Journal of Robotics and Control (JRC)
Volume 6, Issue 3, 2025
ISSN: 2715-5072, DOI: 10.18196/jrc.v6i3.25804 

# Utilization of Convolutional Neural Network for Effective Recognition of Complex and Common Facial Emotions

Ammar Ibrahim Majeed [1], Suhad Qasim Naeem [2], Elaf A. Saeed [3*]
[1, 3] Department of Automation and Artificial Intelligence, College of Information Engineering, Al-Nahrain University, Baghdad, Iraq
[2] Department of Information and Communication Engineering, College of Information Engineering, Al-Nahrain University, Baghdad, Iraq
Email: [1] ammar.alkindy@nahrainuniv.edu.iq, [2] suhadqasim73@nahrainuniv.edu.iq, [3] elaf.ahmed@nahrainuniv.edu.iq
*Corresponding Author

***Abstract*—Facial expression recognition is an important area of computer vision used for human-computer interaction. The convolutional neural network model in this work was tested on the Fer-2013 dataset, and the experimental results demonstrated the superiority of the recognition rate. It is known that the Fer-2013 dataset contains data collected in an experimental environment, and to verify the generalization capability of model recognition, a self-made facial expression data set in a natural state was created, and the models are trained using this dataset to identify emotions from face photos, however, it has biases and limitations, including poor resolution (48 x 48 pixels) and class imbalance, which causes some emotions to be overrepresented. Additionally, it is devoid of demographic data, which may cause some groups to do poorly, furthermore, even though emotions are frequently mixed and context-dependent, it assumes that they are entirely distinct. More varied datasets, better class balance, the addition of demographic data, context, and sophisticated deep learning might all be employed to boost performance. also performed a series of pre-processing on the face images such as cropping, and pixel adjustment. The cropping is used to increase processing efficiency by removing extraneous portions of the image to highlight the crucial area. Normalization and contrast enhancement are examples of pixel manipulation that improves analysis and make the image more readable. The expression recognition results indicate that the model achieved an overall accuracy rate of 85.10% on the self-made natural expression dataset. Recognition accuracy was high for happy, neutral, and surprised expressions, while it was lower for disgust and fear expressions due to their variability and similarity in features. Because they have recognizable facial traits that are simple for models to identify—such as a grin for happiness or an open mouth for surprise—they are more accurate at identifying emotions of pleasure, neutrality, and surprise. On the other hand, the model's accuracy is lower for disgust and fear expressions since some of their characteristics are comparable to those of other emotions (for example, the resemblance between the expressions of fear and surprise) and differ from person to person, making it challenging to tell them apart. The confusion matrix highlights that fear expressions were often misidentified as a surprise, primarily due to pupil dilation in both expressions. The study concludes that the developed pre-training CNN model effectively recognizes facial expressions, demonstrating significant accuracy, particularly with certain emotions. Future work may focus on improving recognition rates for less distinct expressions and expanding the dataset for better generalization.**

***Keywords*—Expression Recognition; Convolutional Neural Network; Deep Learning; Confusion Matrix Analysis; Emotion Variability.***

## I. INTRODUCTION

With the development of artificial intelligence, human-computer interaction has been widely studied. Letting machines learn human emotions based on human expressions [1][2] is an important research part of human-computer interaction. Facial expression recognition is an interdisciplinary subject in a broad sense, and its research involves computer vision, graphics and image processing, and psychology. Research on facial expression recognition can promote better human-computer interaction and is also an indispensable part of human computer interactions. The purpose of studying facial expression recognition may lay in one or more of the following fields:

1) To better understand human emotions in human-computer interaction, thereby improving the experience of human-computer interaction;

2) To track and recognize facial expressions in video clips;

3) To research facial expression encoding mode, which is more conducive to transmitting and storing pictures of facial expressions.

Facial expression recognition has broad application prospects in security, psychology, medical care, customer satisfaction analysis, and online teaching. Human facial expressions have been studied since the 1970s and human expressions have been classified. This process is divided into feature extraction and expression classification in traditional expression recognition systems. The methods for extracting facial features include the Gabor filter [3][4]; direction histogram HOG [5]-[8]; discrete cosine transform (DCT) [9]-[12] and scale-invariant feature transform (SIFT) [13]-[16], etc., and then use SVM [17]-[20] or PCA [21]-[24] to perform facial expression Classification. However, the HOG algorithm has contrast and illumination allergies, which reduces its accuracy under various settings. Additionally, it has trouble with size changes and engineering distortions, and it requires a lot of calculations, which makes it unsuitable for

real-time use, particularly on constrained resources. Regarding the DCT method, excessive pressure causes data loss, which degrades image quality. Additionally, it has trouble representing sharp edges and is prone to confusion, which might result in irregularities in vision. Despite being less complex than some other transfers, using them in real-time applications can be somewhat demanding. Deep learning has been applied in expression recognition with the development of deep learning. The deep neural network model can extract image features and classification simultaneously, so it also brings great convenience to expression recognition. In computer vision [25]-[28], convolutional neural networks have better performance than other neural networks in processing graphics and images due to their own convolutional and pooling operations. Because deep learning can automatically extract features from photos without requiring manual feature construction, as is the case with more conventional approaches like HOG or DCT, it outperforms the other algorithms for face emotion recognition. In order to recognize expressions despite variations in lighting, angles, or particular facial characteristics, deep neural networks, like CNNS, rely on learning the intricate representations of faces. Compared to conventional algorithms, it is more precise and adaptable in handling a range of facial expressions since it can also use the vast amount of data and enhance its performance while expanding the volume of data. This paper designs a new convolutional neural network structure to extract and classify expression features. The model of this work draws on the idea of a VGG [29]-[32] network designs a convolutional network structure, and adjusts the parameters of the network structure. Inspired by the GoogleNet [33]-[36]network, a 1*1 convolution kernel [37]-[40]to the first layer of the convolutional neural network was added to increase the nonlinear representation of the input. Finally, in the fully connected layer, the complexity of the model was simplified by discarding some neurons.

The contributions of this work can be summarized in the following points:

1) A new convolutional neural network structure is designed based on the idea of a VGG network.

2) Train the model through the Fer-2013 dataset [41][42] and check the accuracy of model recognition.

The FER-2013 database suffers from biases that affect the generalizability of the models trained in it. The sample size may not be sufficient to represent all age and ethnic groups, which causes bias in the predictions. The diversity of the data is also limited, as it was collected under uniform imaging conditions, which reduces the model's ability to handle different lighting and angles in real-world environments. Additionally, FER-2013 lacks ecological validity, as its images are often unnatural, which makes the models less accurate when applied to real-world situations. Therefore, combining it with more diverse databases is preferable to improve performance.

3) Created a self-made data set of human facial expressions in a natural state, and verified the recognition generalization ability of the model based on the self-made dataset.

## II. RELATED WORK

The convolutional neural network is divided into two processes, forward propagation and back propagation. Forward propagation performs convolution and pooling operations. These two operations are to extract image features and process image features. Backpropagation uses the BP algorithm to transmit errors, thereby using the optimization algorithm to update model parameters.

In the 2012 ILSVRC challenge, Abraham M. et al. [43] used deep convolutional neural networks for image classification and achieved good results in the challenge. Since then, in image recognition, convolutional neural networks have been used extensively. Mohammed Aly et al. [44] studied convolution and pooling algorithms for facial expression recognition, pointed out some limitations of fixed pooling, and proposed a dynamic adaptive pooling algorithm. Huilin Ge et al. [45] respectively designed a convolutional neural network model for expression recognition, but the accuracy of expression recognition was not ideal. Ahmad B.A [46] et al. respectively designed a convolutional neural network that simultaneously recognizes gender and expression. Mohammad A. et al. [47] designed a parallel convolutional neural network to recognize expressions, reducing the training time during model training. To improve the accuracy of recognition, the convolutional neural network usually integrates another model for expression recognition [48]. For example, John L. et al. [48] merged the convolutional neural network and the support vector machine. The convolutional neural network only extracts features, and the support vector machine is used to replace the fully connected layer for classification. Naresh V. et al. [49] and Kadimi Naveen et al. [50] respectively proposed cross-connection convolutional neural networks. Different convolutional layers extract different features. Cross-connections are used to retain the features of different layers to improve the rate of recognition. Yali Yu et al. [51] used a convolutional neural network to extract facial expression features from different perspectives, so that the extracted features are more accurate and detailed and are more conducive to classifying expressions.

## III. CNN STRUCTURE DESIGN

A convolutional neural network [52][53] is a kind of neural network with unique advantages in extracting image features. Expression recognition belongs to classification supervised learning, which uses labeled expression pictures to train a convolutional neural network classification model. The forward propagation of the convolutional neural network model is convolution and pooling operations, the back-propagation algorithm is used to transfer errors, and the stochastic gradient descent (SGD) optimization algorithm [54]-[57] is used to train and optimize the parameters of the model. The convolutional neural network [58]-[61] designed in this article for expression recognition consists of an input layer, 4 convolutional layers, 3 pooling layers, 2 fully connected layers, and a SoftMax [62]-[65] layer. Its structure is shown in Fig. 1. The parameters of each convolutional layer, pooling layers, and fully connected layers are detailed in Table I.

*A. Convolution Layer*

The **convolutional** layer of the convolutional neural network performs convolution operations on facial expression pictures to extract facial expression features[66]. The input layer directly uses the image pixels as the input value and then performs a convolution operation on the input value. This article uses convolution kernels of different sizes in order to make features extraction. Convolution kernels of different sizes represent different receptive fields. Therefore, different convolution kernels are used to extract the features of expression of different receptive fields. The expression of the convolution layer is as in (1):

$$C_i = f(x * w_i + b_i) \quad (1)$$

Where, $C_i$ represents the output result obtained by the $i^{th}$ convolution, f(.) represents the activation function. The activation function selected the rectified linear unit function (Rectified Linear Units, ReLU), x represents the input image value, * represents convolution operation, and $w_i$ represents the $i^{th}$ Convolution kernel, $b_i$ represents the bias of the first convolution kernel. The expression of the ReLU function is as in:

$$ReLU\ (y) = \begin{cases} y, & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (2)$$

This work uses a total of 4 convolutional layers. The convolution kernel sizes are: 1*1, 5*5, 3*3, 3*3, and the number of convolution kernels is 32-32-64-128. After the convolution layer, the excitation layer is output. A 1*1 convolution kernel before the second layer of convolution is used to increase the nonlinear representation of the input, deepen the network structure of the model, and improves the expression capability of the model. The input of the image is a 48*48 matrix. After convolution with 32 (1*1) convolution kernels, 32 (48*48) feature maps are output. The second layer of convolution uses a 5*5 convolution kernel to first extract features in a large receptive field, and then reduce the size of the convolution kernel to extract features in a smaller area. 5*5 convolution kernels for convolution on the 48*48 feature map was used to obtain 32 (48-5+1)*(48-5+1) feature maps. Using 32 convolution kernels is the extraction 32 different local expression features are included. The third and fourth convolutional layers use 3*3 convolution kernels respectively[67]. The specific parameter values of each layer of the network are shown in Fig. 1. Each layer of the convolution operation of the convolutional neural network performs feature extraction. This paper fuses the features extracted by different convolution kernels in each layer and visually displays the extracted feature maps. A facial expression pictures in the Fer2013 dataset were used for demonstration[68]. The feature extraction for a feature map after one convolution operation is as shown in Fig. 2.

*B. Pooling Layer*

The **pooling layer** [69]-[72] of a convolutional neural network is usually designed after the convolutional layer. The number of feature maps will increase as the number of convolutional layers increases. However, the increase in feature dimension will cause a dimensionality disaster, so it is usually added after the convolutional layer. The pooling layer is used for dimensionality reduction. This work uses the maximum pooling operation to maintain the most salient features in a pooled area[73]. The pooling layer can be expressed as in:

$$S_i = down\left(\max\left(y_{a,b}\right)\right) \quad a, b \in p_i \quad (3)$$

Where, Si represents the maximum pooling result of the $i_{th}$ pooling area, down (·) represents the down sampling process (retaining the maximum value of the pooling area), $y_{a,b}$ represents the value in the pooling area, and pi represents the $i_{th}$ pooling area. The third layer of the network structure in this work is the pooling layer, and the feature map input to this layer is 44*44. The pooling area is 2*2, so in the feature map, 2*2 represents a pooling window, and each pooling window results in a maximum pooling result. Therefore, the final pooling result of the feature map is (44/2)*(44/2).

*C. Fully Connected Layer*

In the **fully connected layer**, the neurons, are connected to the ones in the previous layer, thereby converting the feature dimensions into one-dimensional data. The last pooling layer in this work is connected to the fully connected layer. The last pooling layer outputs 128 convolutional 5*5 feature maps, which are converted into one-dimensional data: 128×5×5 = 3200, and then input 1×3200 data into the fully connected layer[74]. The fully connected layer, as in:

$$Full = f(w \times z + b) \quad (4)$$

Where, Full represents the output result of the fully connected layer, f(·) is the ReLU activation function, w rep-resents the weight value of the connection, z is the value input to the fully connected layer, and b is the bias. To reduce the complexity of the network structure and prevent over-fitting, random deactivation (Dropout) of neurons is used[75].

*D. SoftMax*

The last layer of the network structure in this work is the SoftMax function to classify the 7 facial expressions. There are 7 neurons in this layer, and each neuron represents an expression category. For each input face picture, the 7 neurons in the SoftMax layer input the prob-ability between 0 and 1, and the neuron with the largest input probability value, it means that the expression probability corresponding to this neuron is the highest. The representation of SoftMax classification as in:

$$p(\ y = c|\ m\ ;w) = \frac{e^{w_c \times m}}{\sum_{i=1}^{k} e^{w_i \times m}} \quad (5)$$

Where represents the probability that the input picture m is the expression type *c*, w is the weight parameter value (to be fitted), and k is the total number of categories, 7. The value of expression type c is {0, 1, 2, 3, 4, 5, 6}.

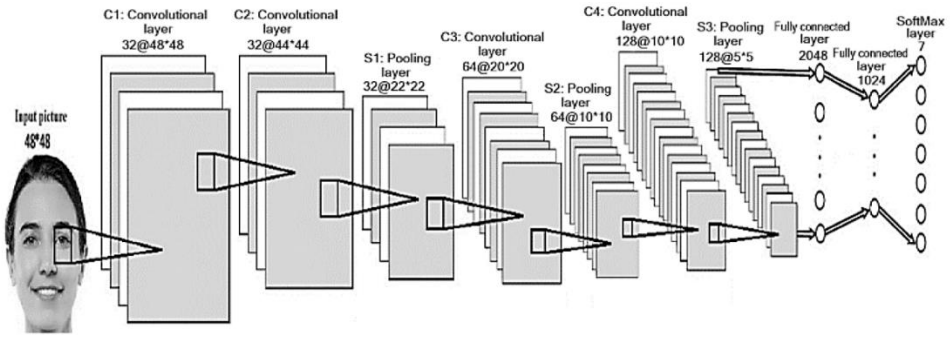


Fig. 1. The proposed structure of the convolutional neural network

TABLE I. PARAMETERS OF CONVOLUTIONAL NEURAL NETWORK STRUCTURE

| *Network layer* | *Input dimension* | *Convolution kernel size* | *Pooling area* | *Step size* | *With or without Filling* | *Output dimension* |
|---|---|---|---|---|---|---|
| Layer-1 (convolution) | 48*48 | 32@1*1 | | 1 | None | 32@48*48 |
| Layer-2 (convolution) | 32@48*48 | 32@5*5 | | 1 | None | 32@44*44 |
| Layer-3 (pooling) | 32@44*44 | | 2*2 | 2 | None | 32@22*22 |
| Layer-4 (convolution) | 32@22*22 | 64@3*3 | | 1 | None | 64@20*20 |
| Layer-5 (pooling) | 64@20*20 | | 2*2 | 2 | None | 64@10*10 |
| Layer-6 (convolution) | 64@10*10 | 128@3*3 | | 1 | Yes | 128@10*10 |
| Layer-7 (pooling) | 128@10*10 | | 2*2 | 2 | None | 128@5*5 |
| Layer-8 | 1*3200 | | fully connected layer Dropout (0.6) | | | 1*2048 |
| Layer-9 | 1*2048 | | fully connected layer Dropout (0.4) | | | 1*1024 |
| Layer-10 | 1*1024 | | SoftMax layer | | | 1*7 |

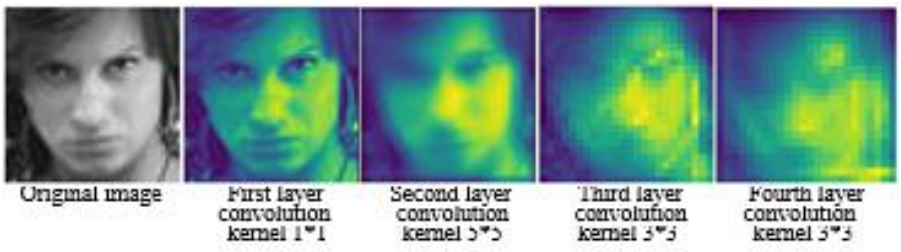


Fig. 2. Feature after convolution

## IV. EXPERIMENT

The experiments in this work are implemented in Python and are based on the deep learning platform of Keras. In addition, Python was also used to reproduce the two models in the two papers for comparison. In order to make a fair comparison of the experimental results, a unified data set was used for training the different models.

### *A. Dataset*

This article uses two data sets, one is the Fer2013 [76] facial expression dataset, and the other is produced as a part of this paper work. The Fer2013 expression data-base has 35,886 facial expression pictures, including 28,708 in the training set, 3589 in the verification set and 3,589 in the test set [77]. The size of each grayscale image is 48*48. There are 7 expressions in the data set: angry, disgusted, fearful, happy, sad, surprised, and neutral[78].Since skin color variations play a part in emotional expression, the grayscale in FER-2013 reduces color information that may aid identify some emotions, including rage or shyness, which impacts the models' accuracy. Additionally, by eliminating RGB channel differences, it decreases the number of visual cues that are available, which may make it more challenging to discern between identical feelings. In comparison to utilizing color photos, grayscale results in some detail loss that might compromise accuracy, even while it speeds up training and reduces the model's sensitivity to light.

The Fer2013 data set was collected in a laboratory environment, so it cannot well verify the model's recognition of human expressions in the natural state. The quality and generalizability of models trained on the FER-2013 database may be impacted by a number of biases and constraints. First, a model that favors some groups that are better represented in the data may result from the sample size not being large enough to capture all potential differences in facial expressions across age and ethnic groups. Second, the model may not be able to identify expressions in more complicated realistic circumstances, including those with varying lighting or varied filming angles, because the data may not be as diverse as it might be because the photographs are taken in uniform shooting conditions or from certain locales. Another issue with ecological validity is that FER-2013 photos are frequently cropped and artificial in comparison to faces in real-world settings. This might result in a model that is unable to generalize when used in actual scenarios like in-person interactions. Because of these factors, FER-2013 should be used with caution, and in order to obtain more accurate and dependable results, it might be essential to integrate it with other, more varied databases.

Therefore, a search for some pictures of human expressions in the natural state from the Internet was performed and then analyzed the pictures. The size, pixels, background, etc. are preprocessed and the pictures are uniformly converted into grayscale images. To ensure that human variances are as well represented as possible, the data was gathered from a range of sources. The photographs were taken from open-source datasets or captured by cameras in public areas or interactive settings like workplaces and schools. This was done in order to capitalize on media like TV interviews, documentaries, or even recordings of unplanned discussions that feature natural face expression situations. This guarantees that the facial expressions are representative of real-life situations rather than only being recorded in simulated settings. Since low-resolution photos can cause the loss of crucial details like small wrinkles and muscle movements, raising the quality of the photographs is crucial to increasing the model's accuracy. High-resolution photos were taken while accounting for various backdrops, lighting situations, and shooting angles. Images that were blurry, clipped, or deformed were also eliminated since they can have an impact on how well the model recognizes expressions. A team of behavioral psychologists or specialists in facial expression analysis helped identify the appropriate categories for every photograph in the new database, which has been categorized more accurately than FER-2013. Since some emotion categories, like happiness and anger, are frequently overrepresented compared to others, like disgust or surprise, the various emotion categories within the data were balanced. Therefore, it is crucial to make sure that the distribution is as balanced as feasible and that there are an equal number of samples for each expression. This is made possible by the Data Augmentation technique, which uses a variety of transformations, including rotation, lighting changes, and color adjustments, to increase the number of images in underrepresented categories without requiring the manual collection of additional images.

Finally, a small data set is formed. The facial expressions in the self-made data set are divided into 7 types of expressions, with a total of 396 pictures. This work uses the above two data sets to jointly verify the performance of the suggested convolutional neural network model[79].

### *B. Model Training*

To train a more accurate model and use expression pictures more efficiently, the expression data library was amplified through a series of random transformations, such as these shown in Fig. 3.

The loss function used in this work is a multi-class cross-entropy loss function, the loss function is as in:

$$loss = -\sum_{i=1}^{n} y_{i1} \log a_{i1} + y_{i2} \log a_{i2} + \ldots + y_{i7} \log a_{i7} \quad (6)$$

Where a is the actual output value of the neuron and y is the expected output value.

The training goal is to minimize the loss value, use the backpropagation algorithm to propagate the error value, and use the SGD optimization algorithm to update the parameter values along the direction of gradient descent. The SGD algorithm is as in:

$$\frac{\partial loss}{\partial \theta_{i1}} = -\sum_{i=1}^{n} \frac{\theta_{i1}}{a_{i1}} \quad (7)$$

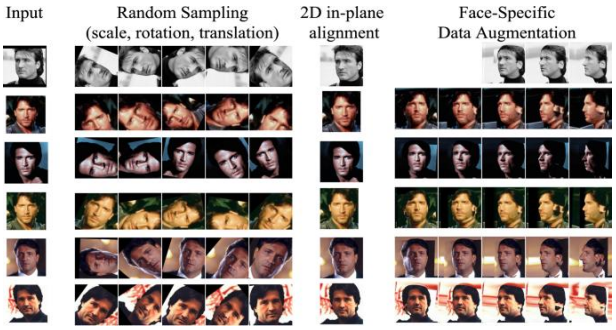


Fig. 3. Data augmentation

Therefore, the parameters are updated as in:

$$\theta_j = \theta_j - a \frac{\partial loss}{\partial \theta_j} \quad (8)$$

Where $\theta_j$ is the parameter to be updated, a is the learning rate, $\frac{\partial loss}{\partial \theta_j}$ is the value to decrease in the gradient descent direction. The learning rate in this model is 0.01. To allow the training to converge to the best result, this work sets the learning rate to gradually decay as the number of training times increases, so the learning step size gradually decreases.

Because they have a direct impact on the CNN model's performance, training speed, and generalization ability, the hyperparameters in its design are carefully set. These comprise fundamental parameters including the number of layers, batch size, learning rate, number of nodes in the hidden layers, kernel size, number of filters, and number of layers. The job and the type of data are taken into consideration while choosing these parameters. The model's capacity to catch features, for instance, is influenced by the kernel size; bigger kernels capture broader patterns, whereas smaller kernels (such as 3x3) identify finer details. How many features are retrieved per layer depends on the number of filters; more filters increase the model's capacity but also raise its computational cost. The model's convergence speed is influenced by the learning rate; a high learning rate expedites training but may cause instability, whereas a low learning rate guarantees stability but slows convergence. Several methods are employed to improve these hyperparameters, including Random Search, which chooses values at random to save time, and Grid Search, which methodically tries every conceivable combination. Hyperband and Bayesian Optimization are more sophisticated techniques that speed up the search for ideal values. Hyperparameter selection may also be automated by utilizing AutoML and Reinforcement Learning. To guarantee an effective and broadly applicable CNN model, it is necessary to balance accuracy, training speed, and computational complexity while selecting and fine-tuning hyperparameters.

This work first uses the training set in the Fer2013 data set to train the model and then uses the verification set to verify the accuracy of the recognition. When the accuracy of the verification set decreases and the loss value increases, the training is stopped.

## V. EXPERIMENTAL RESULTS AND ANALYSIS

This article reproduces the convolutional neural network model proposed by Lu et al.[80] and the LeNet-5 model proposed in [81] using Python. The Fer2013 data set was used to train and the accuracy was calculated on the test set. The training results of the model in this paper, the model of Lu [80] et al, and the model [81], are as shown in Fig. 4 to Fig. 6.

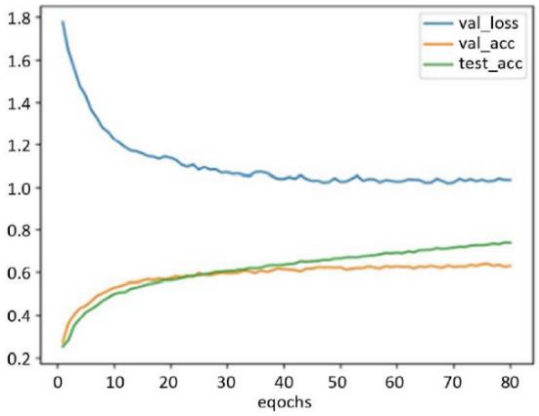


Fig. 4. Training results of the suggested model

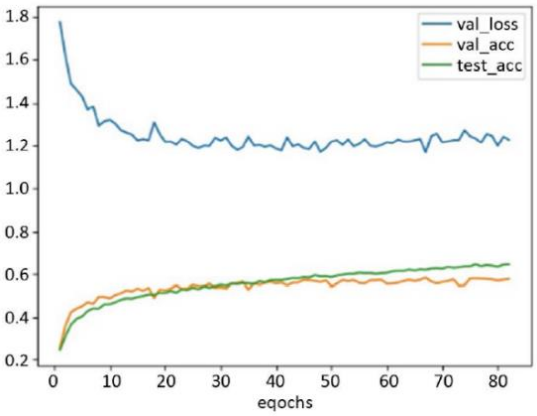


Fig. 5. Training results of the Lu et al. model

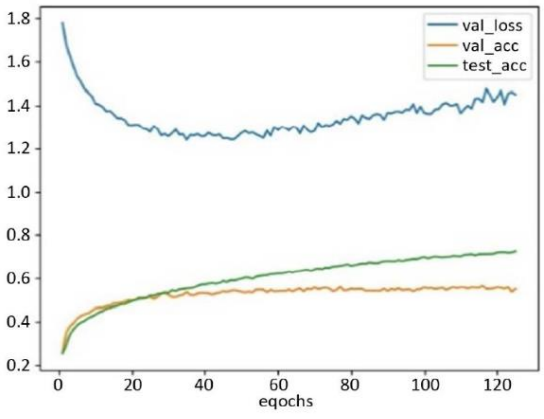


Fig. 6. Training results of the Li et al. model

Due to class imbalance and grayscale image restrictions, the FER-2013 dataset exhibits substantial variability. Model performance is impacted by standard deviation, which highlights discrepancies in pixel intensity and class distributions. Wider intervals suggest instability, and confidence intervals aid in evaluating the accuracy and F1-scores' dependability. High cross-validation variance highlights dataset discrepancies and indicates sensitivity to data splits. Model stability and generalization are enhanced by addressing these problems using balancing strategies and data augmentation.

For the training results, we selected the training models with the highest accuracy on the verification set. Table II summarizes the accuracy of various models in the test set. It can be seen from this table that the suggested model's accuracy on the test set is relatively high, at 72.92%.

The model in this work is trained with the Fer2013 data set saves the trained model, and then uses the trained model to identify the pictures in the self-made data set. For example, the single picture recognition result is shown in Fig. 7.

TABLE II. ACCURACY OF EACH MODEL

| Document | Number of iterations | Validation set accuracy | Test set accuracy |
|---|---|---|---|
| Ref. [7] | 76 | 0.5811 | 0.6455 |
| Ref. [8] | | | 0.7074 |
| Ref. [14] | 116 | 0.5646 | 0.7142 |
| The suggested model | 76 | 0.6400 | 0.7292 |

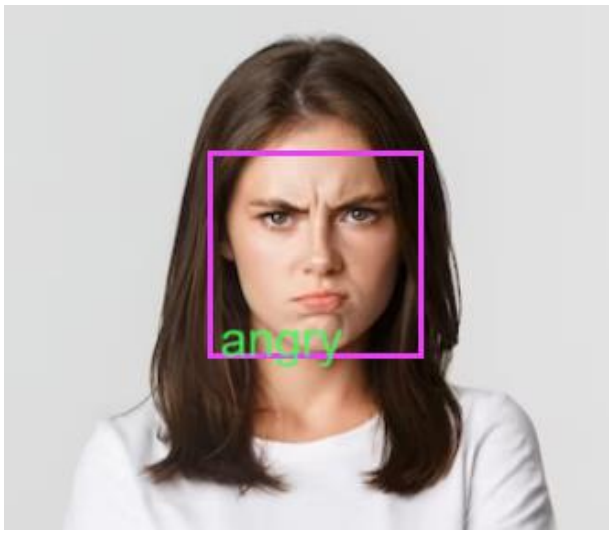


Fig. 7. Expression recognition results

The entire self-made dataset was identified, and the confusion matrix of the identification results is shown in Table III. From the confusion matrix, one can see that the recognition accuracy of happy, neutral, and surprised expressions is relatively high. Still, the recognition effect is relatively poor when identifying disgust and fear expressions. Regarding disgust expressions, everyone has different expressions and facial expressions are also very different. Therefore, when recognizing disgust expressions, the recognition results are relatively scattered and may be recognized into various expressions.

TABLE III. CONFUSION MATRIX FOR IDENTIFYING SELF-MADE DATASETS

| | | Predict | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral | Recognition Rate |
| Actual | Angry | 33 | 0 | 3 | 0 | 0 | 1 | 0 | 89.19% |
| | Disgust | 5 | 10 | 0 | 2 | 3 | 0 | 2 | 45.45% |
| | Fear | 0 | 0 | 22 | 0 | 1 | 13 | 2 | 57.89% |
| | Happy | 1 | 0 | 1 | 134 | 0 | 0 | 5 | 95.04% |
| | Sad | 0 | 0 | 2 | 1 | 20 | 0 | 7 | 66.66% |
| | Surprise | 0 | 0 | 1 | 2 | 0 | 39 | 1 | 90.70% |
| | Neutral | 1 | 0 | 2 | 2 | 1 | 0 | 79 | 92.94% |

When recognizing fear expressions, it is easier to recognize surprise, mainly when extracting features of the eyes. Both fear and surprise tend to make people's pupils dilate, so fear will be recognized as a surprise. The overall accuracy rate on the self-made natural expression data set is 337/396 = 85.10%.

The suggested model outperforms the compared models Because it makes good use of data augmentation techniques, which increase the variety of the training dataset and boost recognition accuracy, the proposed model performs better than the other models. To improve feature extraction and model convergence, it also uses a multi-class cross-entropy loss function and optimizes parameters using the SGD technique. The architecture of the model is also designed to capture unique face features, which increases the accuracy of identifying emotions with recognizable facial features.

In expression recognition, we analyzed several difficulties in facial expression and expression recognition. Human beings are complex animals with a rich inner world. The expressions on human faces are sometimes intertwined with multiple emotions, such as facial expressions may contain multiple expressions such as surprise, anger, and helplessness at the same time, which makes recognition difficult. Sometimes different expressions of human beings may express the same emotion, and the same expression may have different emotions for different people, which requires strict extraction of subtle features of human faces. Finally, human facial features have their own characteristics and cannot be generalized. For example, in the expression recognition process, the expressions of people with big eyes are more likely to be recognized as surprise or fear.

Challenges in facial expression detection, such as misinterpreting fear as a surprise or the wide range of disgust emotions, can affect applications in security, healthcare, and human-computer interaction. Using sophisticated deep learning models (such as multimodal approaches, attention mechanisms, and Transformers), increasing dataset diversity, and refining training methods like discriminative learning, data augmentation, and feature embedding are some solutions. By increasing model accuracy, these enhancements guarantee more trustworthy emotion identification in practical applications.

Overfitting, in which the model memorizes training examples rather than generalizing to new data, is a consequence of a small number of pictures in a dataset. Additionally, it lessens class variety, which makes it more difficult for the model to pick up on differences in angles, lighting, and emotions. Because of this, the model generates erratic and skewed predictions, particularly for emotions that are underrepresented, which reduces overall accuracy and dependability in practical applications.

Future studies should concentrate on growing the dataset to include a more excellent range of facial expressions from various populations and situations in order to enhance the detection of difficult expressions like fear and contempt. By applying sophisticated methods like transfer learning from bigger, more varied datasets, the model's capacity to generalize and precisely detect minute variations in these emotions may be improved. Furthermore, temporal analysis using video data may be able to better distinguish between comparable emotions like surprise and terror by capturing the dynamic character of facial expressions.

## VI. CONCLUSION

Without manual extraction, the convolutional neural network can automatically and implicitly learn the properties of facial expressions. The model can be trained using the image's pixels as input. This paper takes advantage of the advantages of convolutional neural networks in processing images and designs a convolutional neural network structure model for facial expression recognition. The expression dataset Fer2013 is used to train the model, and the experimental results on this dataset demonstrate the superiority of the proposed method. In addition, the model's recognition generalization ability on a self-made natural state data set was tested, and the generalization ability is relatively good. Convolutional neural networks require a large number of data sets to train, and the trained model can learn good classification effects. Therefore, the expression recognition model will be more generalizable if more images of facial expressions in their natural settings are gathered to train it.

## ACKNOWLEDGMENTS

The authors express gratitude to Al-Nahrain University and the College of Information Engineering for their essential support and favorable environment for conducting this research.

## REFERENCES

[1] R. Vempati and L. D. Sharma, "A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence," *Results in Engineering*, vol. 18, p. 101027, 2023.

[2] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput Appl*, vol. 35, no. 32, pp. 23311–23328, 2023.

[3] R. Aliradi and A. Ouamane, "A novel descriptor (LGBQ) based on Gabor filters," *Multimed Tools Appl*, vol. 83, no. 4, pp. 11669–11686, 2024.

[4] U. A. Bhatti *et al.*, "Local similarity-based spatial–spectral fusion hyperspectral image classification with deep CNN and Gabor filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[5] A. Kharat, P. Garje, and R. Wantmure, "Local approaches in face recognition: a case study using histogram of oriented gradients (HOG) technique," *NCRD's Technical Review: e-Journal*, vol. 8, no. 1, pp. 1–12, 2023.

[6] M. Sheriff, S. Jalaja, K. T. Dinesh, J. Pavithra, Y. Puja, and M. Sowmiya, "Face emotion recognition using histogram of oriented gradient (hog) feature extraction and neural networks," in *Recent Trends in Computational Intelligence and Its Application*, pp. 114–122, 2023.

[7] N. R. Pradeep and J. Ravi, "An Accurate Fingerprint Recognition Algorithm based on Histogram Oriented Gradient (HOG) Feature Extractor," *International Journal of Electrical Engineering & Technology (IJEET)*, vol. 12, no. 2, pp. 19–25, 2021.

[8] L. Mao and L. Tang, "Pedestrian detection based on gradient direction histogram," in *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pp. 939–943, 2022.

[9] W. Burger and M. J. Burge, "The discrete cosine transform (DCT)," in *Digital Image Processing: An Algorithmic Introduction*, pp. 589–597, 2022.

[10] C. Scribano, G. Franchini, M. Prato, and M. Bertogna, "DCT-former: Efficient self-attention with discrete cosine transform," *J Sci Comput*, vol. 94, no. 3, p. 67, 2023.

[11] X. Shen *et al.*, "Dct-mask: Discrete cosine transform mask representation for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8720–8729, 2021.

[12] I. F. Ince, F. Bulut, I. Kilic, M. E. Yildirim, and O. F. Ince, "Low dynamic range discrete cosine transform (LDR-DCT) for high-performance JPEG image compression," *Vis Comput*, vol. 38, no. 5, pp. 1845–1870, 2022.

[13] W. Burger and M. J. Burge, "Scale-invariant feature transform (SIFT)," in *Digital Image Processing: An Algorithmic Introduction*, pp. 709–763, 2022.

[14] M. A. Taha, H. M. Ahmed, and S. O. Husain, "Iris Features Extraction and Recognition based on the Scale Invariant Feature Transform (SIFT)," *Webology*, vol. 19, no. 1, pp. 171–184, 2022.

[15] P. Podder, M. R. H. Mondal, and J. Kamruzzaman, "Iris feature extraction using three-level Haar wavelet transform and modified local binary pattern," in *Applications of Computational Intelligence in Multi-Disciplinary Research*, pp. 1-15, 2022.

[16] T. A. Al-Shurbaji, K. A. AlKaabneh, I. Alhadid, and R. Masa'deh, "An optimized scale-invariant feature transform using chamfer distance in image matching," *Intelligent Automation & Soft Computing*, vol. 31, no. 2, pp. 971–985, 2022.

[17] S. Goyal, "Effective software defect prediction using support vector machines (SVMs)," *International Journal of System Assurance Engineering and Management*, vol. 13, no. 2, pp. 681–696, 2022.

[18] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *International Journal of Information Technology*, vol. 13, no. 5, pp. 1–11, 2021.

[19] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting," *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, 2023.

[20] Z. H. Kok, A. R. M. Shariff, M. S. M. Alfatni, and S. Khairunniza-Bejo, "Support vector machine in precision agriculture: a review," *Comput Electron Agric*, vol. 191, p. 106546, 2021.

[21] A. R. Zhang, T. T. Cai, and Y. Wu, "Heteroskedastic PCA: Algorithm, optimality, and applications," *The Annals of Statistics*, vol. 50, no. 1, pp. 53–80, 2022.

[22] J. Liu *et al.*, "A spatial distribution–Principal component analysis (SD-PCA) model to assess pollution of heavy metals in soil," *Science of The Total Environment*, vol. 859, p. 160112, 2023.

[23] D. Huang, F. Jiang, K. Li, G. Tong, and G. Zhou, "Scaled PCA: A new approach to dimension reduction," *Manage Sci*, vol. 68, no. 3, pp. 1678–1695, 2022.

[24] L. C. Lee and A. A. Jemain, "On overview of PCA application strategy in processing high dimensionality forensic data," *Microchemical Journal*, vol. 169, p. 106608, 2021.

[25] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[26] L. Zhou, L. Zhang, and N. Konz, "Computer vision techniques in manufacturing," *IEEE Trans Syst Man Cybern Syst*, vol. 53, no. 1, pp. 105–117, 2022.

[27] L. Yuan *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[28] M. Goldblum *et al.*, "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks," *Adv Neural Inf Process Syst*, vol. 36, 2024.

[29] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: a case study on early detection of a rice disease," *Agronomy*, vol. 13, no. 6, p. 1633, 2023.

[30] G. S. Nugraha, M. I. Darmawan, and R. Dwiyansaputra, "Comparison of CNN's Architecture GoogleNet, AlexNet, VGG-16, Lenet-5, Resnet-50 in Arabic Handwriting Pattern Recognition," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2023.

[31] A. Bagaskara and M. Suryanegara, "Evaluation of VGG-16 and VGG-19 deep learning architecture for classifying dementia people," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 1–4, 2021.

[32] X. Zhang, "The AlexNet, LeNet-5 and VGG NET applied to CIFAR-10," in *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pp. 414–419, 2021.

[33] M. S. Al-Huseiny and A. S. Sajit, "Transfer learning with GoogLeNet for detection of lung cancer," *Indonesian Journal of Electrical Engineering and computer science*, vol. 22, no. 2, pp. 1078–1086, 2021.

[34] S.-H. Chen, Y.-L. Wu, C.-Y. Pan, L.-Y. Lian, and Q.-C. Su, "Breast ultrasound image classification and physiological assessment based on GoogLeNet," *J Radiat Res Appl Sci*, vol. 16, no. 3, p. 100628, 2023.

[35] L. Yang *et al.*, "GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases," *Comput Electron Agric*, vol. 204, p. 107543, 2023.

[36] N. Yang, Z. Zhang, J. Yang, Z. Hong, and J. Shi, "A convolutional neural network of GoogLeNet applied in mineral prospectivity prediction based on multi-source geoinformation," *Natural Resources Research*, vol. 30, no. 6, pp. 3905–3923, 2021.

[37] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11963–11975, 2022.

[38] M. Ansari, S. Homayouni, A. Safari, and S. Niazmardi, "A new convolutional kernel classifier for hyperspectral image classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 11240–11256, 2021.

[39] H. Son, J. Lee, S. Cho, and S. Lee, "Single image defocus deblurring using kernel-sharing parallel atrous convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2642–2650, 2021.

[40] B. Shiri and D. Baleanu, "All linear fractional derivatives with power functions' convolution kernel and interpolation properties," *Chaos Solitons Fractals*, vol. 170, p. 113399, 2023.

[41] "FER-2013." Accessed: Jan. 31, 2025. [Online]. Available: https://www.kaggle.com/datasets/msambare/fer2013

[42] A. A. O. Díaz, S. C. Tamayo, D. N. De Oliveira, and G. A. Abensur, "Models for Real-Time Emotion Classification: FER-2013 Dataset," in *Intelligent Systems Conference*, pp. 289–304, 2024.

[43] A. M. Obeso, J. Benois-Pineau, M. S. G. Vázquez, and A. Á. R. Acosta, "Visual vs internal attention mechanisms in deep neural networks for image classification and object detection," *Pattern Recognit*, vol. 123, p. 108411, 2022.

[44] M. Aly, A. Ghallab, and I. S. Fathi, "Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model.," *IEEE Access*, vol. 11, pp. 121419–121433, 2023.

[45] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Comput Methods Programs Biomed*, vol. 215, p. 106621, 2022.

[46] A. B. A. Hassanat *et al.*, "DeepVeil: deep learning for identification of face, gender, expression recognition under veiled conditions," *Int J Biom*, vol. 14, no. 3–4, pp. 453–480, 2022.

[47] M. A. Hossain and B. Assiri, "Facial expression recognition based on active region of interest using deep learning and parallelism," *PeerJ Comput Sci*, vol. 8, p. e894, 2022.

[48] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics (Basel)*, vol. 11, no. 23, p. 3935, 2022.

[49] N. Vedhamuru, R. Malmathanraj, and P. Palanisamy, "Lightweight deep and cross residual skip connection separable CNN for plant leaf diseases classification," *J Electron Imaging*, vol. 33, no. 3, p. 33035, 2024.

[50] K. N. Kumar Tataji, M. N. Kartheek, and M. V. N. K. Prasad, "CC-CNN: A cross connected convolutional neural network using feature level fusion for facial expression recognition," *Multimed Tools Appl*, vol. 83, no. 9, pp. 27619–27645, 2024.

[51] Y. Yu, H. Huo, and J. Liu, "Facial expression recognition based on multi-channel fusion and lightweight neural network," *Soft comput*, vol. 27, no. 24, pp. 18549–18563, 2023.

[52] M. M. Taye, "Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions," *Computation*, vol. 11, no. 3, p. 52, 2023.

[53] X. Zhang, X. Zhang, and W. Wang, "Convolutional neural network," in *Intelligent Information Processing with Matlab*, pp. 39–71, 2023.

[54] Y. Tian, Y. Zhang, and H. Zhang, "Recent advances in stochastic gradient descent in deep learning," *Mathematics*, vol. 11, no. 3, p. 682, 2023.

[55] A. Sclocchi and M. Wyart, "On the different regimes of stochastic gradient descent," *Proceedings of the National Academy of Sciences*, vol. 121, no. 9, p. e2316301121, 2024.

[56] S. Pu, A. Olshevsky, and I. C. Paschalidis, "A sharp estimate on the transient time of distributed stochastic gradient descent," *IEEE Trans Automat Contr*, vol. 67, no. 11, pp. 5900–5915, 2021.

[57] S.-Y. Zhao, Y.-P. Xie, and W.-J. Li, "On the convergence and improvement of stochastic normalized gradient descent," *Science China Information Sciences*, vol. 64, pp. 1–13, 2021.

[58] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artif Intell Rev*, vol. 57, no. 4, p. 99, 2024.

[59] A. Patil and M. Rane, "Convolutional neural networks: an overview and its applications in pattern recognition," *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2020, Volume 1*, pp. 21–30, 2021.

[60] S. Wei, Y. Chen, Z. Zhou, and G. Long, "A quantum convolutional neural network on NISQ devices," *AAPPS bulletin*, vol. 32, pp. 1–11, 2022.

[61] X. Shen *et al.*, "Deepmad: Mathematical architecture design for deep convolutional neural network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6163–6173, 2023.

[62] S. Mehra, G. Raut, R. Das Purkayastha, S. K. Vishvakarma, and A. Biasizzo, "An empirical evaluation of enhanced performance softmax function in deep learning," *IEEE Access*, vol. 11, pp. 34912–34924, 2023.

[63] S. Raghuram, A. S. Bharadwaj, S. K. Deepika, M. S. Khadabadi, and A. Jayaprakash, "Digital implementation of the softmax activation function and the inverse softmax function," in *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pp. 64–67, 2022.

[64] Y. Zhang, L. Peng, L. Quan, Y. Zhang, S. Zheng, and H. Chen, "High-precision method and architecture for base-2 softmax function in DNN training," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 8, pp. 3268–3279, 2023.

[65] D. Han *et al.*, "Bridging the divide: Reconsidering softmax and linear attention," *Adv Neural Inf Process Syst*, vol. 37, pp. 79221–79245, 2024.

[66] Y. He, "Facial expression recognition using multi-branch attention convolutional neural network," *Ieee Access*, vol. 11, pp. 1244–1253, 2022.

[67] Y. Xie, H. Chen, Y. Ma, and Y. Xu, "Automated design of CNN architecture based on efficient evolutionary search," *Neurocomputing*, vol. 491, pp. 160–171, 2022.

[68] A. Kashef and J. Ma, "Real-time Facial Emotion Recognition Using FER2013 Image Dataset," in *IIE Annual Conference. Proceedings*, *Institute of Industrial and Systems Engineers (IISE)*, pp. 1–9, 2023.

[69] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study," *Neural Comput Appl*, vol. 34, no. 7, pp. 5321–5347, 2022.

[70] Y.-D. Zhang, S. C. Satapathy, S. Liu, and G.-R. Li, "A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis," *Mach Vis Appl*, vol. 32, pp. 1–13, 2021.

[71] D. Grattarola, D. Zambon, F. M. Bianchi, and C. Alippi, "Understanding pooling in graph neural networks," *IEEE Trans Neural Netw Learn Syst*, vol. 35, no. 2, pp. 2708–2718, 2022.

[72] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 887–904, 2022.

[73] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Repvit: Revisiting mobile cnn from vit perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15909–15920, 2024.

[74] S. R. Waheed, M. S. M. Rahim, N. M. Suaib, and A. A. Salim, "CNN deep learning-based image to vector depiction," *Multimed Tools Appl*, vol. 82, no. 13, pp. 20283–20302, 2023.

[75] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, and S. Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," *Multimed Tools Appl*, vol. 83, no. 6, pp. 15711–15732, 2024.

[76] "FER2013 Dataset | Papers With Code." Accessed: Jan. 31, 2025. [Online]. Available: https://paperswithcode.com/dataset/fer2013

[77] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi," in *2020 Fifth international conference on informatics and computing (ICIC)*, IEEE, 2020, pp. 1–9.

[78] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, 2022.

[79] B. E. Santoso and G. P. Kusuma, "Facial emotion recognition on FER2013 using VGGSPINALNET," *J Theor Appl Inf Technol*, vol. 100, no. 7, pp. 2088–2102, 2022.

[80] G. Lu *et al.*, "Video-based neonatal pain expression recognition with cross-stream attention," *Multimed Tools Appl*, vol. 83, no. 2, pp. 4667–4690, 2024.

[81] C. Huang, "Face recognition algorithm based on improved neural network," in *Second Guangdong-Hong Kong-Macao Greater Bay Area Artificial Intelligence and Big Data Forum (AIBDF 2022)*, pp. 110–116, 2023.