

Detecting autism with Vietnamese child facial images using deep learning

Tran Van Thanh¹, Lam Thanh Hien², Do Nang Khoa³, Le Anh Tu⁴, Ha Manh Toan⁵, Do Nang Toan⁵

¹Faculty of Mechatronics and Electronics, Lac Hong University, Bien Hoa, Vietnam

²Faculty of Information Technology, Lac Hong University, Bien Hoa, Vietnam

³Faculty of Information Technology, Hanoi Architectural University, Hanoi, Vietnam

⁴Faculty of Information Technology, Ha Long University, Uong Bi, Vietnam

⁵Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

Article Info

Article history:

Received Feb 12, 2025

Revised Jun 5, 2025

Accepted Jul 3, 2025

Keywords:

Autistic detection

Convolutional neural network

Generalization

Transfer learning

Vietnamese facial child image

ABSTRACT

Deep learning techniques created a significant increase in intelligent systems, especially in the medical field. Among mental problems, autism is a dangerous neurodevelopmental disorder and it needs to be diagnosed early because of the malleability of child brain development. In our study, we focused on autism detection by using the Vietnamese facial child image and studied the role of international data and Vietnamese data when applying deep learning approach to diagnose autism. To do that, we proposed different strategies based on our hypothesis about factors of the transfer learning and training set types. To conduct the experiment, we prepared a Vietnamese facial child image set from several kindergartens in Ho Chi Minh City, Vietnam and we applied different deep architectures such as ResNet, DenseNet, and AlexNet in the autism classification experiment with both Vietnamese and international facial child images. We analyzed important factors from the experiment results with area under the curve (AUC), accuracy, sensitivity, and specificity, including applying transfer learning and the appearance of Vietnamese data in the training set. Besides, we also discussed the difference of international and Vietnamese data domains. The exposure of data distribution differences in the proposed strategies also highlights the importance of collecting facial data of Vietnamese children.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Do Nang Toan

Institute of Information Technology, Vietnam Academy of Science and Technology

18 Hoang Quoc Viet, Cau Giay District, Hanoi 10072, Vietnam

Email: dntoan@ioit.ac.vn

1. INTRODUCTION

Autism is a neurodevelopmental disorder characterized by a lack of social communication and repetitive behaviors, both verbal and physical [1]. Autistic syndrome has difficulty interacting and communicating with others. Symptoms of autism can occur in many situations and in many places' different places. This disease involves abnormal brain development that affects facial expressions and physical states [2]. Children with autism have unusual facial features, which sets them apart from typically developing children. This disease not only causes difficulties for the patient but also for their loved ones in daily life [3]. For parents who have had a child with autism, the risk of having another child with autism is approximately 3 to 10% [4]. There have been recent concerns about rising autism rates. The World Health Organization estimates that the prevalence of autism is 0.67% worldwide, but this figure is only reported in 16% of children globally [5]. The centers for disease control and prevention estimates that about 1.68% of US children aged 8 years and older are diagnosed with autism, which is 1 in 59 children [6].

Early diagnosis of autism is essential. Children showing signs of autism can benefit from early diagnosis due to the malleability of their brain development, which can help them improve their social life. Late diagnosis can reduce the effectiveness of treatments for patients with autism [7]. According to a recent study, children who received medical care before the age of two had a higher intelligence quotient than children who did not receive medical care until later in life [8].

There have been several artificial intelligence studies relating to autism diagnosis. In 2018, Heinsfeld *et al.* [9] studied autism classification with brain magnetic resonance imaging data. In the study, the authors used a custom deep neural network model combined with two stacked denoising autoencoders and experiments using 10-fold cross-validation. Ahmed *et al.* [10] had research aiming at diagnosing autism with various tasks such as removing all noise from the eye path area, extracting the path of eye points on the image, building a classification model, and evaluating. To perform these tasks, the authors used several techniques such as local binary samples, and gray-level co-occurrence matrices. Besides, the authors also conducted experiments on some deep learning models such as Google-Net and ResNet-18. Farooq *et al.* [11] presented a study using federated learning to diagnose autism. In their work, the authors applied two traditional machine learning models, logistic regression and support vector machine, to data from various sources. In the experiment, the authors used more than 600 tabular data records and achieved an accuracy of 98% and 81% for two groups of children and adults, respectively.

In Vietnam, there have been several attempts by the government and community relating to autism. In 2022, the Ministry of Labor - Invalids and Social Affairs announced news responding to world autism awareness day [12]. That news presented a program in Ho Chi Minh City that allows the community to get a better understanding of autism and support autistic children. In this paper, our study is put in the context of autism in Vietnamese children. We aim to autism detection in Vietnamese children with modern deep learning techniques by using facial image data that can be captured by using a normal camera in smartphones.

Several studies were using deep learning methods for autism detection. Karri *et al.* [13] published a study using a deep-learning approach to classify autism based on facial images. In the experiment, the authors used the deep learning model Dense-Net along with a data set of face images downloaded from the Kaggle platform. Rabbi *et al.* [14] presented the work for autism classification tested with a dataset of 2,936 images provided by the Kaggle platform. In that study, the authors applied the transfer learning approach to VGG 19, Inception V3, and Dense-Net 201 models and used accuracy, precision, recall, F1-score, and AUC (Area under the ROC curve) to evaluate their results. Li *et al.* [15] presented a study using the transfer learning approach with MobileNetV2 and MobileNetV3-Large networks to identify autistic children based on facial images. The authors built a framework of facial image classification combining the two-phase transfer learning and the multi-classifier integration. In the experiment, their work reached 0.8833 accuracy for MobileNetV2 and 0.8767 accuracy for MobileNetV3-Large. Another study published in 2023 by Ghazal *et al.* [16]. The authors proposed a convolutional neural network based on Alex-Net with facial images as the input. Their study aimed at the robust extraction of numerous facial features, which is a difficult task because of the subtlety requirements. In the experiment, that work reached a validation accuracy of 0.877, a validation sensitivity of 0.876, and a validation specificity of 0.876. Reddy and Andrew [17] published a paper on using the deep learning approach to identify autistic children. The authors used three models VGG16, VGG19, and EfficientNetB0 with pre-trained weights in the framework of applying transfer learning. From the experiments, the authors achieved an accuracy of 84.66% for the VGG16 model, 80.05% for the VGG19 model, and 87.9% for the EfficientNetB0 model. Also, Ahmad *et al.* [18] published a paper using different models such as AlexNet, MobileNetV2, ResNet34, ResNet50, VGG16, and VGG19 to detect autism from facial images and analyze their results. In the experiment, the training time was approximately 2 hours, and the testing time was nearly 3 minutes. In results, ResNet34 reached an accuracy score surpassing 0.86 with 248x248 resolution and 0.83 with 124x124 resolution.

An important issue we want to explore is the difference in the data domain of Vietnamese and international child facial images. Our hypothesis is that this difference will have a significant impact and therefore the presence of Vietnamese child facial image data in the steps of developing deep learning models for this problem. In fact, many studies on facial morphology and anatomy show that Vietnamese faces have distinct characteristics. The study [19] compared Vietnamese people with North American white people and showed that Vietnamese people have a wider distance between the two eye corners, a larger nose, but a smaller mouth width. Compared with other Asian groups, research [20] shows that Chinese people tend to have smaller foreheads, higher noses, and a smaller ratio between nose length and face height than Vietnamese people.

In our research, we aim to apply a deep learning approach to effectively classify autism from Vietnamese children's data. On the one hand, we often need a large enough data set for effective training in deep learning methods. On the other hand, collecting a large enough Vietnamese children dataset is not easy in real-life conditions. The access requires the permission and support of the children's parents as well as the

support of the teachers at the kindergarten. Therefore, we designed strategies relating to applying international data in generalizing Vietnamese facial images. Besides, because our study aims to build a state-of-the-art solution for the autism detection problem in Vietnamese children's facial images, we also collected data from several local Vietnamese kindergartens that have both autistic children and normal children learning. The Vietnamese dataset would play an important role in evaluating our different strategies. In detail, we design different strategies regarding how to apply international data to the problem of autism diagnosis on Vietnamese children's facial images. Based on separate evaluation evidence on international and Vietnamese data, we delve into the important factors to build an effective solution such as how to use pre-trained weights as well as the role of Vietnamese data in the training phase. The evidence also reveals the limitations of existing international data in generalizing Vietnamese facial images.

In short, the main contributions were described as follows. Firstly, we set up a ready-to-use Vietnamese children's facial image dataset for the autism detection. Secondly, we proposed different strategies relating to applying international data in generalizing Vietnamese facial images. Lastly, we analyzed the experiment results of the strategies to deeply delve into important factors such as options of using pre-trained weights and the difference of international and Vietnamese data domains.

The structure of the paper is presented as follows: Section 2 is the background of deep-learning architecture models for the proposed system. Section 3 shows the proposed study. Section 4 describes and analyzes the results of the experiment. In the end, section 5 concludes our study.

2. DEEP-LEARNING ARCHITECTURES

The proposed method needs some deep-learning architecture models. To conduct experiments, we used some popular deep-learning models such as Res-Net 34, Res-Net 50, Alex-Net, and Dense-Net. These models were applied successfully in a lot of artificial intelligence research.

Krizhevsky *et al.* [21] presented a new deep architecture and this model won the image net large scale visual recognition challenge 2012. The authors reached a top-5 error result of 15.3% and this figure was far higher than the second group, which reached only 26.2%. In architecture, the Alex-Net model consists of five convolution layers followed by three fully connected layers and the end is the SoftMax function. In the medical field, Alex-Net was also used for some research. For example, Mohi ud Din and Jayanthi [22] published a study of classifying autism using Electroencephalography signals in 2022. Studies [16], [18] also used Alex-Net.

He *et al.* [23] had a study about training deep networks to solve the vanishing gradient problem. In their paper, their idea was implemented in the residual block model of Res-Net architecture. With that architecture, the authors won the image net large scale visual recognition challenge 2015 and reached a classification error of 3.57%. Since then, this architecture has become famous in the deep-learning field and has had many variants with different depths. In medical research, Res-Net is applied to many image classification problems, such as [10], [18].

Huang *et al.* [24] proposed Dense-Net with the idea of connecting densely between layers. Their idea was implemented in the Dense block model and was rather similar to Res-Net in the context of the vanishing gradient problem. In medical research, Dense-Net is applied to image classification problems, such as classifying autism [13], [14].

3. PROPOSED STUDY

3.1. Dataset preparation

Our study is designed to focus on the Vietnamese context. One important task is the preparation of a Vietnamese image dataset for the experiment. In this study, all facial images were captured with the support of the camera on the smartphone.

We collected data at several kindergartens in Ho Chi Minh City, Vietnam. Data were collected from some children with autism and some normal children. The collection process was carried out with the support of parents. Accordingly, parents will directly record images of their children's faces using smartphones or parents will babysit the children while someone else records the images. Collected images will be selected to get qualified images. Accordingly, we will remove images in which the child's face is blurred, partially obscured due to the child's facial pose, or obscured by other objects or body parts such as a hand.

To perform data labeling, we cropped the qualified face image areas and arranged them into two groups corresponding to the status of children with autism or normal. To do this, we built a program that automatically localizes the face in the image, and based on that result, we performed operations on the program interface to be able to correct the face area if necessary. Finally, we would annotate the states corresponding to the two groups and the program will show the final resulting images.

In summary, we have prepared a data set of 892 images of Vietnamese children's faces. Among them, there are 444 photos of children with autism and 448 photos of normal children. All of these images were used in the experimental step of this study.

Besides the data of Vietnamese children, we use an international dataset published on the Kaggle platform [25]. This dataset includes 2,936 images of children's faces that are also divided into two groups according to autism criteria. In the dataset, there are 1,468 images marked as being of children with autism and 1,468 images of the opposite case.

To experiment in this study, we would prepare training, validation, and test datasets based on the Vietnamese dataset and the international dataset. So, both the Vietnamese dataset and the international dataset were split. The international dataset was already split. In detail, from the international dataset, the test set has 300 samples, the validation set has 100 samples, and the training set has 2,536 samples. For the Vietnamese dataset, we set up the test set with 90 samples, the validation set with 89 samples, and the training set with 713 samples.

3.2. Proposed strategies relating to apply international data in generalizing to Vietnamese data

To apply international data to the problem of diagnosing autism in Vietnamese children, we analyze the use of the transfer learning approach. Transfer learning is a popular approach to problems using deep learning models. In which, the application of transfer learning will be effective if the domain of the data used for pretraining and the domain of the data in the target problem are similar and this application will be ineffective if the two data domains are sufficiently different. For the problem of diagnosing autism from facial images of Vietnamese children, we have some observations as follows. On the one hand, autism is a specific problem with distinct semantics. That makes it possible to use a model trained from scratch. On the other hand, the data used is facial image data. This is a type of data that has been used in many other problems such as facial recognition, emotion recognition, and gender recognition. From this perspective, the application of the transfer learning approach is feasible. From our perspective, it makes sense to apply a transfer learning approach. To further clarify this, in our experiments, we will train the model from scratch and train the model with pre-trained weights. The choice used is the pre-trained model weights provided by the PyTorch platform [26]. These are models that have been pre-trained on the popular ImageNet dataset.

Digging deeper into the problem, we found that with autism diagnosis based on children's facial images, the signs of autism will be expressed through children's facial expressions. Therefore, it can be said that the data domain of this problem will have a certain similarity with the data domain of the problem of recognizing children's facial expressions. On this basis, we further designed the application of transfer learning using a pre-trained model on the children's facial expression dataset. Our experimental choice is to use pre-trained models on the facial expression recognition (FER) children's facial expression dataset [27]. We hypothesize that the model pre-trained on the FER children's facial expression dataset will have a higher fit than the model pre-trained on the ImageNet dataset.

Thus, in designing strategies to test the transfer learning application, we use three different options. Specifically, the option with models trained from scratch, the option with models pre-trained on ImageNet and available on the PyTorch platform, and the final option is models pre-trained on the FER child facial expression dataset. The trained models will be evaluated on both the international dataset and the Vietnamese dataset for the autism diagnosis from child facial images. According to our hypothesis, the order of performance of the cases will be similar on both test sets with the best results belonging to the models pre-trained on the child facial expression dataset and the worst results belonging to the models trained from scratch.

Another option we have set out to apply international data to the problem of diagnosing autism in Vietnamese children is to combine international data and Vietnamese data into the training set as described in Table 1. Using this option will give us a clear assessment of the difference between the two data domains, Vietnamese children's facial images and international children's facial images, in the problem of diagnosing autism from children's facial images. Accordingly, we design strategies with training data options including the first option of training only with international data and the second option of training with data combining international data with Vietnamese data. Similarly, the trained models will also be evaluated separately on the international data set and the Vietnamese data set. According to our hypothesis, the evaluation results on the international data set will not change much due to the participation of international data in the training set. However, we believe that the evaluation results on the Vietnamese dataset will have a big difference. In particular, the results of models trained on the combined dataset of international data and Vietnamese data will be higher than the results of models trained only on international data due to the difference in data domain. The strategies are described in detail as follows:

In Figure 1, Strategy 1A trains deep learning models from scratch on international data, and strategy 1B also trains deep learning models from scratch but on combined data. Accordingly, the model weights will be randomly initialized. Training from scratch can be considered because autism is a rather specialized

content. However, according to our hypothesis, the training results from scratch will likely be worse than using transfer learning within the experimental scope set. The evaluation results on international and Vietnamese test data sets of the two strategies are expected to also reflect the difference in data domains on the two sets.

Next in Figure 2, strategy 2A trains deep learning models on international data with parameters learned by transferring from pretrained models with ImageNet and strategy 2B is similar but performs training on combined data. Accordingly, the model weights will be taken from models provided by the PyTorch platform and trained on the ImageNet dataset. Transfer learning is performed on the assumption that the autism data used here is facial image data - a fairly common type of image data. According to our hypothesis, the evaluation results of the two strategies will be better than those of strategies 1A and 1B in case of the same training set. Similarly, the evaluation results on the international and Vietnamese test sets of these two strategies are also expected to reflect the difference in data domains on the two sets.

Finally in Figure 3, strategy 3A trains deep learning models on the international dataset with parameters learned by transferring from models pre-trained on the FER dataset, and strategy 3B is similar but trains on the combined dataset. Accordingly, we will train the pre-models on the FER dataset with facial expression labels. Transfer learning here is performed on the assumption that autism features on children's facial images will also be expressed through facial expressions. According to our hypothesis, the evaluation results of these two strategies will be best when considered in the case of the same training set. Similar to the previous cases, the evaluation results on the international and Vietnamese test sets of these two strategies are also expected to reflect the difference in the data domains on the two sets.

Table 1. International data and combined data in training phase

Case	International data	Vietnamese data
International training set	2536	0
International validation set	100	0
Combined training set	2536	713
Combined validation set	100	89

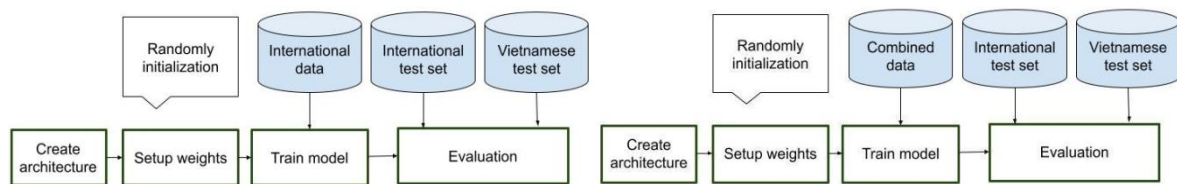


Figure 1. Strategy 1A (left) and strategy 1B (right)

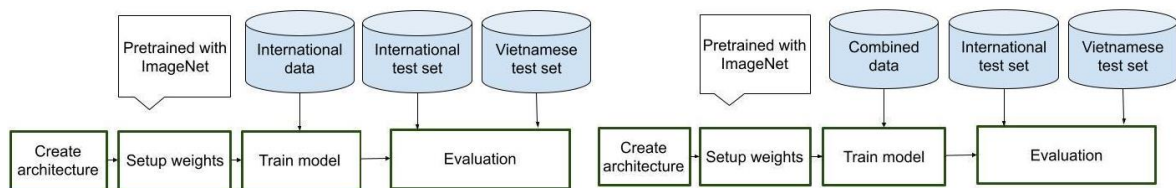


Figure 2. Strategy 2A (left) and strategy 2B (right)

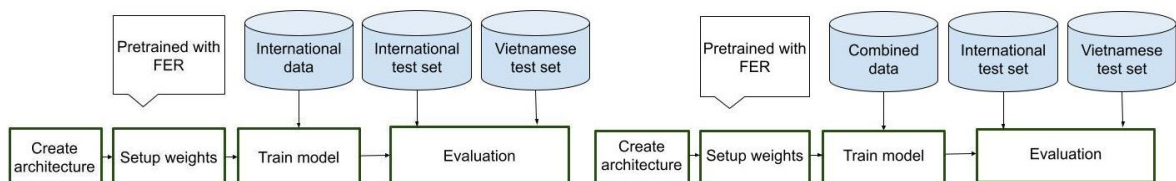


Figure 3. Strategy 3A (left) and strategy 3B (right)

4. RESULTS AND EVALUATION

4.1. Performance measures

In this study, we used photos of children's faces to assess autism status. Theoretically, this is a 2-class classification problem. With a photo, the program would conclude "Normal" or "Autism". For evaluation, we used sensitivity, specificity, accuracy, and AUC. AUC is the abbreviation for the area under the receiver operating characteristic curve, which is a graph that shows the relationship between True Positive Rate and False Positive Rate over a set threshold. AUC is a measure of the overall quality of a binary classifier. For the others, sensitivity and specificity are two commonly used measures in medical research. These measures along with the accuracy are calculated by combining predicted and reference values through the values: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The calculation is performed as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity} = \frac{FP}{FP + TN} \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

4.2. Experiments and results

This section would clarify our hypotheses relating to apply international data to the problem of diagnosing autism in Vietnamese children by discussing the results obtained from the experiments. In the training phase, the models are learned using the training and validation sets as described above. The learning process will be stopped based on the evaluation of the measured score from the validation set. More specifically, the accuracy score will be used as an indicator. After being trained, the models will be put into the evaluation step with test sets. Through these metrics, we will be able to analyze and discuss the effectiveness of the designed strategies and highlight important characteristics. To experiment, we use the Google Colab deep learning server supported by the Nvidia graphics processing unit (GPU) computation power. The test program is built on the Python programming language and is supported by the PyTorch deep learning library. Four models ResNet34, ResNet50, Alex-Net, and DenseNet121 were chosen for each strategy. After that, the trained models of strategies were tested on the test sets of the Vietnamese dataset and the international dataset.

To have a fair and overall assessment of the quality of the strategies, we evaluate the average of each measure from the four models. From Figure 4, it is clear that the average scores for all four measures, including accuracy, sensitivity, specificity, and AUC, are in the same order. Specifically, strategy 1A with training deep learning models from scratch has the worst average scores, and strategy 3A with transfer learning from pre-trained weights on the FER facial expression dataset achieves the best results. Specifically, the strategy 3A when evaluated with the international test dataset achieved an average accuracy of 0.870833, an average sensitivity of 0.856667, an average specificity of 0.885 and an average AUC of 0.943178. The results of such strategies also accurately reflect our hypothesis about initializing the weights of deep learning models before training. Besides, strategy 3A also outperforms with an average true positive of 128.5, an average true negative of 132.75, an average false positive of 17.25, and an average false negative of 21.5.

Figure 5 shows the average scores on the test set of the international data set with strategies training on the combined data. When looking at the evaluation results of these strategies, we also see a clear similarity to the results in Figure 4. Training the deep learning models from scratch gives the worst results while training with transfer learning from the training weights with FER gives the best results. Specifically in this case, strategy 3B achieved an average accuracy of 0.870833, an average sensitivity of 0.866667, an average specificity of 0.875 and an average AUC of 0.939244. Similarly, strategy 3B also shows effectiveness with average true positive and average true negative being higher than the other strategies while average False Positive and average false negative being the lowest among the strategies. Specifically in this case, strategy 3B achieved an average true positive of 130, an average true negative of 131.25, an average false positive of 18.75, and an average false negative of 20. Therefore, the hypothesis of the feasibility of using facial expression data was clearly demonstrated through the evaluation results of strategies 3A and 3B on the international test set.

Another point to note is that although the strategies using pre-trained weights with FER have better results than the strategies using pre-trained weights with ImageNet in each respective case, the difference in results is not large. This also reflects that both options are feasible. We observe more specifically about the difference in results in Table 2. Table 2 shows the difference in the evaluation data on the international test set between the strategies using pre-trained weights on the facial expression dataset FER and pre-trained

weights on the general dataset ImageNet in each case of the training data. Obviously, the difference in results is not large. Specifically, when comparing strategy 3A with strategy 2A, the results of 3A are slightly higher than strategy 2A with an increase in average accuracy of 0.011666, an increase in average sensitivity of 0.003334, an increase in average specificity of 0.02, and an increase in average AUC of 0.0121. Similarly, when comparing strategy 3B with strategy 2B, the results of 3B also have a slight increase compared to strategy 2B with an increase in average accuracy of 0.016666, an increase in average sensitivity of 0.023334, an increase in average specificity of 0.01, and an increase in average AUC of 0.007911. Therefore, in our opinion, when conducting future studies, both of the above cases are candidates.

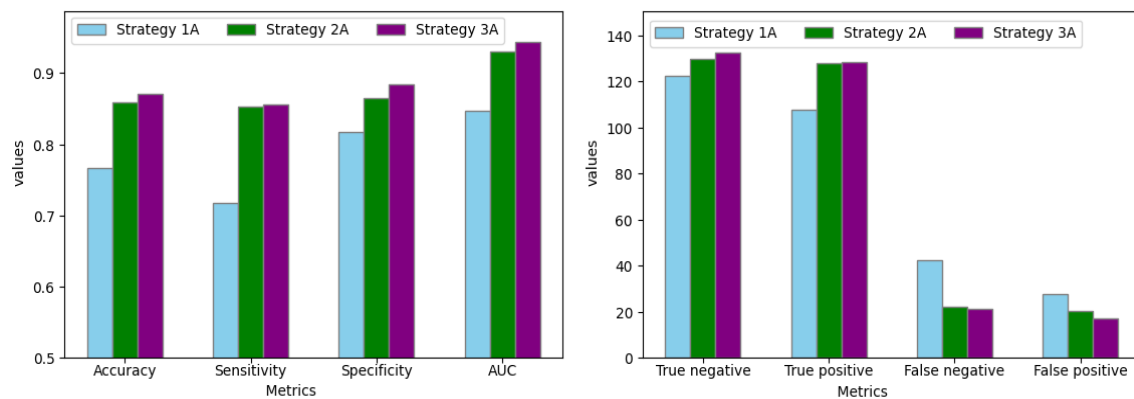


Figure 4. Average scores of strategies 1A, 2A, and 3A on the international test set: accuracy, sensitivity, specificity, AUC (left) and true positives, true negatives, false positives, and false negatives (right)

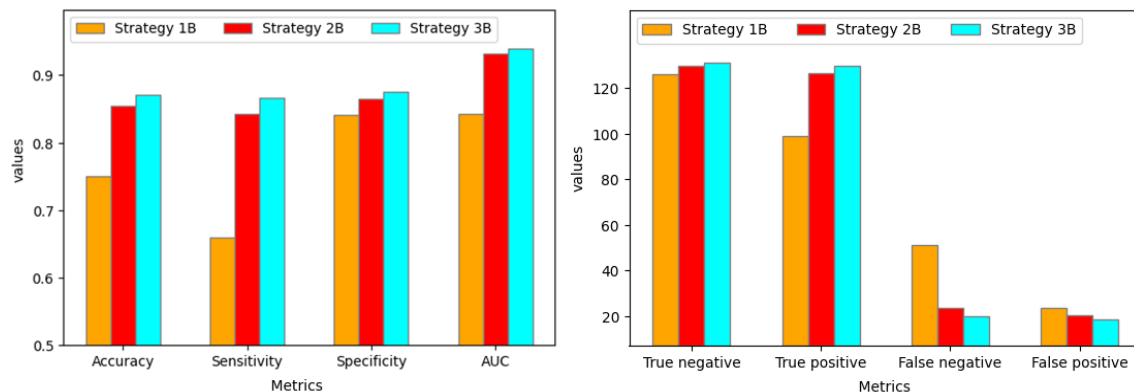


Figure 5. Average scores of strategies 1B, 2B, and 3B on the international test set: accuracy, sensitivity, specificity, AUC (left) and true positives, true negatives, false positives, and false negatives (right)

Table 2. Compare the average scores of strategies 2A with 3A and 2B with 3B on the international test set

Case	Average accuracy	Average sensitivity	Average specificity	Average AUC
Strategy 2A	0.859167	0.853333	0.865	0.931078
Strategy 3A	0.870833	0.856667	0.885	0.943178
Compare 3A to 2A	+0.011666	+0.003334	+0.02	+0.0121
Strategy 2B	0.854167	0.843333	0.865	0.931333
Strategy 3B	0.870833	0.866667	0.875	0.939244
Compare 3B to 2B	+0.016666	+0.023334	+0.01	+0.007911

Next, we see the average scores on the test set of Vietnamese data set. Figure 6, we can clearly see that the test results on the Vietnamese test set are clearly low when the models are trained purely on the international data set. Specifically, the highest average accuracy value is 0.413889, the highest average specificity value is 0.05, and the highest average AUC value is 0.264938. Although the highest average sensitivity value is 0.8, because the average specificity is too low, the models after training are biased

towards one class label and ignore the other class label, so the significance of the average sensitivity value is not large. Such low results do not reflect the advantages and disadvantages of the weight initialization methods for deep learning models. However, these results accurately reflect our hypothesis about the differences in data domains and clearly reveal the limitations of the existing international dataset in generalizing Vietnamese facial images. With strategies 1B, 2B and 3B, due to the influence of Vietnamese children's facial image data samples in the combined training dataset, the average scores have significantly improved compared to the training cases with pure international data. Specifically, the highest average accuracy value is 0.775, the highest average sensitivity value is 0.9, the highest average specificity value is 0.661111, and the highest average AUC value is 0.867531. It is worth noting that in these cases, the order of results of the strategies has changed compared to the case of evaluating on purely international test data. Although strategy 1B with deep learning models trained from scratch still gives the lowest scores, the best position among the scores is not fixedly belonging to strategy 2B or strategy 3B. Strategy 3B wins with average sensitivity of 0.9 and average AUC of 0.867531 while strategy 2B wins with average accuracy of 0.775 and average specificity of 0.661111. We also observe more specifically about the difference between strategy 2B and 3B in results in Table 3.

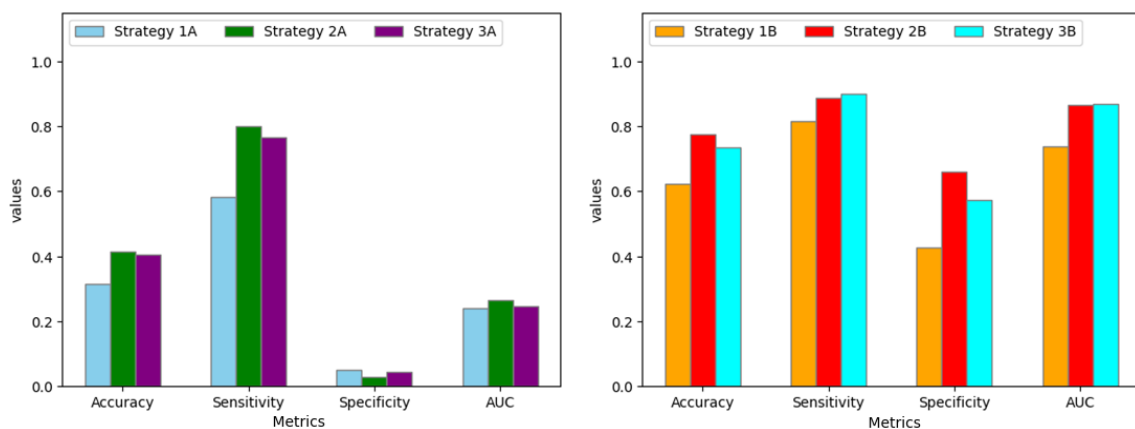


Figure 6. Average accuracy, average sensitivity, average specificity, and average AUC on Vietnamese test set of strategies 1A, 2A, 3A (left) and 1B, 2B, 3B (right)

Table 3. Compare the average scores of strategies 2B with 3B on Vietnamese test set

Case	Average accuracy	Average sensitivity	Average specificity	Average AUC
Strategy 2B	0.775	0.88889	0.661111	0.866296
Strategy 3B	0.736111	0.9	0.572222	0.867531
Compare 3B to 2B	-0.038889	+0.011111	-0.088889	+0.001235

Thus, although the results have shown the advantage of using transfer learning, the difference between using pre-trained weights on the facial expression dataset FER or pre-trained weights on the general dataset ImageNet is no longer clearly evident. In our opinion, this is due to the difference in the Vietnamese and the international dataset that were pointed out in our hypothesis. Although strategy 3B uses pre-trained weights from facial expression data, the dataset FER is not Vietnamese data while the test data considered here is purely Vietnamese data. Therefore, strategy 3B no longer has a clear advantage as in the above experiment when evaluating on the international test set.

From the above evidence, we once again see that there is value in initializing the weights of deep learning models from models pretrained on facial expression data such as FER or on general data such as ImageNet. When conducting further research, especially when working with local data sources such as Vietnamese children's facial data, both of these options should be considered. To get a more detailed view, we will look at the evaluation results on each individual deep learning model in detail.

Table 4 describes the detailed comparison of the results of deep learning models between strategy 3A and strategy 2A when evaluating on the international test set. When comparing the scores of strategies 3A with strategy 2A, we see some changes as follows. For accuracy score, the increase occurs in 3 models with the largest difference being 0.023333 of DenseNet121 model and the only degradation case is ResNet50 with the level of 0.006667. For sensitivity, the increase also occurs in 3 models with the largest

difference being 0.013334 of DenseNet121 model and the only degradation case is AlexNet with the level of 0.013333. For specificity score, the increase also occurs in 3 models with the largest difference being 0.04 of AlexNet model and the only degradation case is ResNet50 with the level of 0.02. Finally, the AUC also increases across the three models with the largest difference being 0.022222 for the DenseNet121 model and the only decline being AlexNet with 0.002044. In general, the score difference is not large with the absolute value not exceeding 0.04 and usually fluctuating around 0.01. Overall, the impact of the two strategies on each deep learning model when evaluated on two international test sets does not have a clear difference.

Table 4. Detailed score comparison of strategies 2A with 3A on international test set

Case	Architecture	Accuracy	Sensitivity	Specificity	AUC
Strategy 2A	ResNet34	0.873333	0.886667	0.86	0.944756
Strategy 3A	ResNet34	0.89	0.893333	0.886667	0.953778
Compare 3A to 2A	ResNet34	+0.016667	+0.006666	+0.026667	+0.009022
Strategy 2A	ResNet50	0.863333	0.833333	0.893333	0.923911
Strategy 3A	ResNet50	0.856667	0.84	0.873333	0.943111
Compare 3A to 2A	ResNet50	-0.006667	+0.006667	-0.02	+0.0192
Strategy 2A	AlexNet	0.84	0.84	0.84	0.926133
Strategy 3A	AlexNet	0.853333	0.826667	0.88	0.924089
Compare 3A to 2A	AlexNet	+0.013333	-0.013333	+0.04	-0.002044
Strategy 2A	DenseNet121	0.86	0.853333	0.866667	0.929511
Strategy 3A	DenseNet121	0.883333	0.866667	0.9	0.951733
Compare 3A to 2A	DenseNet121	+0.023333	+0.013334	+0.033333	+0.022222

Next, Table 5 describes the detailed comparison of the results of deep learning models between strategy 3B and strategy 2B when evaluating on the international test set. When comparing the scores of strategy 3B with strategy 2B, we see some changes as follows. For accuracy score, the increase occurs in 2 models with the largest difference being 0.053333 of AlexNet model and the only degradation case is ResNet34 with the level of 0.006666. For sensitivity, the increase also occurs in 2 models with the largest difference being 0.086666 of DenseNet121 model and the only degradation case is ResNet34 with the level of 0.073333.

Table 5. Detailed score comparison of strategies 2B with 3B on international test set

Case	Architecture	Accuracy	Sensitivity	Specificity	AUC
Strategy 2B	ResNet34	0.883333	0.933333	0.833333	0.934622
Strategy 3B	ResNet34	0.876667	0.86	0.893333	0.940044
Compare 3B to 2B	ResNet34	-0.006666	-0.073333	+0.06	+0.005422
Strategy 2B	ResNet50	0.863333	0.86	0.866667	0.943378
Strategy 3B	ResNet50	0.863333	0.86	0.866667	0.938222
Compare 3B to 2B	ResNet50	0	0	0	-0.005156
Strategy 2B	AlexNet	0.83	0.773333	0.886667	0.916311
Strategy 3B	AlexNet	0.883333	0.853333	0.913333	0.931733
Compare 3B to 2B	AlexNet	+0.053333	+0.08	+0.026667	+0.015422
Strategy 2B	DenseNet121	0.84	0.806667	0.873333	0.931022
Strategy 3B	DenseNet121	0.86	0.893333	0.826667	0.946978
Compare 3B to 2B	DenseNet121	+0.02	+0.086666	-0.046666	+0.015956

For specificity score, the increase also occurs in 2 models with the largest difference being 0.06 of ResNet34 model and the only degradation case is DenseNet121 with the level of 0.046666. Finally, the AUC also increases across the three models with the largest difference being 0.015956 for the DenseNet121 model and the only decline being ResNet50 with 0.005156. We see that the absolute values of the score deviations have increased significantly when compared to the above evaluation case of strategy 3A and strategy 2A. In which, the largest difference is 0.086666 and it is more than twice as high as the largest difference of a deep learning architecture between strategy 3A and strategy 2A. This also clearly reflects the difference in data domain between Vietnam data and international data. Here we can see the impact of Vietnamese children's facial image data when participating in the training data and it clearly has an impact on changing the distribution of data in the training set. Next, we consider the detailed comparison of the results of deep learning models between strategy 3B and strategy 2B when evaluating on the Vietnamese test set in Table 6.

Table 6. Detailed score comparison of strategies 2B with 3B on Vietnamese test set

Case	Architecture	Accuracy	Sensitivity	Specificity	AUC
Strategy 2B	ResNet34	0.744444	0.777778	0.711111	0.788148
Strategy 3B	ResNet34	0.822222	0.888889	0.755556	0.913086
Compare 3B to 2B	ResNet34	+0.077778	+0.111111	+0.044444	+0.124938
Strategy 2B	ResNet50	0.777778	0.911111	0.644444	0.901728
Strategy 3B	ResNet50	0.7	0.888889	0.511111	0.880494
Compare 3B to 2B	ResNet50	-0.077778	-0.022222	-0.133333	-0.021234
Strategy 2B	AlexNet	0.777778	0.933333	0.622222	0.888395
Strategy 3B	AlexNet	0.711111	0.911111	0.511111	0.823210
Compare 3B to 2B	AlexNet	-0.066667	-0.022222	-0.111111	-0.065185
Strategy 2B	DenseNet121	0.8	0.933333	0.666667	0.886914
Strategy 3B	DenseNet121	0.711111	0.911111	0.511111	0.853333
Compare 3B to 2B	DenseNet121	-0.088889	-0.022222	-0.155556	-0.033581

Clearly, we have seen strong differences between the resulting scores. For the accuracy score, there is only one increase in ResNet34 with 0.077778 and the remaining three decrease with the largest decrease being 0.088889 for DenseNet121. For the sensitivity score, we see the same with one increase in ResNet34 with 0.111111 and three decrease with the overall decrease being 0.022222. For the specificity score, we also see one increase in ResNet34 with 0.044444 and three decrease with the largest decrease being 0.155556 for DenseNet121. For the AUC score, there is also only one increase in ResNet34 with 0.124938 and the remaining three decrease with the largest decrease being 0.065185 for AlexNet. Clearly, the results have a stronger discrepancy with the highest difference being 0.155556 and it is significantly higher than the two estimates given above when using international test data. This has strongly demonstrated the difference between the Vietnamese child facial data domain and the international one. The results were generated with the Vietnamese test set and the training data with Vietnamese data in the minority. Therefore, the collection of local data such as Vietnamese child facial data for the autism diagnosis system from facial images in Vietnam has clearly demonstrated its role and importance through the evidence of the results.

5. CONCLUSION

Our article focuses on analyzing and discussing the role of international and Vietnamese children's face data and the influence of different pretrained weights of deep learning models in the autism classification problem. The data for the experiments in this study were collected from several kindergartens in Ho Chi Minh City, Vietnam along with an international dataset downloaded from the Kaggle platform. The proposed research focuses on designing strategies and analyzing and evaluating results to build an autism classification application with the facial data of Vietnamese children. The experimental results achieved on different metrics such as accuracy, sensitivity, specificity, and AUC. The figures pointed out the necessary of pretrained weights of deep learning models and the role of international and Vietnamese children's face data in the training phase. We also deeply discuss the exposure of data distribution differences in the proposed strategies to highlight the importance of collecting facial data of Vietnamese children for next researches.

ACKNOWLEDGEMENTS

The authors wish to thank Lac Hong University for the financial support.




REFERENCES

- [1] H. Hodge, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Translational Pediatrics*, vol. 9, pp. 55–65, 2020.
- [2] WHO, "Autism," *World Health Organization (WHO)*, 2023. <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders> (accessed Feb. 12, 2025).
- [3] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, Aug. 2018, doi: 10.1016/S0140-6736(18)31129-2.
- [4] S. B. Sulkes, "Autism spectrum disorder," *MSD*, 2025. <https://www.msdmanuals.com/professional/pediatrics/learning-and-developmental-disorders/autism-spectrum-disorders> (accessed Feb. 12, 2025).
- [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, Feb. 2015, doi: 10.1017/S003329171400172X.
- [6] 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, Apr. 2018, doi: 10.15585/mmwr.ss6706a1.
- [7] R. J. Landa, "Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders," *International Review of Psychiatry*, vol. 30, no. 1, pp. 25–39, Jan. 2018, doi: 10.1080/09540261.2018.1432574.




- [8] K. E. Zuckerman, S. Broder-Fingert, and R. C. Sheldrick, "To reduce the average age of autism diagnosis, screen preschoolers in primary care," *Autism*, vol. 25, no. 2, pp. 593–596, 2021, doi: 10.1177/1362361320968974.
- [9] 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.
- [10] I. A. Ahmed *et al.*, "Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques," *Electronics (Switzerland)*, vol. 11, no. 4, 2022, doi: 10.3390/electronics11040530.
- [11] 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, p. 9605, Jun. 2023, doi: 10.1038/s41598-023-35910-1.
- [12] Ministry of Labour, "Respond to the world autism awareness day: stronger engagement of the community in supporting children with autism," *Ministry of Labour*, 2022. <https://english.molisa.gov.vn/topic/231259> (accessed Feb. 12, 2025).
- [13] V. S. Karri, S. Remya, A. R. Vybhav, G. S. Ganesh, and J. Eswar, "Detecting autism spectrum disorder using DenseNet," in *ICT Infrastructure and Computing*, 2023, pp. 461–467.
- [14] M. F. Rabbi *et al.*, "Autism spectrum disorder detection using transfer learning with VGG 19, inception V3 and DenseNet 201," in *Communications in Computer and Information Science*, 2023, pp. 190–204.
- [15] 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.
- [16] T. M. Ghazal, S. Munir, S. Abbas, A. Athar, H. Alrababah, and M. Adnan 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.
- [17] P. Reddy and A. J., "Diagnosis of autism in children using deep learning techniques by analyzing facial features," in *RAiSE-2023*, Jan. 2024, p. 198, doi: 10.3390/engproc2023059198.
- [18] 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, Aug. 2024, doi: 10.1049/htl2.12073.
- [19] T. T. Le, L. G. Farkas, R. C. K. Ngim, L. S. Levin, and C. R. Forrest, "Proportionality in Asian and North American Caucasian faces using neoclassical facial canons as criteria," *Aesthetic Plastic Surgery*, vol. 26, no. 1, pp. 64–69, Jan. 2002, doi: 10.1007/s00266-001-0033-7.
- [20] H. Li, S. Yang, H. Chen, L. Liu, Y. Zhang, and C. Dai, "Morphometric variations and growth of the profile of the face in Chinese boys aged 4–15 years," *HOMO*, vol. 71, no. 2, pp. 83–90, Apr. 2020, doi: 10.1127/homo/2020/1196.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [22] Q. Mohi ud Din and A. K. Jayanthi, "Automated classification of autism spectrum disorder using EEG signals and convolutional neural networks," *Biomedical Engineering: Applications, Basis and Communications*, vol. 34, no. 02, Apr. 2022, doi: 10.4015/S101623722250020X.
- [23] 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)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
- [25] Kaggle, "Kaggle platform," *Kaggle*, <https://www.kaggle.com> (Feb. 12, 2025).
- [26] PyTorch, "PyTorch library," *PyTorch*, <https://pytorch.org>, accessed (Feb. 12, 2025).
- [27] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science*, vol. 8228. Springer, Berlin, Heidelberg. 2013, pp. 117–124, doi: 10.1007/978-3-642-42051-1_16.

BIOGRAPHIES OF AUTHORS






Tran Van Thanh    received his master's degree in control and automation engineering from the University of Transport, Ho Chi Minh City, in 2012. He is currently a lecturer in the Faculty of Electrical and Electronics Engineering at Lac Hong University. His research interests include computer vision, image processing, machine learning, industrial control, and communication systems. He can be reached at email: thanhtran@lhu.edu.vn.






Lam Thanh Hien    joined a M.Sc. in applied informatics at the INNOTECH Institute, France, and received a degree in 2004. In 2017, he earned a Ph.D. degree at the Vietnam Academy of Science and Technology. Now, he works at Lac Hong University in the headmaster's role. His studies interests relate to machine learning, computer vision, and deep learning. He can be reached at email: lthien@lhu.edu.vn.






Do Nang Khoa    is a final-year student of the Faculty of Information Technology, Hanoi Architectural University. The graduation project, he is working on is about image processing and deep learning. He can be contacted at email: dnangkhoa@gmail.com.






Le Anh Tu    graduated with a Master's degree in 2007 from Thai Nguyen University, and a Ph.D. in 2017 from the Institute of Information Technology, Vietnam Academy of Science and Technology. He is currently a lecturer at Ha Long University. His main research areas are data mining, machine learning, and artificial neural networks. He can be reached at email: leanhtu@daihochalong.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 reached 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 reached at email: dntoan@ioit.ac.vn.