FACIAL EMOTION RECOGNITION USING SHALLOW CONVOLUTIONAL NEURAL NETWORK AND IMPROVED FER-2013 DATASET

Qalandarov Aziz Abdukayumovich[1], Kurbanov Abdurahmon Alishboyevich[2]

[1]GulDPI Vice-Rector for Academic Affairs, Doctor of Philosophy, Associate Professor
[2]PhD student at Jizzakh branch of National University of Uzbekistan named after Mirzo Ulugbek
https://orcid.org/0000-0002-6840-3522
mr.kurbanov144@gmail.com

KEYWORDS

CNN, fer-2013, Machine learning, Deep learning, model, CNN architecture, OpenCV.

ABSTRACT

It is very easy and simple for a person to sense his inner feelings by looking at his face. That is, in the process of evaluating the emotional state of the person standing in front of him, the human brain sees the facial structure of the other person and can quickly analyze it. However, the ability of a computer to understand and respond to human emotions is considered one of the most difficult problems in the fields of modern computer vision and deep learning. Despite the fact that many studies have been carried out on the evaluation of the emotional state of a person, the proposed solutions are not effective enough. Several convolutional neural network models developed in this field can also solve the problem in a rather narrow scope. In this article, we proposed a shallow convolutional neural network and an augmented and improved fer-20103 dataset in order to speed up the training process and improve previous results. The proposed architecture was tested and analyzed on an updated dataset.

Introduction.

Detecting human emotions is of practical importance in several areas of society. In the medical field, identifying emotions helps in early detection and treatment of psychological problems. In the educational field, information about students ' emotions helps them to master the lesson being taught and to strengthen the individual approach (Kurbanov, 2024). The introduction of emotion recognition systems into online learning platforms will lead to high efficiency in distance learning education by using lesson content that is appropriate for the student 's personal needs. In security and surveillance systems, facial expression analysis is important for early detection of dangerous situations in public places. From the perspective of human - computer interaction, the ability of a computer to communicate with a person by analyzing the emotional state of a person will take the creation of humanoid robots to a new level.

The FER-2013 dataset is one of the most popular and widely used datasets for studying facial emotions. However, its shortcomings, such as labeling errors and poor image quality, can lead to a decrease in the accuracy of the results. By improving this dataset, for example, by correcting mislabeled images and removing poor quality data, the accuracy and reliability of the emotion recognition system can be significantly improved.

Researchers have taken various approaches to facial emotion recognition. Early studies used classical machine learning algorithms. These methods extracted features from facial images using, for example, HOG, LBP, or Gabor filters. Support vector machine (SVM) and k - nearest neighbors (k-NN) have been widely used in emotion classification (Kurbanov A. A., 2024). However, since classical approaches require manual feature extraction, their efficiency in detecting complex emotions has been limited (Figure 1).

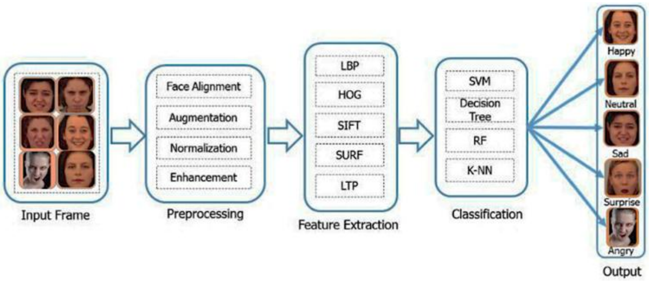


Figure 1. Emotion recognition based on traditional machine learning

Later, deep learning approaches have made significant progress in facial emotion recognition. Models built on convolutional neural networks (CNNs), including the LeNet, VGG, and ResNet architectures, have achieved high accuracy in emotion recognition (Figure 2). Transfer learning methods, such as VGGFace or EfficientNet, have been successfully applied to small datasets with pre - trained models.

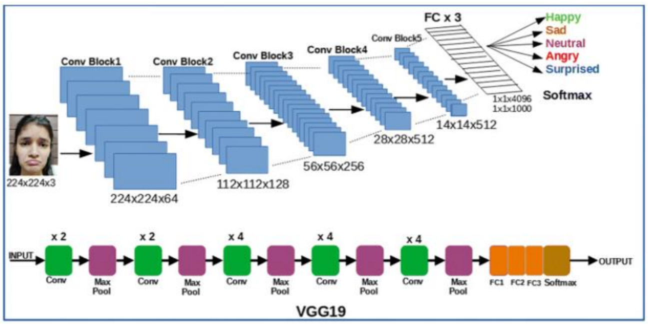


Figure 2. Emotion detection using the VGGNet 19 model

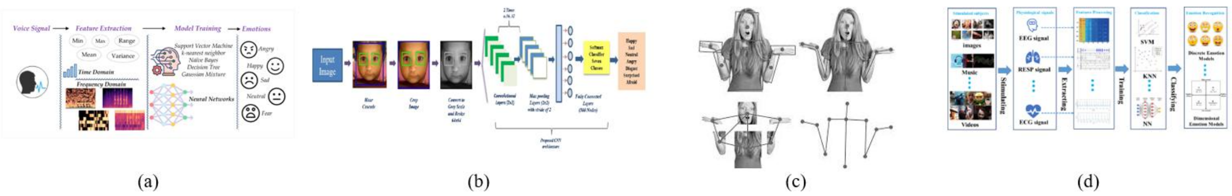
RNN and LSTM networks were also used to analyze the dynamics of facial expressions over time (Chouhayebi H, 2024). Currently, some scientists have also proposed hybrid methods for emotion recognition. In the hybrid method, facial image extraction is performed by performing the feature recognition process in CNN, and classifying the obtained features into emotions using traditional machine learning algorithms such as SVM or k- NN methods.

Shallow CNN architecture has several advantages over conventional deep learning models. These include low resource requirements, relatively fast training speed, and real - time adaptability. This approach performs well on small datasets and works well on devices with limited computing power, such as smartphones or IoT systems. In practice, this technology allows for the creation of automated analysis systems that do not require human intervention. Working with the improved FER - 2013 dataset makes the results more accurate and provides flexibility for more user cases. This makes this approach a relevant solution for various software products and research projects.

Materials and methods.

The classification of emotions is based on the approach proposed by psychologist Paul Ekman. He paid great attention to the study of the relationship between human facial expressions and emotions. In the 1970s, the scientist studied the relationship between facial expressions and emotions and developed models that provide recognition of basic emotional expressions. Paul Ekman's research not only studied the facial expressions of people from one culture, but also conducted research in distant tribes. He conducted empirical studies in other tribes, such as Papua New Guinea. These studies confirmed the cross-cultural properties of emotions proposed by Ekman. Ekman and his team concluded that the basic emotions, happiness, anger, fear, disgust, surprise and sadness, are expressed in the same way by all people, even if they belong to different cultures and tribes (Ekman, 1978).

It has been shown that identifying human emotions can be determined by analyzing several key sources or processes (Ahmed, 2023). Figure 3 shows the basic sources of emotion recognition, using facial images, speech sounds, text content, and body movement analysis (Kurbanov A. A, 2024) and emotions can be identified by analyzing data from physiological signals such as breathing rate and heart rate.



(a)

(b)

(c)

(d)

Figure 3. Emotion detection. (a) using speech sounds, (b) using facial images, (c) using body movement analysis, (d) using physiological signals

In this research paper, we propose an approach to solving the problem of assessing human emotional states by analyzing facial images. Several popular datasets that are widely used for detecting emotions from facial images have been developed by researchers and leading companies. **AffectNet** is one of the largest and most widely used facial expression datasets, created in 2017. This dataset contains more than 1 million facial expression images and is divided into 8 basic emotions (happiness, anger, fear, disgust, surprise, sadness, neutral, and disgust). The images were collected over the past 8 years, with 1.5 million analyzed. The dataset contains different emotions for each image, and as a result is very useful for recognizing facial expressions and emotions (A. Mollahosseini, 2019). **CK+ (Extended Cohn-Kanade)** – Developed by Cohn and Kanade in the 2000s. This dataset consists of 593 videos and 123 individuals. Each video represents a specific emotion, meaning that a visual representation of each emotion can be observed. The CK+ dataset provides high- quality images and is used by many researchers to study facial expressions and emotion recognition. This dataset contains 7 basic emotions: happiness, anger, fear, disgust, surprise, sadness, and neutral (P. Lucey, 2010). **EmoReact** – is used to detect facial expressions and emotions from video images. This dataset contains over 31,000 video clips, each clip showing a different emotion. Using video images, emotions can be observed and analyzed with high precision. This dataset is especially useful for studies that study emotion detection from dynamic facial expressions in video (Nojavanasghari, 2016). **JAFFE (Japanese Female Facial Expression)** – This dataset is focused on studying facial expressions, mainly filled with facial expressions of Japanese women. This dataset contains 213 images depicting 10 emotions (happiness, anger, fear, disgust, surprise, sadness, anxiety, fatigue, disgust, and neutral). The images are of high quality and make it easy to identify and analyze the image. This dataset may have small data sets, but it is effective for accurate emotion detection (Arafin, 2024).

FER-2013 (Facial Expression Recognition 2013) – The FER-2013 dataset is one of the most popular datasets for emotion recognition. It was created as part of the 2013 Facial Expression Recognition Challenge organized by Kaggle. The dataset contains 35,887 facial images, and these images are categorized into 7 basic emotions: happiness, anger, fear, disgust, surprise, sadness, and neutral. The images are 48x48 pixels in size, and each image represents only one emotion. The FER-2013 dataset has become the main source of information for many studies and model training (Minaee, 2021). Using this dataset, several researchers have conducted research on emotion recognition using various machine learning and deep learning models.

Zahara et al. attempted to detect micro expressions using the FER -2013 dataset. They explored the possibilities of analyzing emotions in real time using a CNN model, cameras, and a Raspberry Pi platform. (Zahara, 2020)

Oguine et al. studied the combination of CNN and Haar Cascade methods to achieve high accuracy in real- time emotion recognition, combining models and optimizing them for real - time systems. (Oguine, 2022)

Y. Khaireddin and Z. Chen tested the VGGNet model on the FER -2013 dataset and reported good results. (Khaireddin, 2021)

Tutuianu et al. compared different deep learning architectures and studied their performance in real - world facial expression recognition. They conducted experiments to evaluate the performance of different datasets and models under different conditions. (Tutuianu, 2024)

In the above analysis, it is clear that the main test set of many studies is FER -2013. No matter how accurate and loss- tolerant the architecture of the convolutional neural network is in the research process, the defects in the data set will reduce the efficiency of the result. The efficiency of the data set depends on several of its parameters.

Image clarity and precision. The dataset should be free of mislabeling, duplicates, or ambiguous data. Mislabeled data can lead to inaccurate training of the model.

Balance. The images in the dataset should be evenly distributed across all classes. If there is a large difference between classes (**some** classes have too many and others have too few), **the** model may learn biasedly.

Divergence (species diversity). The dataset should cover a variety of conditions, facial expressions, lighting, facial positions, skin tones, gender, and age. This helps the model better adapt to real - world conditions.

Sufficient size. The dataset must be large enough to be trained on. Smalldatasets are not suitable for deep learning models because they reduce the model 's generalization ability.

High image quality. Images should be clear, high- resolution, and free of artifacts (such as noise or distortion). High quality images help to accurately identify facial expressions.

Contextual information. If there are other objects or contexts in the image along with the face, they need to be taken into account. For example, elements other than the face can cause incorrect emotion detection.

Although the FER-2013 dataset is very popular, during the course of this research, we found that it did not sufficiently meet the above parameters, and we cleaned and updated FER - 2013. The image size in FER-2013 is 48x48, and the input image size for our proposed shallow CNN model is 224x24. We can usually scale 48x48 to 224x224, but the resulting image will not meet the **High Image Quality** parameter. Therefore, we use the ESPCN artificial intelligence model to scale the image size.

ESPCN (Efficient Sub- Pixel **Convolutional Neural Network)** is an efficient model for high-quality image upscaling, which implements superresolution (SR) using sub - pixel convolutions. This model was introduced by Chao Dong and colleagues in 2016. ESPCN has a simple structure and uses computational resources economically. (Ruan, 2022) The main idea of the ESPCN model is to upscale the image using traditional methods, such as bicubic interpolation, and then train the neural network on a low - resolution basis, and finally generate a high-resolution image through a sub-pixel layer. Figure 4 shows the structure of the ESPCN network. This increases the efficiency of the model.

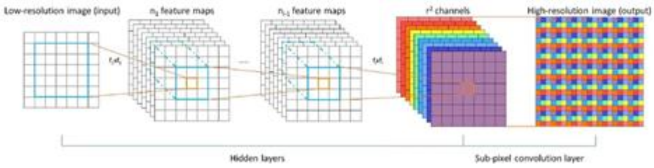

Low-resolution image (input)
n_1 feature maps
n_{L-1} feature maps
r^2 channels
High-resolution image (output)
Hidden layers
Sub-pixel convolution layer

Figure 4. ESPCN network structure

Loss function of ESPCN network

$$(W_{1:L}, b_{1:L}) = \frac{1}{r^2HW}\sum_{x=1}^{rH}\sum_{x=1}^{rW}(I_{x,y}^{HR} - f_{x,y}^{L}(I^{LR}))^2 \qquad (1)$$

Here, I^{HR} represents each original image in the dataset ; I^{LR} represents each subsampled LR image ; r represents the upscaling factor ; H represents the height value of the image ; W represents the width value of the image, $W_{1:L}$ represents all the network weights to be learned, and $b_{1:L}$ represents all the possible learning values (Arafin, 2024). In Figure 5, we can see the image size being enlarged in the usual way and the enlargement using the ESPCN model.

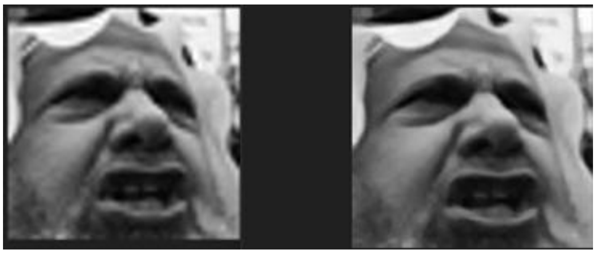


Figure 5. Image size increase in the usual case and using the ESPCN model

The FER-2013 dataset contains 35,887 facial images categorized into 7 main emotions, but a closer look at these facial images reveals that there are many duplicates. Using the Python programming language hashlib library, we hash image files to identify identical files. All the code in this article is available on the github repository https://github.com/abdurakhmonkurbanov/article.git. Figure 6 shows a diagram of the facial images in FER- 2013.

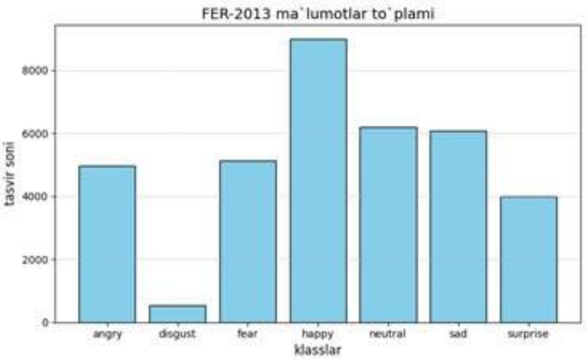


Figure 6. Segmentation of images from the FER-2013 dataset

It is clear that the **Balance in the dataset** is not correct. To correct this deficiency, we add additional images from another dataset to the disgust class. In FER-2013, we delete images that are of poor quality, are not facial images, but rather other images, and are incorrectly labeled. To do this, we examine all the images in the dataset one by one. Figure 7 shows examples of the deficiencies identified.

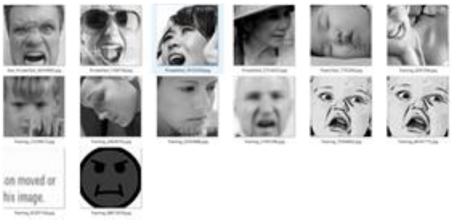


Figure 7. Some of the shortcomings of FER-2013 include images that are too blurry, images that have been merged into another class, and images that do not have facial images.

After eliminating the shortcomings in all parameters, we present the updated FER-2013 dataset, and in the new dataset, a total of 35,896 facial images are grouped into 7 emotion types : angry, disgust, fear, happy, neutral, sad, and surprised. The number of classes is also relatively equal (Figure 7).

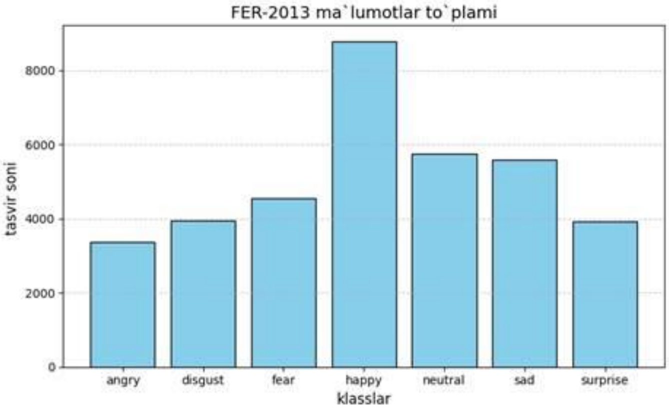


Figure 8. Segmentation of images from the updated FER-2013 data set

In the next step, we will build a shallow convolutional neural network. The model we propose accepts a 224x224 image with 1 channel. It consists of three convolutional layers, and after each layer, a ReLU activation function and MaxPooling layers are used. The convolutional layers help extract the features of the image. The model has 2 fully connected layers. After the first fully connected layer, ReLU activation and dropout are used, which increases the generalization ability of the model and reduces overfitting. The final layer will have an output corresponding to 7 emotion classes. The general detailed form of the model is given in Table 1.

Table 1:

architecture of shallow convolutional neural network

Layer	Output Shape	Param #	Description
Conv2d (conv1)	(None, 224, 224, 32)	320	Convolution of a 1-channel input image (grayscale) with 32 filters.
ReLU (activation)	(None, 224, 224, 32)	0	Applying ReLU activation to the output from the convolutional layer.
MaxPooling2d	(None, 112, 112, 32)	0	Reduce the image to 2x2 size using max pooling.
Conv2d (conv2)	(None, 112, 112, 64)	18,496	Add the next convolution layer with 32 filters.
ReLU (activation_1)	(None, 112, 112, 64)	0	Applying ReLU activation to the output from the convolutional layer.

MaxPooling2d	(None, 56, 56, 64)	0	Apply a Max pooling layer and resize the image again.
Conv2d (conv3)	(None, 56, 56, 128)	73,856	Add another convolution layer with 64 filters.
ReLU (activation_2)	(None, 56, 56, 128)	0	Applying ReLU activation to the output from the convolutional layer.
MaxPooling2d	(None, 28, 28, 128)	0	Reduce the size of the image using max pooling.
Dropout	(None, 28, 28, 128)	0	Using a dropout layer to prevent overfitting.
Flatten	(None, 100352)	0	Convert an image to a one - dimensional vector.
Dense (fc1)	(None, 512)	51,795,456	Convert a vector of size 100352 to size 512.
ReLU (activation_3)	(None, 512)	0	Applying ReLU activation to the output from the fully connected layer.
Dropout	(None, 512)	0	Randomly " turning off " neurons using dropout.
Dense (fc2)	(None, 7)	3,591	In the final layer of the model, 512 neurons are divided into 7 emotion classes.
Softmax	(None, 7)	0	Calculating final probabilities for emotion classification with softmax activation.

Cross-Entropy Loss was used as the loss function in building the model. Cross-Entropy Loss is defined by the formula (2).

$$L = -\sum_{i=1}^{C} y_i * log(p_i) \quad (2)$$

In this

L – loss value,

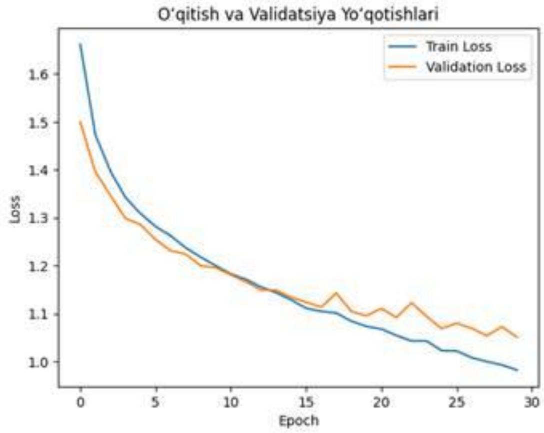
y_i – real characters, i.e., true or false

p_i is the probability predicted by the model.

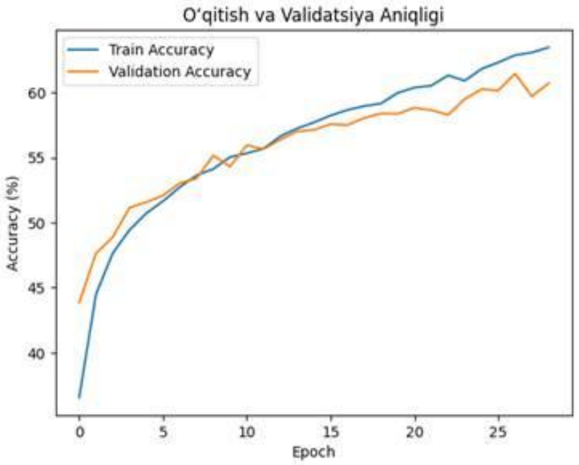
The loss function measures the difference between the model's likely predictions and the actual labels and is used in optimization to further improve the model. (Mao, 2023)

Research results.

The model training process continued for 30 epochs. The results improved with each epoch. In the last epoch, we obtained the results **Epoch 30/30, Train Loss: 0.9821, Train Accuracy: 63.47%, Val Loss: 1.0508, Val Accuracy: 60.71%.** We saved the model for further use. Figure 8 shows a diagram of our results for each epoch.



(a)



(b)

Figure 8. Results obtained with the improved FER-2013 dataset of a shallow convolutional neural network : (a) losses, (b) accuracies.

Analysis of research results.

In our study, the results of the training process of our model were observed in terms of training and validation losses and accuracy. The results show that the model improves during training and also improves significantly on the validation data. Training Loss**: It can** be seen that the model’s loss decreases significantly during the training process. From the first epoch (1.6618) **to** the 30th epoch (0.9821), the **training** loss decreases by about 60 %. This shows that the model is capable of learning and performs robustly during the training process. **Validation Loss**: The validation loss also follows a similar trend. We can see a decrease from 1.4997 in the first epoch to 1.0508 in the last epoch. This indicates that the model also adapts to the validation data during the overall optimization process. **Accuracy:** The training accuracy of the model was 36.55% at the beginning, and reached 63.47 % by the last epoch (epoch 30). This shows that the model has significantly improved in detecting emotional states during the training process. **Validation** Accuracy: **The** validation accuracy also increased in the same way. The increase from 43.85 % in the first epoch to 60.71% in the 30th epoch indicates that the model is successfully learning on the validation data and has good generalization ability to new data. The similarity between the training and validation losses and accuracy of the model indicates **that** the model **avoids overfitting**, as the training and validation losses show the same decrease. Early stages of training (Epochs 1-10): The training and validation losses and accuracies changed rapidly in the first 10 epochs. This indicates a rapid increase in the initial learning process. Middle stages (Epochs 10-20): In the middle stages, the learning rate of the model slowed down a bit, but the loss decreased and the accuracy continued to increase. This indicates that the model has started to learn steadily. Late stages (Epochs 20-30): In the last 10 epochs, i.e., the last training stages of the model, the loss and accuracy showed minimal changes. This indicates that the model has finished improving and has reached the end of the optimization process.

Conclusion.

This study presents approaches to improve the performance of facial emotion recognition using a shallow convolutional neural network architecture and the improved FER - 2013 dataset. The following main results and conclusions were reached during the study :

1. **Shallow Convolutional Neural Network (Shallow CNN):** The Shallow CNN model has a simple but effective architecture and was trained for 30 epochs. Each layer of the model, in particular, convolutional layers, ReLU activation function, max pooling and dropout layers, ensured accurate data analysis. Analysis of the changes in the accuracy and loss of the model during training showed the successful learning process and the effectiveness of the model in identifying emotional states. The decrease in the training loss from 1.66 to 0.98, and the increase in accuracy from 36.55% to 63.47% confirmed the improvement of the model.
2. **Improved FER-2013 dataset:** The FER-2013 dataset has been supplemented and improved to address issues in its original version. Face image size differences, mislabeling, and low-**quality** images have been corrected. These improvements have improved the model's generalization ability and allowed it to achieve superior results on both validation and testing data.
3. **Final Results:** The proposed shallow CNN model and the improved FER-2013 dataset have achieved significant accuracy in facial emotion recognition. The decrease and increase in the accuracy and loss of the model confirm its learning efficiency. The new approach has shown to have higher accuracy and better generalization ability than previous studies. This

method makes a great contribution to the development of new approaches in facial emotion recognition and can be used in many computer vision and deep learning applications.

However, there are possibilities to further improve the model by architectural changes, using more complex networks, and using other optimization techniques. We will continue in this direction in our future research.

1. A. Mollahosseini, B. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing, vol. 10*, 18-31.
2. Ahmed, N. a. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. Appl. 17, 200171.*
3. Arafin, S. a. (2024). Enhancing the ability to recognize facial expressions in young children: A method using few-shot learning and cross-dataset validation. *6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, (pp. 640-645).
4. Chouhayebi H, M. M. (2024). Human Emotion Recognition Based on Spatio-Temporal Facial Features Using HOG-HOF and VGG-LSTM. *Computers*, 101.
5. Ekman, P. a. (1978). Facial action coding systems. *Consulting Psychologists Press*.
6. Khaireddin, Y. a. (2021). Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588.*
7. Kurbanov A. A. (2024). Inson tanasi harakatlarini tahlil qilishda zamonaviy modellar va algoritmlarni qo ‘llashni o ‘rganish. *Al-Farg’oniy avlodlari*, 169-175.
8. Kurbanov A. A. (2024). Inson yuz tasviridan hissiyotlarni aniqlash uchun geomertik xususiyatlarini va tashqi ko ‘rinishga asoslangan xususiyatlarini ajratib olish. *Al-Farg’oniy avlodlari*, 61-67.

urbanov, A. (2024). Chuqur o‘rganishga asoslangan yuz tahlili: xususiyatlarni ajratib olish va his-tuyg‘ularni tushunish. *Международный Журнал Теоретических и Прикладных Вопросов Цифровых Т*

10. Lucey, P. e. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *IEEE.*
11. Mao, A. M. (2023). Cross-entropy loss functions: Theoretical analysis and applications. *International conference on Machine learning*, (pp. 23803-23828).
12. Minaee, S. a. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 30-46.
13. Nojavanasghari, B. B. (2016). Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. *Proceedings of the 18th acm international conference on multimodal interaction*, 137-144.
14. Oguine, O. C. (2022). Hybrid facial expression recognition (FER2013) model for real-time emotion classification and prediction. *arXiv preprint arXiv:2206.09509.*
15. P. Lucey, J. F. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA*, 94-101.
16. Ruan, H. T. (2022). Efficient sub-pixel convolutional neural network for terahertz image super-resolution. *Optics letters*, 3115-3118.
17. Tutuianu, G. I. (2024). Benchmarking deep facial expression recognition: An extensive protocol with balanced dataset in the wild. *Engineering Applications of Artificial Intelligence*.
18. Zahara, L. M. (2020). The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi. *Fifth international conference on informatics and computing (ICIC)* (pp. 1-9). IEEE.