Convolutional Neural Network-Based Classification of Facial Emotional Expressions and Computational Complexity Analysis

Berkay Cakmak* and Ibrahim Develi

Faculty of Engineering, Electrical & Electronics Engineering, Erciyes University, 38039, Kayseri, Türkiye

*brkayckmak@gmail.com Email of the corresponding author

Abstract **–** This study presents a novel convolutional neural network (CNN) model and a detailed complexity analysis with other models available in the literature for an accurate classification of facial emotional expressions. Human beings have been defined by seven basic emotions, which are anger, fear, happiness, sad, contempt, disgust, and surprise. Model accuracy plays an important role in emotion detection studies with deep neural networks, as the high model accuracy is directly related to the accuracy of the predicted emotions. A 23-layer CNN model was created that classifies 7 different emotions. The CNN model we trained with the FER2013 dataset has a higher accuracy performance than other studies trained with the same dataset in the literature. The accuracy performance of our CNN model is 98.83% in training data and 83.52% in validation data. The complexity of the algorithms used in other studies is compared with the proposed study. Although the accuracy performance of our CNN model is higher than other studies in the literature, the complexity of our model is also higher than most other studies. The CNN model we obtained is used in an algorithm that we have created to increase the efficiency of online courses, which performs sentiment analysis 4 times per second.

Keywords – Convolutional Neural Networks, Identifying Facial Expressions of Emotions, FER2013 Data Set, Complexity Analysis, Online Education Efficiency.

I. INTRODUCTION

Sentiment analysis is a large and important research area that aims to identify human emotions with image processing methods. Facial changes during a communication are the first signs that convey the emotional state [1]. Among the many non-verbal components, facial expressions are one of the main channels of information in communication because of their emotional meaning. Since the twentieth century, human beings have been defined by seven basic emotions, which are anger, fear, happiness, sad, contempt, disgust, and surprise, regardless of the culture they grew up in [2]. Sentiment analysis approaches from facial expressions consist of three steps which are face and face component detection, feature extraction, and expression classification. First, a face is detected from an input image and starts with detecting face components or landmarks from the face region. Then, various features are extracted from the facial components. Finally, recognition results are generated using features extracted by deep learning algorithms. [3].

CNN is a class of deep neural networks widely applied for analyzing visual images. CNN is a feedforward neural network, also called multilayer perceptrons (MLPs). CNN's main field of work is image recognition and image classification. The convolution layer extracts the features of an input

image from the used data set. A CNN model consists of convolution, concatenation, flattening, and dense layers. When training a CNN model with large datasets, more convolutional layers are used for better accuracy [4].

In general, the complexity of a neural network structure is measured by the number of free parameters in the network; that is, the number of neurons and the number and strength (weights) of connections between neurons. Network complexity analysis plays an important role in the design and implementation of artificial neural networks because complexity can significantly affect the neural network's ability to learn and generalize [5]. Prediction process of the networks used in real-time applications requires very low latency, so the heavy computational burden is a major problem with these systems. The success of convolutional neural networks in real-time applications is limited by heavy computational loads [6]. The recent trend in "deep learning" to use more complex multi-parameter models requires large amounts of computational resources, in particular for training the models but also for applying the learned ConvNets [7].

In this study, a CNN model that can detect people's emotions through facial expressions was designed using image processing technology. The designed CNN model works more efficiently than other algorithms trained with the existing FER2013 data set in the literature. The emotional states obtained in real-time in live lectures or trainings conducted over the Internet and the intense emotional states of the students according to time are analyzed. As a result of the analysis, a study was carried out to increase the efficiency of the lecture by notifying the lecturer of the change in the instantaneous emotional intensity and giving a report at the end of the lecture.

II. Literature Review

Facial expressions are the common signal for all people to express emotions. Since there are application areas in many fields such as robotics, medicine, and driving support systems, many attempts are made to make facial expression analysis tools with image processing [8].

In the literature, some studies perform sentiment analysis using the FER2013 data set. In [9], the performance of the Xception algorithm trained with the FER2013 data set was obtained as 61.7%, and then the accuracy was increased to 63% with the facial image threshing process. In the study conducted in [10], a CNN model was designed for sentiment analysis, and an accuracy of 57.1% was achieved. In another sentiment analysis study, 65% accuracy was achieved by training the mini Xception architecture with the FER2013 data set. In that study, they also demonstrated the accuracy of the real-time CNN model [11]. The 11-layer CNN model prepared in [12] was trained for 106 epochs with the FER2013 data set and achieved 70% accuracy. In that study, they also conducted to find out how similar facial expressions are to which expression in percentage terms. In [13], several CNN models, pre-trained models, and training procedures were studied, and the studies were compared. In [13], the k-nearest neighbour (KNN) model was applied to select the closest training examples for an input image, then the support vector machine (SVM) classifier was then trained on selected training samples and used to predict the class label for the test image on which it was trained. With this method, 75.42% accuracy was achieved on the FER2013 data set. In another study, an 11-layer CNN model was designed and the Scale Invariant Feature Transform (SIFT) method was applied to this model. The SIFT method aims to achieve high accuracy with a low-volume data set, but the accuracy rate was determined as 73.4% [14]. In another study, the analysis of the emotions of the students in the lecture with image processing algorithms was analyzed by taking a one-time image at the end of the training [15]. Our study differs from [15] in that it analyzes the emotions of the students 4 times every second, presents the lecturer with instantaneous emotional change notifications and has a high accuracy rate as stated.

III. Method

As input data, an algorithm is designed that takes the facial images of the listeners 4 times per second and creates a time-dependent graph of the emotions. The algorithm consists of 4 main parts.

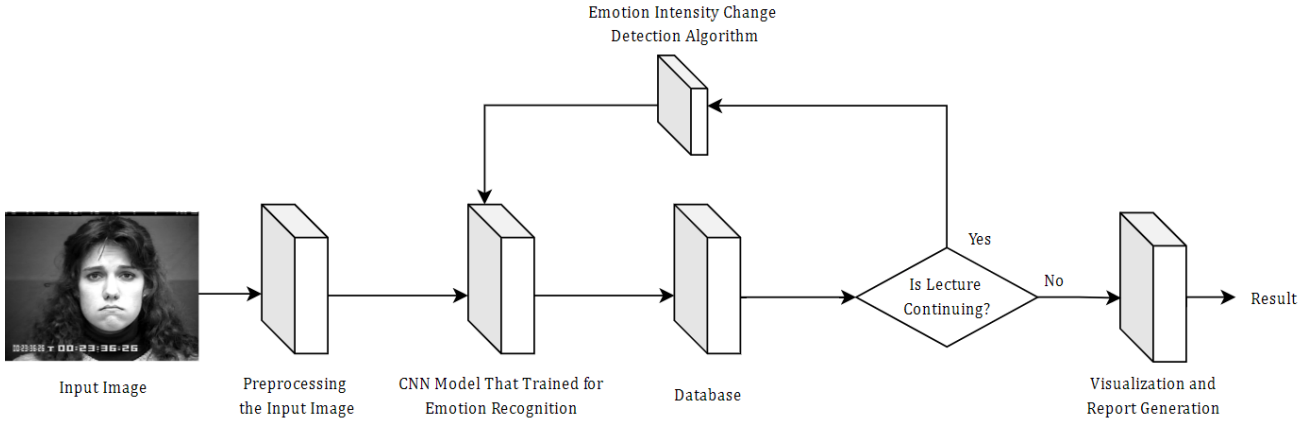

Fig 1. The flowchart of the algorithm.

These sections are processing the input image, emotion recognition, transferring the emotion data to the database, and finally visualizing the data and creating a report. Figure 1 shows the flowchart of the algorithm.

In this study, the input image goes through a pre-processing stage consisting of 3 steps. In the first step, the input image, which has 3 dimensions, red, green, and blue, is reduced to one dimension by going through the grayscale process. After this process, in the second step, the region of the face is determined using the Haar Cascade library to eliminate the possibility that pixels other than the face region will affect the prediction success of the developed CNN model. In the third step, the face area is cropped from the input image and converted to 48x48 size. In Figure 2, the image pre-processing stages are visualized and presented. Emotion detection is performed by processing the pre-processed image with the trained CNN model.

Our CNN model has 23 layers. The first layer consists of a 48x48 input image. This layer is followed by a block of convolution, normalization, and activation (ReLU) layers consisting of 32 feature maps. After this block, the model structure continues with another block consisting of 64 feature maps and containing normalization and activation layers. In the continuation of this block, there is a 2x2 max pooling layer. Therefore, the size of the resulting image is 23x23x64. Similarly, following this layer, a block consists of convolution, activation (ReLU), and normalization layers consists of 64 feature maps and a block consisting of convolution, activation (ReLU), and normalization layers consists of 128 feature maps, followed by another 2x2 size pooling (max. pooling) layer. The size of the image that will appear after this layer is 10x10x128. After this layer, the data is made one-dimensional with a smoothing layer. Afterward, a model was created with a 200-dimensional dense, then a dropout, and finally a 7-dimensional dense layer. The created CNN model can predict 7 different emotions as anger, confused, fearful, happy, neutral, sad, and disgusted.

Figure 3 represents the architecture of the CNN model. 80% of the FER2013 data set was used for training and 20% for testing of the CNN model. The created CNN model was trained for 60 epochs with the FER2013 data set, which was separated for training and testing, and reached an accuracy of 97.83% in the training data set and 83.52% in the test data set.

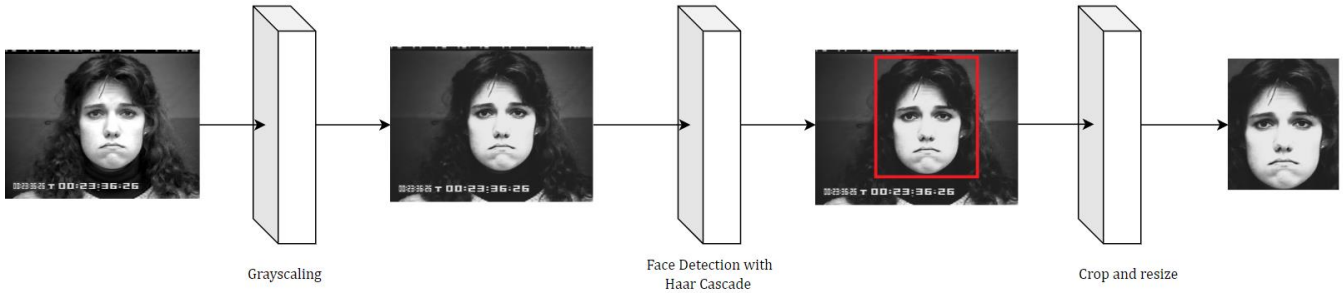

Fig 2. Pre-processing stages of the input image.

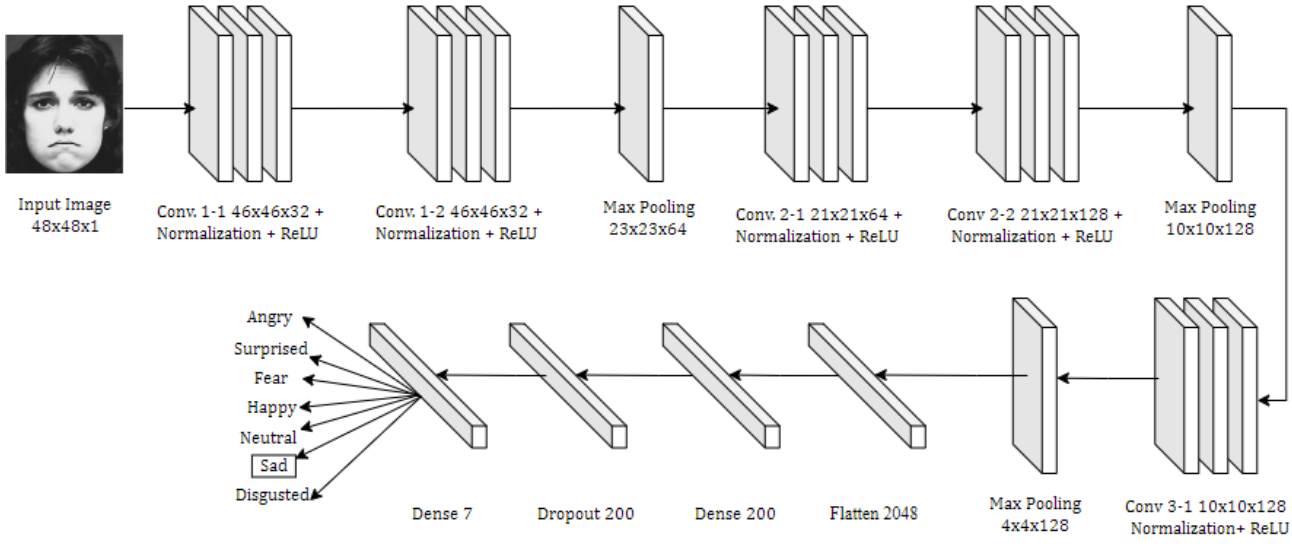

Fig 3. The architecture of the created CNN model.

Table 1 shows the confusion matrix in the training data set of the CNN model. When the confusion matrix is examined, it is seen that most of the images given as input to the trained CNN model are predicted correctly.

Table 1. Confusion matrix in the training data of the CNN model.

Confusion Matrix								
True Classes	Angry	7919 (13.98%)	4 (0.01%)	47 (0.08%)	23 (0.04 %)	31 (0.05%)	10 (0.02%)	20 (0.04%)
	Disgusted	0 (0%)	8094 (14.29%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Fear	28 (0.05%)	2 (0%)	7900 (13.95 %)	10 (0.02 %)	57 (0.1%)	68 (0.12%)	44 (0.08%)
	Happy	3 (0.01%)	0 (0%)	1 (0%)	8056 (14.23 %)	4 (0.01%)	6 (0.01%)	13 (0.02%)
	Sad	35 (0.06%)	0 (0%)	39 (0.07%)	21 (0.04 %)	7954 (14.05%)	4 (0.01%)	48 (0.08%)
	Surprised	10 (0.02%)	5 (0.01%)	24 (0.04%)	13 (0.02 %)	7 (0.01%)	8059 (14.23%)	2 (0%)
	Neutral	23 (0.04%)	0 (0%)	28 (0.05%)	31 (0.05 %)	37 (0.07%)	6 (0.01%)	7944 (14.04%)
		Angry	Disgusted	Fear	Happy	Sad	Surprised	Neutral
		Predicted Classes						

The established algorithm analyses the faces of the listeners and makes emotional predictions 4 times per second. The data from the CNN model is stored for visualization and reporting by transferring 4 emotional data per second to the database during the training period. The instantaneous data is processed by another algorithm simultaneously with the database connection, and certain changes in the intensity of emotion are presented to the lecturer in the form of notification while the lecture continues.

IV. Complexity Analysis

In the study conducted in [9], a pre-trained CNN model consisting of greater than 26 layers, the Xception model, and a 13-layer facial image threshing (FIT) machine are used. In the study conducted in [10], a CNN model with 1 input, 4 blocks; each consisting of a convolution, ReLU and max pooling layer, and 3 fully connected layers at the end of the network was prepared. This CNN model consists of 17 layers in total and no other algorithm is used to increase the performance of the model. In [11], a pre-trained CNN model named the mini Xception model was used. This model consists of greater than 10 layers and is implemented in a camera via Raspberry Pi. With this way, emotion measurement can be made live. In [12], a total of 12-layer CNN model was created, consisting of 1 input layer, 4 blocks; each consisting of a convolution and a max pooling layer, and 2 fully connected layers behind these blocks. In [13]; four different models, bag-of-visual-words (BOVW), VGG-face, VGG-F and VGG-13 were trained. The BOVW model, which is a KNN model, built the feature representation by extracting dense SIFT descriptors from all training images, and by later quantizing the extracted descriptors into visual words using k-means clustering [13]. The VGG-face model is a pre-trained 16-layer CNN model. The VGG-F model is a pre-trained 8-layer CNN model. The VGG-13 model, on the other hand, is not pre-trained and is a CNN model consisting of 13 convolutions and 4 max poolings, a total of 17 layers. A CNN model

powered by the SIFT method was designed in [14]. SIFT is used to extract the key-points from the facial images. CNN model has 3 blocks, each consisting of 2 convolutions, 1 max pooling and 1 dropout layer. After these blocks, the output meets the output of the SIFT model in a fully connected layer. This model has a total of 17 layers. The complexities of the mentioned studies are summarized in Table 2. In this study, the model complexity was calculated by counting each of the layers such as convolution, ReLU, max pooling, fully connected layer and dropout layers in CNN models as 1 layer unit.

Table 2. Studies in the literature and accuracy performances of the proposed approach.

Studies in the Literature	**Algorithm**	**Complexity**	**Accuracy Performance**
[9]	Xception (CNN)	Greater than 27-layer CNN + FIT with 13 layers	63,00%
[10]	CNN	17-layer CNN	57,10%
[11]	mini Xception (CNN)	Greater than 10-layer CNN	65,00%
[12]	CNN	12-layer CNN	70,00%
[13]	CNN, KNN + SVM	KNN+SVM, 16-layer CNN, 8-layer CNN, 17-layer CNN	75,42%
[14]	CNN + SIFT	17-layer CNN+SIFT	73,40%
Proposed	CNN	23-layer CNN	97,83%

As can be seen, the proposed model is more complex than the other studies except the study given by [7]. Although we have presented a highly complex model, the accuracy of the CNN model we trained is much higher than the models established in other studies.

V. Results

The accuracy of the CNN model we obtained is higher than other studies in the literature. On the other hand, the complexity of our CNN model has been revealed to be higher than the CNN models in most studies in the literature. With the established algorithm, an increase in efficiency has been achieved in online lectures.

VI. Discussion

The study differs from other studies in the literature with the high accuracy of the CNN model and its time-based data processing. When other studies in the literature are examined, no study can reach 97.83% training success and 83.52% test success between the deep neural networks that trained with FER2013 dataset. The complexity of our CNN model is higher than most other studies found in the literature. As the complexity of the CNN model increases, the computational load also increases. Despite this, there is no problem in the operation of the emotion recognition algorithm, which we perform live 4 times per second. In CNN models, complexity is often sacrificed for high accuracy. There are many studies aiming to increase the accuracy rate by reducing the complexity in CNN models. In other studies, in the literature on the use of deep neural networks and emotion analysis to increase educational efficiency, there are analysis problems arising from not processing students' emotions over time. Emotion analysis only from the image taken at the end of the lesson does not provide a large amount of useful data according to our study, since it does not keep the data about the emotions of the students during the lesson. Since 4 emotions are detected per second for each student with the algorithm prepared in our study, both the efficiency of the lecturer is increased before the end of the lesson, and the efficiency of the instructor's future lessons can be increased based on the detailed report shared at the end of the lecture.

VII. Conclusion

In this study, a CNN model with high classification accuracy was prepared to detect seven different emotions in real-time. The prepared CNN model was integrated into the algorithm designed for the analysis of the emotional states of the participants in online education platforms. With the designed system, the lecturer can monitor the emotional states of the participants during the lecture and can take actions that will positively affect the efficiency, quality, and performance of their training with the useful information and graphics in the report sent at the end of the lecture. This report increases efficiency in future lectures.

REFERENCES

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

[2] P. Ekman and W. V. Friesen, *Constants across cultures in the face and emotion.* Journal of personality and social psychology, 1971, 17.2: 124.

[3] B. C. Ko, *A brief review of facial emotion recognition based on visual information*. sensors, 2018, 18.2: 401.

[4] F. Altekin and H. Demir, *Emotion Detection from Facial Expression Using Different Feature Descriptor Methods with Convolutional Neural Networks*. European Journal of Engineering and Applied Sciences, 4.1: 14-17.

[5] H. Yu, *Network complexity analysis of multilayer feedforward artificial neural networks.* Applications of Neural Networks in High Assurance Systems, 2010, 41-55.

[6] Y. Zhao, et al., *A faster algorithm for reducing the computational complexity of convolutional neural networks.* Algorithms, 2018, 11.10: 159.

[7] R. J. Cintra, et al. *Low-complexity approximate convolutional neural networks.* IEEE transactions on neural networks and learning systems, 2018, 29.12: 5981-5992.

[8] N. Mehendale, *Facial emotion recognition using convolutional neural networks (FERC).* SN Applied Sciences, 2020, 2.3: 446.

[9] J. H. Kim, Alwin Poulose and D. S. Han., *The extensive usage of the facial image threshing machine for facial emotion recognition performance.* Sensors, 2021, 21.6: 2026.

[10] V. Tumen, O. F. Soylemez and B. Ergen, *Facial emotion recognition on a dataset using convolutional neural network.* In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, 2017. p. 1-5.

[11] L. Zahara, et al. *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. p. 1-9.

[12] I. Lasri, A. R. Solh and M. El Belkacemi, *Facial emotion recognition of students using convolutional neural network.* In: 2019 third international conference on intelligent computing in data sciences (ICDS). IEEE, 2019. p. 1-6.

[13] M. I. Georgescu., R. T. Ionescu and M. Popescu, *Local learning with deep and handcrafted features for facial expression recognition.* IEEE Access, 2019, 7: 64827-64836.

[14] T. Connie, et al. *Facial expression recognition using a hybrid CNN–SIFT aggregator.* In: Multi-disciplinary Trends in Artificial Intelligence: 11th International Workshop, MIWAI 2017, Gadong, Brunei, November 20-22, 2017, Proceedings 11. Springer International Publishing, 2017. p. 139-149.

[15] W. Wang, et al. *Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation.* Complexity, 2020, 2020: 1-9.