Comparison of CNN-Based Models in Facial Micro-Expression Classification

Ruoxuan Liu *

School of Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangzhou, China

* Corresponding Author Email: 20223801043@m.scnu.edu.cn

Abstract. With the rapid development of deep learning, especially the application of Convolutional Neural Networks (CNNs), significant progress has been made in the identification and recognition of facial micro-expressions. This paper explores and compares three widely used models in this field: VGG, DenseNet, and ResNet, across seven expression labels—Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. The comparison is based on several key evaluation metrics commonly used in deep learning classification tasks. To ensure consistency, critical training parameters such as epochs, learning rate, and data preprocessing steps are kept the same across all models. Additionally, a dropout layer is incorporated to address issues such as overfitting and improve generalization in each model. These works indicate that VGG has the highest performance in RAFDB dataset because it as an F1-score of 0.76, an AUC value of 0.95, and a 75% accuracy rate. Additionally, VGG uses only 1243.31MB of memory, making it the most efficient model compared to the other two models. This shows that VGG excels not only in performance but also in memory efficiency.

Keywords: Facial Expression; Deep Learning; CNNs; VGG.

1. Introduction

Facial micro-expression is involuntary, fleeting, and subtle [1]. It is said that 55% of daily communication between people is through facial expression [2], which means that micro-expression is an important part of daily life. In order to find the true feelings people, attempt to control or suppress, facial micro-expression is important of the real life. Last decade, facial micro-expression analysis has been used in various fields for its development of automatic analysis in computer vision. Facial micro-expression classification and recognition demonstrate widespread applicability and significance for their transience, faintness, and complexity. Traditional micro-expression recognition methods usually rely on manual feature extraction and rule-based algorithms such as those based on Facial Action Coding Systems (FACS). These methods analyze the micro-expression changes by coding facial muscle movements. Although FACS has been extensively studied and considered to be a reliable method for the effective detection of facial expressions, manual FACS coding process is labor-intensive and impractical for real-world applications [3].

Deep learning and machine learning are key parts within the field of artificial intelligence, both playing a critical role in the advancement of technology. As it could handle large data sets, predict data, and provide insights that were previously impossible to obtain, these two technologies are used in various fields and more and more people choose them for training [4]. Nowadays, with the increasing trend of data sets, AI methods are widely used. Machine learning has the advantages of structured data, while deep learning does well in unstructured data, like images, speech, and text [4]. So, deep learning is more suitable for complex tasks in handling data sets, like recognizing and classifying facial micro-expressions than machine learning. Deep learning is capable of autonomously identifying intricate features within micro-expressions, and enhances the precision and resilience of recognition through various models [5].

Convolutional Neural Networks (CNNs), a powerful architecture in the field of computer vision, is one of the effective models for recognizing and classifying facial micro-expressions in deep learning [6]. Compared to other models of image recognition and classification in deep learning, CNN

involves less pre-processing and can extract key features automatically from the initial data sets in the process of training [7]. Besides, CNN reduces the complexity of the model because it uses the same convolutional kernel parameters in the process of training, which could reduce computational cost and storage requirements [8]. At the same time, the multilayer structure of CNN could gradually extract features from low to high levels, which is suitable for processing complex data, especially images [9]. With the popularity and development of CNN, there are more and more different versions and developments of it. These changes are based on three basic network layers, the convolution layer, the pooling layer, and the complete connection layer. This paper aims at providing more effective assistance for subsequent facial micro-expression recognition and classification studies by comparing different deep learning models with the base of CNN structure.

2. Methodology

The paper will detail three classical convolutional neural networks—DenseNet121, ResNet34 and VGG16. These models have different structures and features, which are suitable for classifying and recognizing the facial micro-expression tasks. The following sections will explore their architectures and optimization strategies, to provide a theoretical basis for subsequent experiments.

2.1. VGG-16

VGG represents a stable performance in multiple visual recognition tasks, especially in areas such as facial recognition. In this way, The VGG-16 architecture is composed of five convolutional blocks, three dense layers, and a softmax activation layer for output [10]. The above structure is suitable to extract more intricate facial characteristics. It mainly improves the accuracy of image recognition and classification by using a depth-widened convolutional neural network, which employs numerous small 3x3 convolutional kernels in place of a single larger one, thereby enhancing the network's depth to capture more nuanced features, and avoiding explosive growth of model parameters.

First of all, a 48*48-pixel color image is inputted in the model with three RGB color channels. Then, two continuous convolution layers with the activation function ReLU extract features and a normalized layer Accelerate the learning phase and mitigate the issue of vanishing gradients. At the same time, the maximum pooling layer employs a 2x2 window, advancing with a step size of 2, to reduce the dimensionality of the feature map. The convolution block is used four times, with different numbers of convolution kernels in 64, 128, 256, and 512. Furthermore, in the final block, there is a flattened layer, which flat the output of the convolution layer to form a one-dimensional vector for input to the fully connected layer. This is followed by two consecutive fully connected layers with ReLU function, two dropout layers to discard 10 percent of neuron output randomly for preventing overfitting and an output layer. Finally, there is an output layer with 7 neurons, correspondings to the 7 classes of anger, disgust, fear happiness, neutral, sadness, and surprise, which are classified using the softmax activation function.

2.2. DenseNet-121

DenseNet is designed by reduced links between layers near the input and layers near the output to prevent the vanishing gradient, improve the ability of facilitate feature dissemination, promote the recycling of features, and reduce the number of parameters [11].

These features indicate that it has a high computational efficiency and generalization, which could deal with classifying and recognizing pictures efficiently, especially in a huge data set.

In this context, DenseNet-121 will be shown as an example. DenseNet is divided into three main architectures: Dense Layer, Dense Block, and Transition Layer. Dense Layer includes a 1*1 Convolutional Layer and a 3*3 Convolutional Layer. It identifies fundamental characteristics like edges, corners, and textures, which are crucial for identifying the subtle changes in facial muscles in the recognition of micro-expression. Dense Block of DenseNet-121 includes a 1*1 Convolutional Layer and a 3*3 Convolutional Layer, which helps pick up the nuances of micro-expressions and

translate them. Transition Layer is composed of a 1*1 Convolutional Layer, a 2*2 Average Pooling Layer, and a 1*1 Convolutional Layer, aiming at consisting of width of the feature layers to make the network focus on the key features. The DenseNet-121 has a Dense Layer, four Dense Blocks, each of them cycling 6, 12, 24, 16, and the first three of them followed by a Transition Layer and a Classification Layer with a 7*7 Global Average Pooling Layer and a fully connected Layer.

2.3. ResNet-34

ResNet aims at remitting gradient vanishing in training of deep network by using residual learning so that the network can learn the deep features in different types of facial micro-expressions effectively.

Initially, ResNet includes a 7*7 Convolutional Layer and a 3*3 Max Pooling Layer to extract basic image features [12]. The core component is four Residual Blocks, aiming at learning the residual mapping between input and output to gradient mitigation disappearance. These blocks further extract and combine micro-expression features. Each residual block consists of a convolutional block and an Identify Block. As the number of data sets is around 20,000, a model with few layers is chosen, including 34 convolutional and fully connected layers. In ResNet34, the Convolutional Block has different numbers of Convolutional Layers with the width of 64, 128, 256, 512, and the number of layers is 6, 8, 12, 6. In different network stages, downsampling is performed by convolutional layers with a stride of 2, which results in an increased number of channels in the feature map, preserving the depth of facial micro-expression features. Finally, an average pooling reduces the space dimension of the feature map, followed by a Flattened Layer to output the results.

2.4. Dropout Layer

To enhance the performance of the model and prevent overfitting, a Dropout Layer was added to the fully connected layers in each model. This layer makes the output of neurons randomly and temporarily be 0 when the neural network is training [13], which means that for each neuron, the dropout layer randomly decides with a certain probability p whether this neuron is activated in this iteration or not. Consequently, each training example uses a different network structure, which is the same as consisting of different networks, to avoid extracting special but not usual facial micro-expression features when training. To reduce the model's dependence on any training samples and improving the generalization of these models, this layer improves the robustness of the network learning

By testing different values of the dropout rate p, it was set to 0.2, which may affect the model's fitting capacity and its ability to generalize. This adjustment allows the models to more accurately and efficiently capture the features of facial micro-expression.

3. Results

3.1. Data Set

This research uses a high-quality dataset called RAFDB, which contains 10,000 face images representing multiple ethnicities, including Asian, African, Caucasian, Latin American, and so on, providing powerful support for training the model on a larger scale. The dataset comprises 48x48 pixel RGB images representing seven distinct emotional expressions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Additionally, the dataset is already classified, allowing for easier and faster model training. There are some examples of each classification in Fig. 1.

(a)Anger (b)Disgust (c)Fear (d)Happiness (e)Neutral (f)Sadness (g)Surprise

Fig. 1 Examples of Data sets

3.2. Experimental Environment

In the process of training, overfitting is a significant challenge. One common solution is adjusting the learning rate throughout the process. There is an Adam optimizer, which has been set a learning rate of 0.0001. The Adam optimizer will be responsible for adjusting the weights of the model according to the gradient, so that the model will be gradually optimized and the loss function will be minimized. The experiment configuration and parameter setting are shown in Table 1.

Table 1. Experiment Configuration and Parameter Setting

Experimental Configuration	Parameters
Operating System	Windows 11
CPU Version	11th Gen Intel(R) Core(TM) i5-1135G7@2.40GHz
RAM (Memory)	16.0 GB
Software	Pycharm
Learning Rate	0.0001(with Adam)
Epoch	75
Batch size	512

3.3. Experimental Result

In this paper, precision, recall, F1 score, accuracy, AUC value, and Memory are used as performance evaluation metrics for models. These metrics are key indicators for evaluating the model in deep learning, especially in the case of unbalanced distribution of datasets. As a reason for that, this paper uses them to evaluate the performance of the models in classifying and recognizing facial micro-expression.

3.3.1 Evaluation Indicators

Before these three evaluation indicators are introduced, there are some parameters that need to be known.

True Positive (TP): Instances where positive samples were accurately identified. These are instances where both the actual and predicted classifications are positive. True Negative (TN): Correctly identified negative samples. These occur when both the actual and predicted classifications are negative. False Positive (FP): Incorrectly identified positive samples. This occurs when the actual classification is negative, yet the prediction labels it as positive. False Negative (FN): Incorrectly identified negative samples. This happens when the actual classification is positive, but the prediction incorrectly labels it as negative. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

(1) Precision

Precision indicates the proportion of samples with positive predictions that are positive. For example, in classifying facial micro-expressions, a false positive case on behalf of a non-Anger expression is errorly predicted to be Anger in the classification Anger.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

(2) Recall

Recall reflects the ratio of the correctly predicted positive instances to the total number of actual positive instances within the dataset. For example, in classifying facial micro-expressions, a false negative case represents that an Anger expression is predicted to be a non-Anger expression.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

(3) F1-score

Precision and Recall, are generally used to evaluate the analytical effectiveness of binary classification models, F1-score is used to synthesize these two values with different weights. F1-score is a Weighted Average of Precision and Recall Rates. Here, Precision and Recall have the same weights, so the formulation is:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

(4) Accuracy

Accuracy represents the ratio of correctly classified instances to the overall sample count. The correctly categorized samples have two components, instances where the forecast is positive coinciding with the actual being positive, denoted as TP, as well as instances where the forecast is negative aligning with the actual being negative, denoted as TN.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

(5) Weighted Average

As there are seven classes in the model and there are different numbers of images in different classes, the Weighted Average is used to calculate the overall value of these four parameters mentioned before. P' is the weighted average parameter, P1, P2… is set to be the parameters in each class, and N1, N2… is set to be the images of numbers in each class.

$$P' = \frac{P1*N1+P2*N2+P3*N3+P4*N4+P5*N5+P6*N6+P7*N7}{N1+N2+N3+N4+N5+N6+N7} \tag{5}$$

(6) AUC

AUC value is an important indicator in evaluating the classification models. It evaluates the likelihood that a randomly chosen positive example is assigned a higher rank than a randomly chosen negative example according to their predicted probabilities.

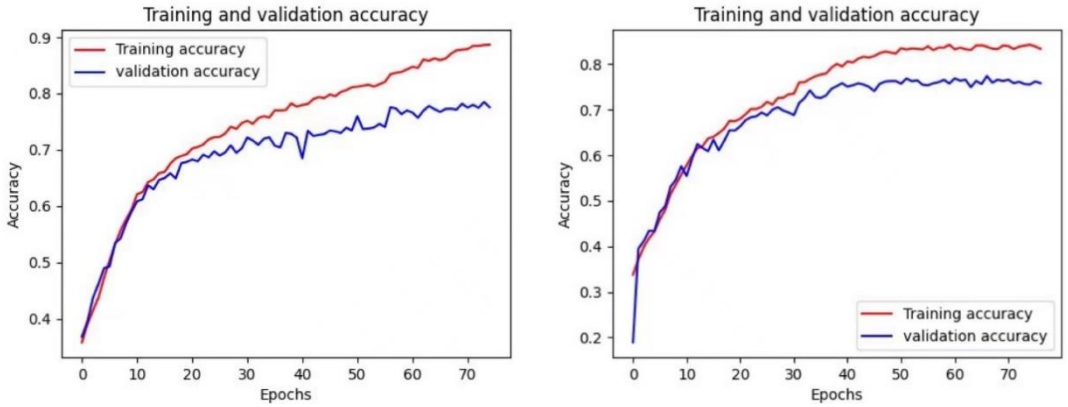
Consider M positive samples, N negative samples, and a total of n samples. Calculate the rank value of each sample in the prediction result, and its position after the ascending order ranking, the sample with the largest PROBABILITY rank is n.

When a positive sample is ranked kth in the ascending order of the positive class prediction result, it is shown that it forms a correctly ranked pair with the negative samples that follow it, and it is the sum of all correctly ranked pairs of samples.

$$AUC = \frac{CorrectPair}{M*N} \tag{6}$$

3.3.2 Model Performance Comparison

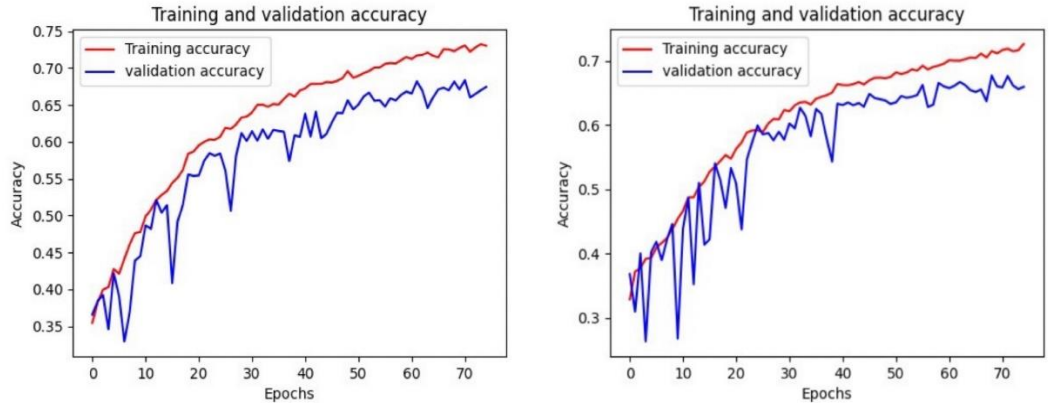
The paper uses the evaluation metrics of 3.3.1 to validate the experiment in a harmonized way. First of all, it compares VGG16, DenseNet121, and ResNet34 with the initial model and with a dropout layer. Then, each model is compared with the final version of it with the dropout layer separately.



(a)without dropout layer (b)with dropout layer

Fig. 2 Training and Validation Accuracy in VGG-16

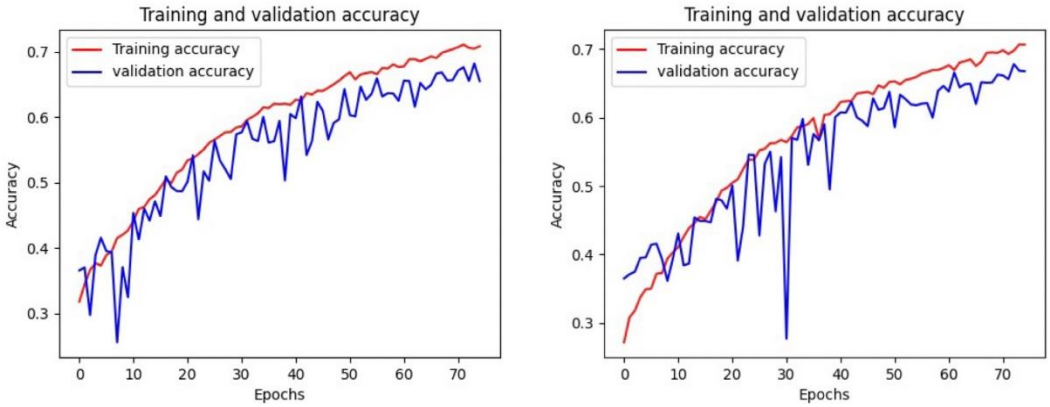
The curve in Fig. 2(a) illustrates the accuracy trends for training and validation during learning process of the VGG-16 model without a dropout layer. The curve in Fig. 2(b) illustrates the accuracy trends for training and validation during learning process of the VGG-16 model with a dropout layer. Observing Fig. 2, it is evident that the accuracy curve for both the training and validation sets with a dropout layer converge gradually when training while the accuracy curve without the dropout layer still has an increasing trend. Without the dropout layer, the difference between these two curves builds up gradually, which shows that the figure without the dropout layer has an overfitting problem, which indicates that the model is overfitting to the training dataset and lacks sufficient generalization to unseen data. At the same time, the accuracy curve of training set and validation set are very close to each other with the dropout layer, indicating that the model fits better, has a strong generalization ability, and has less risk of overfitting.



(a)without dropout layer (b)with dropout layer

Fig. 3 Training and Validation Accuracy in DenseNet-121

The curve in Fig. 3(a) illustrates the accuracy trends for training and validation during learning process of the DenseNet-121 model without a dropout layer while the curve in Fig. 3(b) illustrates the accuracy trends for training and validation during learning process in the same model with a dropout layer. In Fig. 3, the discrepancy between the validation accuracy trend and the training accuracy trend with dropout layer is smaller than that without dropout layer, which suggests that the knowledge acquired by the model from the training set is effectively applied to the validation set and it performs well on a dataset that has never been seen before. At the same time, as the model is trained, the validation accuracy with the dropout layer becomes smoother and smoother, while the validation accuracy without the dropout layer oscillates only slightly less, which indicates that the performance of DenseNet-121 with the dropout layer becomes stable and well.



(a)without dropout layer (b)with dropout layer

Fig. 4 Training and Validation Accuracy in ResNet-34

The curve in Fig. 4(a) illustrates the accuracy trends for training and validation during learning process of the ResNet-34 model without a dropout layer while the curve in Fig. 4(b) illustrates the accuracy trends for training and validation during learning process in the same model with a dropout layer. In Fig. 4, the validation set accuracy curve without a dropout layer, fluctuates more, indicating that the model's performance in different rounds of validation is not stable enough. In particular, there are more decreases in the middle and late stages, which means that the model may be overfitting and generalizing poorly. The validation accuracy curve with the dropout layer, on the other hand, is relatively smooth with small oscillations in most cases and it is gradually close to the training accuracy, which signifies that the model's efficacy on the validation dataset remains consistent across various training epochs, demonstrating improved generalization capabilities.

Table 2 shows the comparison of each model's performance with the RAFDB data set. The first five parameters are mentioned in the 3.3.1 section, and the memory is the overall memory used in this model run and prediction process. In order to demonstrate the model's forecasting ability, the test outcomes are employed to highlight its efficacy.

Table 2. Model Performance Evaluation Comparison

Model	Precision	Recall	F1-score	AUC	Accuracy	Memory
ResNet-34	0.70	0.68	0.69	0.89	0.68	1443.31MB
DenseNet-121	0.68	0.67	0.67	0.90	0.67	1322.56MB
VGG-16	0.77	0.75	0.76	0.95	0.75	1243.31MB

It shows that VGG-16 has the highest accuracy when the dropout layer is added, with an F1-score of 0.76, an AUC value of 0.95, an accuracy of 0.75, and less memory of 1243.31MB. This is because VGG-16 has a consistent structure and the feature map of the network gradually shrinks with increasing depth. This consistency makes VGG more regular, with the convolution operations on almost all layers following the same structure. While the convolutional operations and connectivity of DenseNet and ResNet are more complex and variable, the comprehensibility of these models is reduced. During the training phase, VGG networks could quickly modify the number of layers or structures, while DenseNet and ResNet require more consideration when modifying and adapting due to the complexity of their connection. So, the simple structure of VGG here becomes an advantage compared to the other two models. As the simple structure of VGG, it doesn't need to adjust some parameters such as convolutional kernel size, number of network layers, and step sizes, which are appropriate for limited data sets like RAFDB. Compared to the VGG, ResNet and DenseNet have a complex connectivity, which leads to more memory access in computing the model.

4. Discussion

The experiments in Section 3 show that VGG-16 performs well in the facial micro-expression recognition task, although it has a simple structure. It is mainly because of the simple structure of it, which reduces the complexity of adjusting the parameters and the ubiquity of it, especially in the small data sets. The DenseNet fails to adequately capture changes in facial micro-expressions. It mainly because the intensive connections in DenseNet require more complex tuning of how the model is connected, the number and the depth of the layers in the model. Therefore, the design of it requires more data and computing resources to show its benefits.

Furthermore, the paper adds a dropout layer following the pooling layer within the fully connected layer architecture, which could remit the overfitting and enhance the efficacy of all these models. The optimal hyperparameter setting was found by adjusting the dropout rate through several experiments.

This article aims at providing a further reference toward advancing the recognition of facial micro-expressions. To further this paper, it could be considered to be fine-tuned using transfer learning, which could accelerate training and improve model performance. Besides, the Ensemble is a common approach to improving performance. By training multiple VGG, DenseNet, or ResNet models and averaging their predictions during inference, the prediction bias and variance of individual models can be effectively reduced, thus improving overall performance [14], [15]. There are some methods including bagging or bosting, which could improve the performance of the models though it increases the computing costs. In recent times, the Attention mechanism has yielded outstanding outcomes within the domain of facial micro-expression classification and recognition. [16]. It significantly enhances the network's functionality through adaptively adjusting the network's attention to make the models concentrate on the essential characteristics to enhance the precision of classifying and recognizing.

5. Conclusion

This paper presents an experimental comparison of several commonly used models based on CNN shows the advantages and disadvantages of each model. Specifically, VGG has the most excellent performance in facial micro-expression feature extraction with a deep network structure. ResNet has a better performance than DenseNet in F1-score and Accuracy, while DenseNet uses a lower memory and gets a better AUC compared to ResNet in this data set. More importantly, adding a Dropout Layer in these three models could not only improve the training effects but also help models converge better and faster.

This paper aims to show the advantages and disadvantages in facial micro-expression in deep learning based on CNN, helping other researchers to find a certain choice when training these models with different types of data sets. In order to expand the application scope of facial micro-expression recognition to include a broader array of fields, such as medical care, fatigue driving detection, crime detection, and so on, it could be developed by multimodal hybrid recognition models or architectures. Besides, the Ensemble and Attention mechanism could be utilized to enhance the precision of these models.

References

[1] Zhao G, Li X, Li Y, et al. Facial Micro-expressions: an overview[J]. Proceedings of the IEEE, 2023, 111(10): 1215-1235.

[2] Mellouk, W. and Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. Procedia Computer Science, 175, 689-694

[3] Chen Z, Ansari R, Wilkie D. Automated pain detection from facial expressions using facs: A review[J]. arXiv preprint arXiv:1811.07988, 2018.

[4] Sharifani K, Amini M. Machine learning and deep learning: A review of methods and applications[J]. World Information Technology and Engineering Journal, 2023, 10(07): 3897-3904.

[5] Zeng X, Zhao X, Zhong X, et al. A survey of micro-expression recognition methods based on lbp, optical flow and deep learning[J]. Neural Processing Letters, 2023, 55(5): 5995-6026.

[6] Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F. and Wang, H. (2020). Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Systems with Applications, 149, 113305

[7] Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., Li, X., et al. (2020). Deep convolution network-based emotion analysis towards mental health care. Neurocomputing, 388, 212-227, 2020

[8] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014: 818-833.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. [J]. CoRR,2014, abs/1409.1556

[11] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[14] Das A, Jalal M S, Bari A S M S, et al. Facial Emotion Recognition by Ensemble-DenseNet Networks[C]//2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2023: 608-613.

[15] Tang J, Su Q, Su B, et al. Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition[J]. Computer Methods and Programs in Biomedicine, 2020, 197: 105622.

[16] Liu H, Cai H, Lin Q, et al. Adaptive multilayer perceptual attention network for facial expression recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6253-6266.