focus on important facial regions, which are regions that are understood to be marked by landmarks commonly used in facial expression studies.



(a)

(b)

Figure 5. Some Grad-CAM heatmaps using the proposed augmentation strategy for (a) the EfficientNet-B4 model and (b) the ResNet-18 model

(a)

(b)

Figure 6. Some Grad-CAM heatmaps using the default augmentation strategy for (a) the ResNet-50 model and (b) the ResNet-101 model

However, the two cases in Figure 6 are the two worst-performing cases, so even in the best results there are cases where the model's attention is distracted. In the case of ResNet-50, in the first row, the heatmap color regions are quite spread out. Meanwhile, in the case of ResNet-101, which is the worst case, even in the first row, including the models with the highest diagnostic performance, there are cases where the heatmap is concentrated outside the face image region. This may also be part of the answer to why the worst performance belongs to the case of ResNet-101 model in the default augmentation choice. This further demonstrates the importance of focusing the model's attention on the main face region and the diagnostic quality of deep learning models.

To illustrate how the model responds to small changes in the face, image data consisting of two rows is introduced. The first row is the pairs of original images and landmark-based transformed images. The second row is the Grad-CAM maps corresponding to the images of the first row. The details are shown in Figure 7. In Figure 7, on the one hand, the results show that the transformations change the heatmap distribution, and this will further affect the final prediction results. On the other hand, the main regions of attention on the deformed images still maintain an intersection area. This also partly demonstrates the model's ability to learn from our data augmentation data and maintain a certain stability of attention to important feature regions.

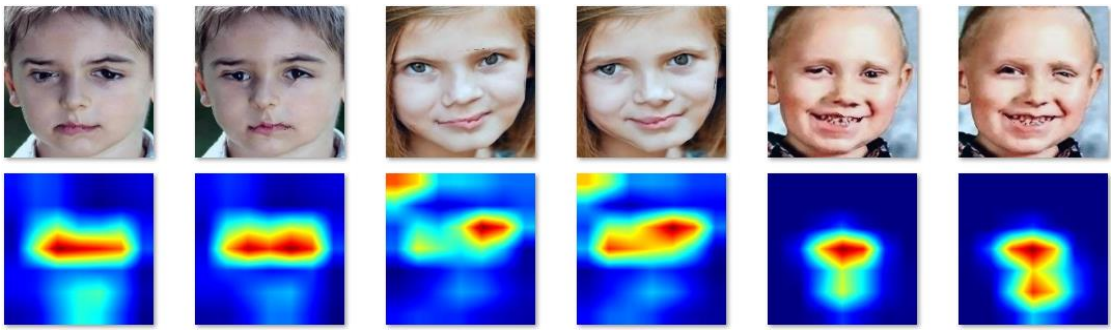


Figure 7. Some saliency maps relating to changes in the image affect predictions

Next, Grad-CAM alignment with facial landmarks is also quantified using attention overlap assessment. Specifically, the attention overlap score is calculated as the overlap ratio between the Grad-CAM focus area and each landmark area, defined as a circular range surrounding each landmark. The analysis is performed on seven main areas of the face. In which, the jawline includes landmarks from 0 to 16, the right eyebrow includes landmarks from 17 to 21, the left eyebrow includes landmarks from 22 to 26, the nose includes landmarks from 27 to 35, the right eye includes landmarks from 36 to 41, the left eye includes landmarks from 42 to 47, and the lips include landmarks from 48 to 67. The details are shown in Table 5.

Table 5. Attention overlaps scores on main regions of the face

Facial region	Landmarks	Default augmentation	Proposed augmentation	Percent change (%)
Jawline	0-16	0.002807	0.002596	-6.99991
Right eyebrow	17-21	0.003634	0.003905	7.315448
Left eyebrow	22-26	0.003805	0.004363	14.99447
Nose	27-35	0.004919	0.004934	0.578789
Right eye	36-41	0.004125	0.004498	9.00641
Left eye	42-47	0.004291	0.004863	13.39065
Lip	48-67	0.00473	0.004305	-9.06066

The results show that the rate of change of attention overlap score after applying our data augmentation by landmark deformation is different between regions. Many regions show an increase in attention overlap score when comparing the proposed augmentation technique with the default case. Notably, the two eye regions have the strongest increase, in which the strongest increase index is 14.99% of the left eyebrow region. Besides, there are also regions that record a decrease in attention overlap score, which are the jawline and lip regions, when the focus of the model has shifted to another range. This also reflects that the autism classification model in this study has learned many expression features from important regions of the face, and especially the eye region which contains many distinctive features related to gaze communication.

Our study on autism detection from child facial images has highlighted the role of facial landmarks in child facial analysis. This is similar to some previous studies such as [16] with the infant face (INFACE) model. While in the context of autism, landmarks help the model learn features related to facial expressions and morphology, in INFACE, landmarks are exploited to build a 3D shape model and disentangle expression and identity. Another similar work is [17] with the exploitation of facial landmarks for clinical diagnosis. Specifically, in the study, the authors used landmarks to assess mandibular deformities. This approach is also similar to our study with the use of landmarks to exploit facial expression features in children. All these approaches show that landmarks are not only geometric reference points, but also have biological and clinical relevance in children.

4. CONCLUSION

In this paper, the focus is on analyzing the influence of facial landmark features within the framework of building a comprehensive and effective solution for autism diagnosis from child facial image data. The data used in this study is collected from the Kaggle platform, including normal and autistic child facial images. A facial augmentation algorithm based on facial landmark transformation was proposed, and the effectiveness of the method was confirmed through comprehensive comparative evaluation with well-known models such as EfficientNet-B0, EfficientNet-B4, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2, DenseNet-121, and DenseNet-201. In addition, visual evidence and discussion based on Grad-CAM heatmaps of model facial image regions are provided. Although good results were achieved in terms of accuracy, AUC, precision, recall, and F1-score, many issues still need to be addressed to create practical autism diagnosis product. The hypothesis regarding the role of facial expression features and their correlation with the performance of popular deep learning models applied to autism recognition tasks was clarified. This confirms that facial expression features are important research issue in facial image data analysis in general and autism diagnosis in particular. One challenge when working with small datasets is domain specificity and difference from original data used to train pre-trained models. Models are often trained on large-scale natural image datasets such as ImageNet, while research data are problem-specific. This difference may cause learned features to transfer poorly to target domain. One solution is domain-based data augmentation rather than general augmentation. Contribution of this study focuses on development of landmark-based augmentation as novel artificial intelligence technique, improving model generalization and interpretability. This approach improves training efficiency and provides clearer understanding of decision-making basis for autism diagnosis. Another aspect to consider is ethical implications of medical research. Artificial intelligence applications must support early diagnosis while respecting privacy and avoiding stigmatization. Positioning artificial intelligence as tool to support doctors helps harness benefits without compromising medical humanities. In future work, integrating attention mechanisms or transformer-based models for autism recognition from facial images will be considered. These architectures can focus on important facial regions, enabling interaction modeling across facial structure, capturing nuanced facial patterns with improved accuracy and interpretability in medical application context.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nguyen Trong Tung		✓	✓			✓	✓	✓	✓		✓			✓
Ngo Duc Vinh	✓				✓	✓				✓			✓	
Ha Manh Toan	✓		✓				✓	✓	✓					
Do Nang Toan	✓	✓		✓	✓					✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com.

REFERENCES

[1] WHO, "Autism," *World Health Organization*. Accessed: Sep. 20, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[2] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Translational Pediatrics*, vol. 9, no. S1, pp. S55–S65, Feb. 2020, doi: 10.21037/tp.2019.09.09.

[3] S. B. Sulkes, "Autism spectrum disorder," *MSD Manual*. Accessed: Sep. 10, 2025. [Online]. Available: https://www.msdmanuals.com/professional/pediatrics/learning-and-developmental-disorders/autism-spectrum-disorders

[4] J. Baio *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2014," *MMWR Surveillance Summaries*, vol. 67, no. 6, pp. 1–23, 2018, doi: 10.15585/mmwr.ss6706a1.

[5] A. J. Baxter, T. S. Brugha, H. E. Erskine, R. W. Scheurer, T. Vos, and J. G. Scott, "The epidemiology and global burden of autism spectrum disorders," *Psychological Medicine*, vol. 45, no. 3, pp. 601–613, 2015, doi: 10.1017/S003329171400172X.

[6] MOLISA, "Respond to the world autism awareness day: stronger engagement of the community in supporting children with autism," *Ministry of Labour, Invalids and Social Affairs*. Accessed: Sep. 10, 2025. [Online]. Available: https://english.molisa.gov.vn/topic/231259

[7] I. A. Ahmed *et al.*, "Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques," *Electronics*, vol. 11, no. 4, 2022, doi: 10.3390/electronics11040530.

[8] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018, doi: 10.1016/j.nicl.2017.08.017.

[9] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-35910-1.

[10] S. M. M. Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, no. 2, pp. 15038–15057, 2023, doi: 10.1109/ACCESS.2022.3232490.

[11] V. S. S. Karri, S. Remya, A. R. Vybhav, G. S. Ganesh, and J. Eswar, "Detecting autism spectrum disorder using DenseNet," in *ICT Infrastructure and Computing*, Singapore: Springer, 2023, pp. 461–467, doi: 10.1007/978-981-19-5331-6_47.

[12] T. M. Ghazal, S. Munir, S. Abbas, A. Athar, H. Alrababah, and M. A. Khan, "Early detection of autism in children using transfer learning," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 11–22, 2023, doi: 10.32604/iasc.2023.030125.

[13] Y. Li, W.-C. Huang, and P.-H. Song, "A face image classification method of autistic children based on the two-phase transfer learning," *Frontiers in Psychology*, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1226470.

[14] I. Ahmad, J. Rashid, M. Faheem, A. Akram, N. A. Khan, and R. ul Amin, "Autism spectrum disorder detection using facial images: a performance comparison of pretrained convolutional neural networks," *Healthcare Technology Letters*, vol. 11, no. 4, pp. 227–239, 2024, doi: 10.1049/htl2.12073.

[15] P. Reddy and J. Andrew, "Diagnosis of autism in children using deep learning techniques by analyzing facial features," in *Engineering Proceedings*, 2023, vol. 59, no. 1, doi: 10.3390/engproc2023059198.

[16] T. N. Schnabel *et al.*, "Large-scale 3D infant face model," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Cham, Switzerland: Springer, 2024, pp. 217–227, doi: 10.1007/978-3-031-72384-1_21.

[17] X. Xu *et al.*, "DiRecT: diagnosis and reconstruction transformer for mandibular deformity assessment," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 2024, pp. 141–151, doi: 10.1007/978-3-031-72384-1_14.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 17, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[19] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational geometry: algorithms and applications*. Berlin, Heidelberg: Springer, 2008, doi: 10.1007/978-3-540-77974-2.
[20] Kornia, "Geometric computer vision library for spatial AI," *GitHub*. Accessed: Oct. 20, 2025. [Online]. Available: https://github.com/kornia/kornia/
[21] B. Koonce, "EfficientNet," in *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, California: Apress, 2021, pp. 109–123, doi: 10.1007/978-1-4842-6168-2_10.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[23] A. G. Howard *et al.*, "MobileNets: efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.
[24] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.
[25] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay," 2018, *arXiv:1803.09820*.
[26] H. M. Toan, "autism-dlm," *GitLab*. Accessed: Oct. 20, 2025. [Online]. Available: https://gitlab.com/hamanhtoan/autism-dlm

BIOGRAPHIES OF AUTHORS

Nguyen Trong Tung graduated from the University of Danang in 2010 with a master's degree in Computer Science. He has 20 years of experience working in education, with special interests in computer vision, deep learning, and cloud computing. He has published numerous papers on deep learning, resource allocation in distributed systems. Now he is a lecturer at Dong-A University, Danang, Vietnam. He joined the Ph.D. program at Ho Chi Minh City Open University in 2024. He has participated in many cooperation programs in the field of training, international conferences. He can be contacted at email: tungqn@donga.edu.vn.

Ngo Duc Vinh received his master of IT in 2006 from the Military Technical Academy, and his Ph.D. in 2016 at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology. Currently, he is a lecturer at Hanoi University of Industry. His research interests involve computer vision, image processing, and machine learning. He can be contacted at email: ngoducvinh@haui.edu.vn.

Ha Manh Toan learned applied mathematics and informatics at the College of Science, Vietnam National University, Hanoi, and received a degree in 2009. In 2015, he earned an M.Sc. degree at the University of Engineering and Technology, Vietnam National University, Hanoi. Now, he is a researcher at the Vietnamese Academy of Science and Technology. His studies interests relate to machine learning, computer vision, and deep learning. He can be contacted at email: hmtoan@ioit.ac.vn.

Do Nang Toan studied Applied Mathematics and Informatics at Hanoi University and received a degree in 1990. In 2001, he earned a Ph.D. degree at the Vietnam Academy of Science and Technology. Now, he is an associate professor at the Vietnamese Academy of Science and Technology. His studies interests relate to machine learning, computer vision, and virtual reality. He can be contacted at email: dntoan@ioit.ac.vn.