Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/22/20231162

A critical examination of deficiencies in the convolutional neural network model for facial emotion recognition

Yuhui Tao

Department of Computer Science, East China University of Science and Technology, Shanghai, 200237, China

20002341@mail.ecust.edu.cn

Abstract. The challenge of addressing the issue of low accuracy in specific scenarios encountered during the implementation of facial emotion recognition systems arises due to the wide array of environments and varying conditions. In this study, the Facial Expression Recognition-2013 (FER-2013) dataset sourced from the Kaggle serves as the basis for training the models, with subsequent analysis conducted on the experimental outcomes. The dataset comprises a training set and a testing set, each annotated with labels representing seven distinct emotions, ranging from "angry" to "surprise". The models developed for facial emotion classification, tasked with automatically recognizing emotions based on provided images, consist of a MobileNet-based model and a self-built model employing convolutional neural networks. Both models exhibit an accuracy of approximately 60%, yet demonstrate deficiencies in predicting the "neutral" label. Additionally, the utilization of techniques such as confusion matrix and saliency map enable the comparative evaluation of model performance across different emotion labels and facilitates an analysis of their corresponding dominant facial regions. Based on a comparison of results obtained from representative cases, two potential factors contributing to these limitations are identified: a paucity of training data and the presence of ambiguous features. The findings of this study are expected to inform future directions for improvement and modification of facial emotion recognition models in order to enhance their applicability in diverse scenarios.

Keywords: facial emotion recognition, convolutional neural network, confusion matrix, saliency map

1. Introduction

Facial emotion recognition, an automated system capable of distinguishing the emotional states of individuals by analyzing facial images, such as happy, sad or angry, by analyzing the given facial images. According to a psychological survey in 1970s, human facial emotions can be classified into basic categories for measurement, namely happiness, anger, surprise, fear, disgust and sadness [1]. It is now widely applied, together with an additional neutral emotion, as the facial emotion classification criterion.

Facial Expression Recognition-2013 (FER-2013) is a kind of commonly employed facial emotion dataset, which divides all kinds of emotions into seven different categories according to experiment statistics – “angry”, “disgust”, “fear”, “happy”, “neutral”, “sad” and “surprise”. It contains over 32,298 labeled grayscale images, including train, validation and test dataset. The utilization of FER-2013

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

presents an intriguing challenge within the domain of deep learning, particularly in the context of conducting facial emotion recognition across diverse environments and under varying conditions. It is also an important project worth investigating in many facial emotion related fields – judging whether a man is fatigue driving through surveillance cameras in order to avoid potential traffic accidents, or helping analyze psychological health problems effectively and efficiently for a patient with mental diseases [2], etc.

In recent years, significant advancements have been made in the domain of facial emotion recognition by leveraging the combined capabilities of neural networks and algorithms. In 2015, extreme sparse learning is introduced to solve the uncontrolled environment problem. With a specific discriminative criterion added to the traditional dictionary training, the classification ability of the model is directly improved [3]. And, just two years before, in 2021, the model networks and dataset are promoted to three dimensions for a better prediction result. After applying 3D facial data with more continuous information to the training process, the final convolutional neural networks have reached a more satisfying accuracy of 69% [4].

In this special period when the corona virus is prevalent worldwide, a targeted modified dataset, masked FER-2013, is founded. This specialized dataset automatically synthesize an actual mask on the initial FER-2013 images as if the man is wearing a mask [5]. With this dataset progress, the facial emotion recognition can be extended to fit the current society and become more suitable for the real application scenarios. However, most research focus on optimizing the model structure, network layers or dataset to increase the accuracy and breadth of facial emotion recognition rather than analyzing the results themselves. Such lack of result analysis may lead to a problem that the overall accuracy is high while some particular cases have a rather low correct prediction rate, making the model only works well in certain conditions.

To solve the limitation mentioned above, this paper aims to use the basic convolutional neural networks and lays emphasis on exploiting the test results and directions for improvement. Two convolutional neural networks are introduced, namely MobileNet based model and a self-build model, which both have typical defects in certain facial emotions to train the FER-2013 dataset and collect their outputs. More importantly, this paper employs confusion matrix and saliency map. Taking advantage of these two methods, it can be visually shown that, among all facial emotions listed in FER-2013, which one works best, and which one works worst. The possible reasons, additionally, are also proposed and concluded combined with the saliency map of each emotion label.

The rest part of this essay is organized into three sections. The second section will provide the methods to carry out this project, including dataset description and preprocessing, overview of the two Convolutional Neural Network (CNN) based model and implementation details. The third part will show the experimental results, from epochs-accuracy curve to confusion matrix and saliency map, and then discuss the defects of the models, causes of their formation and possibly feasible improvements. The last part will cover the conclusion and prospects of the whole project.

2. Methods

2.1. Dataset description and preprocessing

The dataset used to train the two models in this paper is FER-2013 collected on the public website named kaggle [6]. The dataset consists of two main parts, including train data and test data. The training dataset contains 28,709 image examples while the public test dataset consists of 3,589 image examples. Each example represents a grayscale image of a face, measuring 48×48 pixels. Prior to inclusion in the dataset, the facial images were automatically aligned, ensuring that the face is approximately at the centre of the image and occupies a consistent amount of space. The train and test folder both encompass seven categories indicating the true facial emotions of the images.

The data preprocessing employed includes normalization and augmentation in this study This procedure contributes to the improvement of the network's generalization capacity by mitigating the impact of variations in feature scales. As for the FER-2013 dataset in this paper, it is rescaled by 1./255 because each image pixel is an integer between 0 and 255 and it will be transferred into 0 to 1 through rescaling in order to avoid the problem of significant numerical fluctuations. Data augmentation is a preprocessing technique, which applies various transformations to the existing data, like rotation, shifting, flipping and adding noise to the input images, so that the number of training data is increased artificially. It is especially useful when the available training data is limited. As for the FER-2013, the width shift range and the height shift range are 0.2, the rotation range is 5, the zoom range is 0.2 and the horizontal flip and vertical flip are also both applied.

In order to facilitate the classification of facial expressions based on the underlying emotions depicted, it is necessary to convert the specific emotion labels into abstract numerical representations that can be readily interpreted by the model. Therefore, the seven emotion categories in FER-2013 are redefined by the rule – 0 for Angry, 1 for Disgust, 2 for Fear, 3 for Happy, 4 for Sad, 5 for Surprise, and 6 for Neutral. The sample images of the dataset are shown in Figure 1.

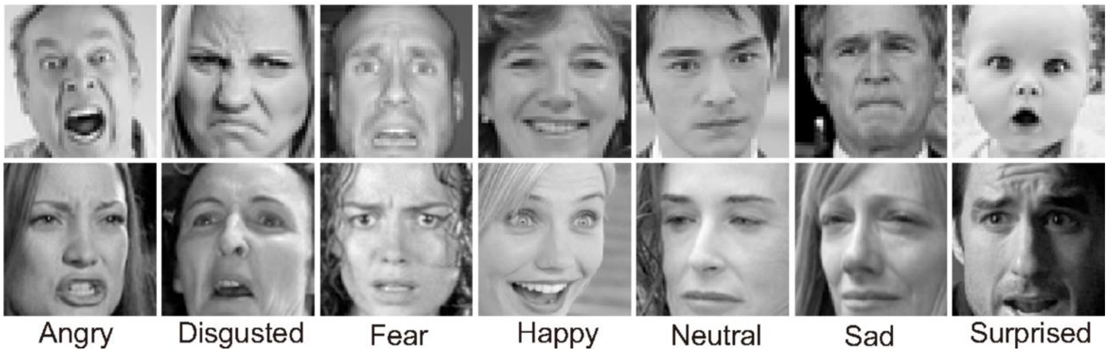

Figure 1. Sample images of FER-2013 dataset.

2.2. CNN-based model

Convolutional neural networks, a machine learning model which can learn from the input images with true labels automatically and predict the most possible labels of new images by the features it learned before, works really well in image classification [7-10]. It is perfectly suitable for the task of facial emotion recognition. MobileNet is a deep learning architecture designed for efficient and lightweight models, which is especially suitable for mobile and embedded devices. It strikes a balance between model size and computational efficiency without compromising accuracy too much. Its ability to achieve good performance on resource-constrained devices helps raise the efficiency of the image classification task.

The convolutional neural network built in this study is composed of convolutional layer, dropout layer, flatten layer, dense layer, activation layer and max-pooling layer. The convolutional layer computes the dot products between the weights and local receptive field of the input image data by the given kernel. It helps capture local patterns and spatial dependencies in the data, making the network extract features from the input images automatically. The dropout layer is a regularization technique used to prevent overfitting in neural networks. The flatten layer is used to transform the output of a convolutional layer into a one-dimensional vector, which can be fed into a dense layer. The dense layer, namely fully connected layer, is a classic neural network layer receiving input from all the neurons in the previous layer and producing output by performing a weighted sum of those inputs followed by an activation function. The activation layer introduces non-linearity into the neural network by applying a non-linear activation function. The max-pooling layer is used for feature extraction and dimensionality reduction by exporting the maximum of each region according to the pooling size so that the computational efficiency can be greatly improved.

The MobileNet based model firstly adds a dropout layer and a flatten layer after the MobileNet model, and then it is followed by three groups of dense layers, activation layers and dropout layers before the final output layer using “softmax” as its activation function. The self-build model uses three groups of convolution layers with kernel size of 3×3, activation function of "relu", padding type of "same", and max-pooling layers with the pooling size of 2×2 to replace the MobileNet part. The detailed structure of the two models introduced in this paper are shown in Figure 2 (the MobileNet based model) and Figure 3 (the self-build model).

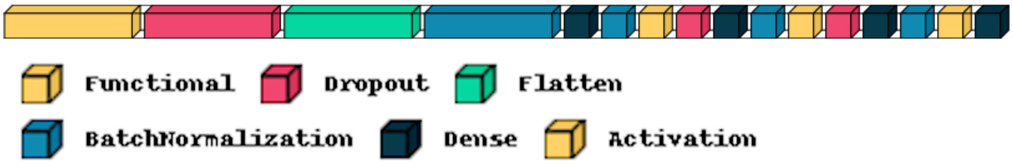

Figure 2. The structure of the MobileNet based model.
Photo/Picture credit: Original

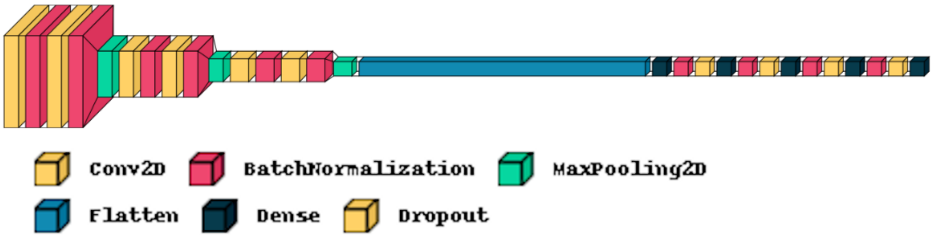

Figure 3. The structure of the self-build model.
Photo/Picture credit: Original

2.3. Implementation details

TensorFlow has gained widespread adoption in the field of the machine learning frameworks due to its versatility, scalability, and comprehensive community support. As for the model compiling and fitting process, the loss function is categorical crossentropy and the optimizer is Adam, particularly suitable in classification tasks where the goal is to assign input data to multiple mutually exclusive classes. Accuracy is chosen as the evaluation metric so that the performance of the model can be assessed precisely. Furthermore, a learning rate of 0.01 is utilized, with a batch size of 64 and an initial epoch setting of 100. Additionally, the early stopping technique was also applied. The patience is 10, which means if the validation accuracy does not improve in 10 epochs, the model fitting will stop.

In terms of data results analysis, confusion matrix and saliency map are employed. Confusion matrix is a standard format to depict the accuracy of image classification by using an n×n matrix – each row represents a true label, and each column represents a predicted label. The diagonal of this matrix, comparatively, indicates the correct prediction rate of each label. Saliency map is a similarly visual attention system integrating multi-scaled feature map. Marking different regions of the image with different colors, according with their weights, how the model works is clearly displayed – pointing out what part of the image plays the vital role in judging the emotion.

3. Results and discussion

3.1. Experimental results

After implementing the methods of MobileNet and convolutional neural network, setting the parameters as 64 batch size, 100 epochs and 10 early stopping patience, the validation accuracy of the MobileNet based model and the self-build model reaches 64.327% in 74 epochs and 59.752% in 68 epochs, respectively. In terms of the testing accuracy, the MobileNet based one has 63.263%, while the self-build one has 59.348% comparatively. The final performance of each model is shown in Table 1.

Table 1. The performance of the MobileNet based model and the self-build model based on the FER-2013 dataset.

	Performance				
Model	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Testing Accuracy
MobileNet based model	0.9334	0.6751	1.0420	0.6433	0.6326
Self-build model	1.0343	0.6146	1.0779	0.5975	0.5935

The progression of loss and accuracy metrics with increasing epochs, in which the MobileNet-based model is applied, are visually represented in Figure 4 and Figure 5, in which the blue curve indicates the performance of training set and the orange curve indicates the performance of testing set. Additionally, the two types of curves using the self-build model are also shown in Figure 6 and Figure 7.

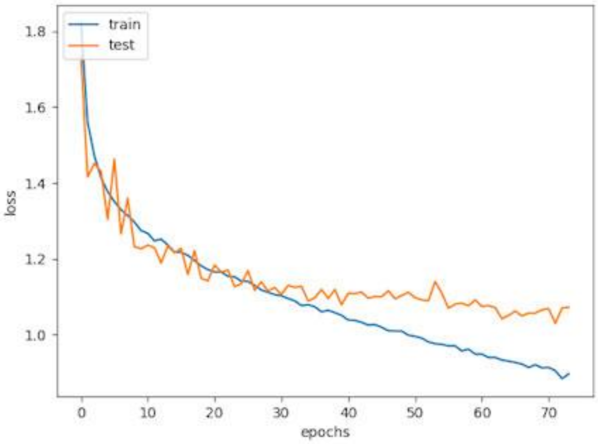

Figure 4. The epochs-loss curve of the MobileNet based model.

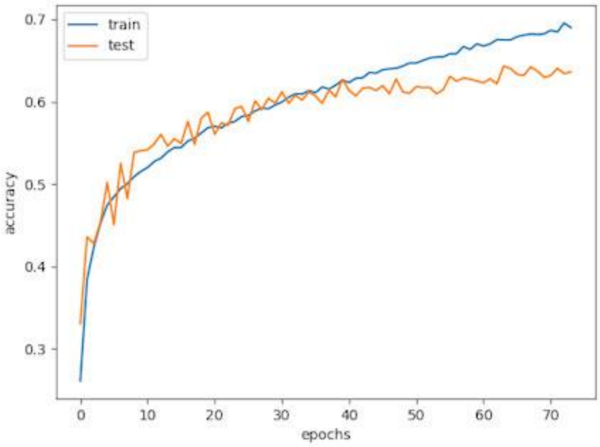

Figure 5. The epochs-accuracy curve of the MobileNet based model.

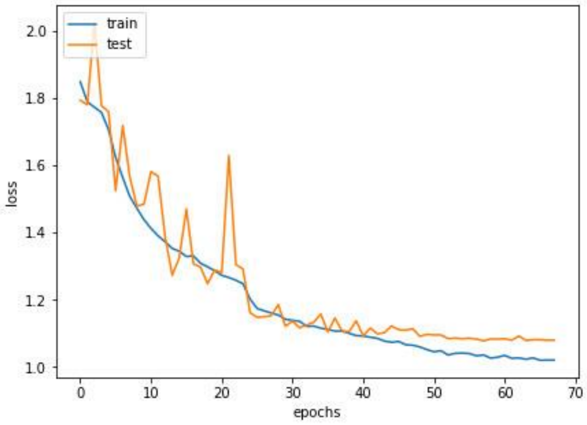

Figure 6. The epochs-loss curve of the self-build model

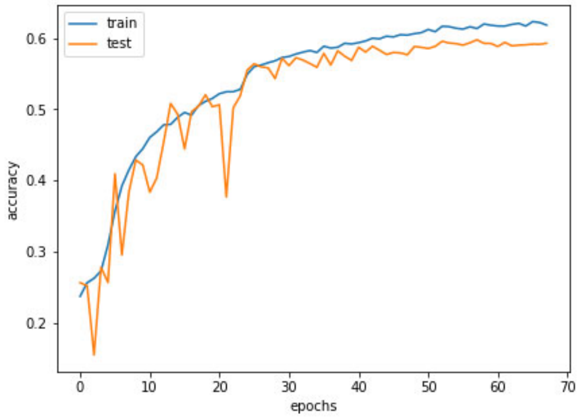

Figure 7. The epochs-accuracy curve of the self-build model

Based on the technique of confusion matrix, the performance of each label is shown in Figure 8 for the MobileNet based model and in Figure 9 for the self-build model. Both models exhibit the highest accuracy in predicting the "happy" label, with 1, 520 and 1, 434 correctly classified instances, respectively. However, it is noteworthy that neither model performs effectively in predicting the "disgust" label, as indicated by the absence of correct predictions for this particular category.

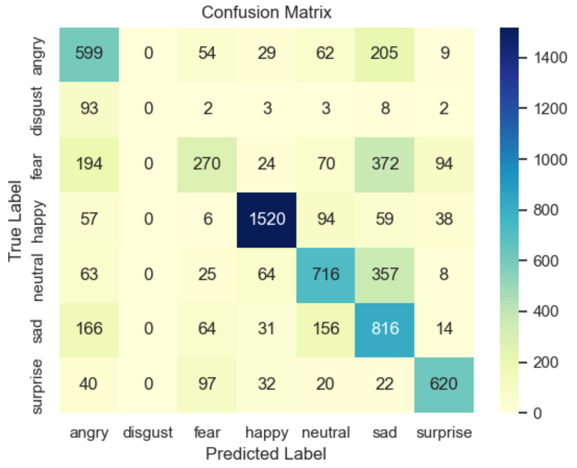

Figure 8. The confusion matrix of the MobileNet based model.

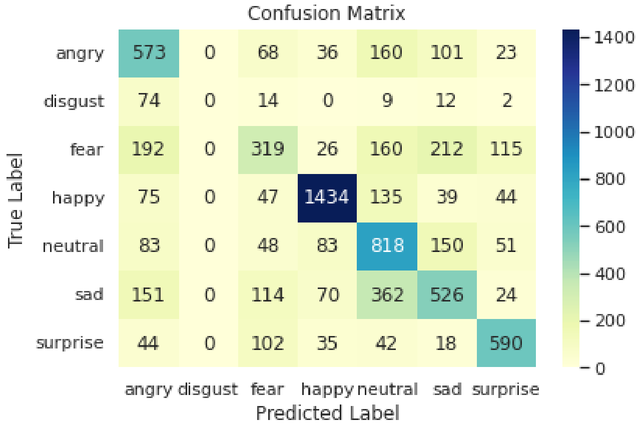

Figure 9. The confusion matrix of the self-build model

The saliency map is made according to the self-build model, shown in Figure 10. The weight-based profile is selected one for each label to show the vital part when predicting. For the "angry" label, the dominant location is the chain. For the "disgust", "fear", "sad" and "surprise" label, the dominant locations are the eyes and mouth. For the "happy" label, the dominant location is the part between nose and mouth. And for the "neutral" label, the dominant location is the whole forehead.

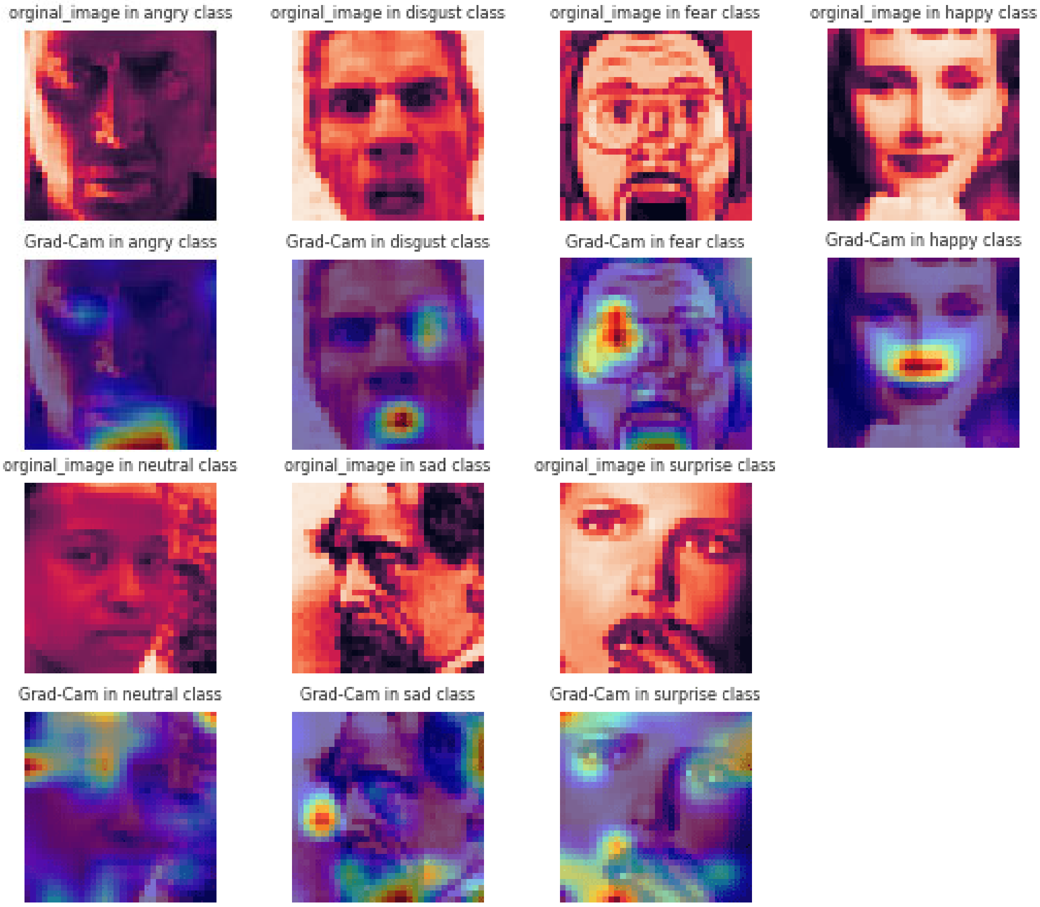
orginal_image in angry class
orginal_image in disgust class
orginal_image in fear class
orginal_image in happy class
Grad-Cam in angry class
Grad-Cam in disgust class
Grad-Cam in fear class
Grad-Cam in happy class
orginal_image in neutral class
orginal_image in sad class
orginal_image in surprise class
Grad-Cam in neutral class
Grad-Cam in sad class
Grad-Cam in surprise class

Figure 10. The saliency map of the self-build model.

3.2. Discussion

When considering the overall testing accuracy, both models demonstrated an accuracy about 60%, which can be regarded effective models in facial emotion classification. But when concentrating on each label separately, it can be easily found that the prediction of the “disgust” label never succeeds, having 0 correct prediction data. Due to this defect in particular label, the models can hardly be applied to practical applications, although they have a really good performance in predicting the “happy” label. The marked contrast between the best performance observed for the "happy" label and the worst performance for the "disgust" label warrants a detailed analysis of these two labels as representative cases.

By integrating the insights derived from the confusion matrix and the saliency map, two potential reasons contributing to the observed limitations in the models' performance can be identified. The first relates to the scarcity of training examples. By retrieving the training dataset, there are only 436 “disgust” label profiles, but there are up to 7215 profiles labelled “happy”, which has a nearly 16 times data volume difference. Despite data augmentation, the data volume difference between them still exists, leading to a huge accuracy discrepancy. For the improvements, the “disgust” label profiles should be expanded or augmented as many as the “happy” label has. The second reason can be attributed to intrinsic difficulties associated with certain emotion classifications. If a person is happy, for example, it is easy to imagine that he is smiling or laughing, the corner of his mouth is raised, it is extremely easy to identify. And this part, the region between nose and mouth, also has a high weight according to the saliency map. But talking to disgust, there is very little change on the facial emotions when a man feels disgust. And what’s worse is that the “disgust” feature, eyes and mouth, are vague as well, easy to be confused with other similar emotions like “angry” or “surprise”, which is consistent with the confusion matrix and saliency map. Addressing this issue necessitates the development of novel network architectures tailored to effectively capture and differentiate the intricate features associated with the "disgust" label. Further research and modification efforts are warranted to explore and overcome this challenging problem.

4. Conclusion

This study centers around the task of facial emotion classification, with a specific focus on analyzing the performance of the models employed and identifying areas for improvement. A MobileNet based model and a self-build model are applied in this paper to distinguish facial emotions and collect experimental results via the techniques of confusion matrix and saliency map. Through comprehensive examination and interpretation of these results, the defect of low accuracy in certain conditions is revealed, and two possible reasons are proposed, including few training data and vague feature. Although not offering specific improvement measures, the work of this paper provides deeper sight and more aspects for the future modification so that the constructed models can be adapted to more application scenarios.

References

[1] Ekman P Friesen W V 1978 Facial Action Coding System (FACS): a Technique for the Measurement of Facial Actions Rivista Di Psichiatria.

[2] Jonitta C et al 2020 Deep Learning based Facial Expression Recognition for Psychological Health Analysis 2020 International Conference on Communication and Signal Processing (ICCSP) (Chennai, India) pp 1155-1158.

[3] Shojaeilangari S et al 2015 Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning in IEEE Transactions on Image Processing vol 24 no 7 pp 2140-2152.

[4] Cao H et al 2021 Facial Expression Study Based on 3D Facial Emotion Recognition 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS) London, United Kingdom pp 375-381.

[5] Han B et al 2023 Masked FER-2013: Augmented Dataset for Facial Expression Recognition 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (Shanghai, China) pp 747-748.

[6] Kaggle 2020 FER-2013 https://www.kaggle.com/datasets/msambare/fer2013?select=test.

[7] Yu Q Chang C S Yan J L et al 2019 Semantic segmentation of intracranial hemorrhages in head CT scans 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) IEEE pp 112-115.

[8] Sharif Razavian A Azizpour H Sullivan J et al 2014 CNN features off-the-shelf: an astounding baseline for recognition Proceedings of the IEEE conference on computer vision and pattern recognition workshops pp 806-813.

[9] Wang S Y Wang O Zhang R et al 2020 CNN-generated images are surprisingly easy to spot. for now Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 8695-8704.

[10] Kayalibay B Jensen G van der Smagt P 2017 CNN-based segmentation of medical imaging data arXiv preprint arXiv:1701.03056 2017.