REAL-TIME FACIAL EMOTION RECOGNITION FOR VISUALIZATION SYSTEMS

CEREN ÖZKARA

Master's Thesis

Graduate School
Izmir University of Economics
Izmir
2022

REAL-TIME FACIAL EMOTION RECOGNITION FOR VISUALIZATION SYSTEMS

CEREN ÖZKARA

A Thesis Submitted to
The Graduate School of Izmir University of Economics
Master's Program in Electrical and Electronics Engineering

Izmir
2022

ABSTRACT

REAL-TIME FACIAL EMOTION RECOGNITION FOR VISUALIZATION SYSTEMS

Özkara, Ceren

Master's Program in Electrical and Electronics Engineering

Advisor: Assoc. Prof. Dr. Pınar OĞUZ EKİM

October, 2022

The camera systems must be able to support meaningful results for users due to the advancement in technology. The camera systems are usually used in areas with humans or robots, so facial emotion recognition results chosen as supporting information are the favourite choice in the camera systems. Thus, this thesis aims to review the most popular deep learning algorithms and their performances in camera systems based on real-time facial emotion recognition and suggest a new model for future applications. Firstly, convolutional neural network (CNN) algorithms that recognize human emotions, such as AlexNet, GoogleNet, and VGG19, are investigated according to their performances. Then, the CNN algorithm with the best numerical performance is chosen for enhancement. After, the new hybrid model is constructed via chosen CNN and long short-term memory (LSTM). Lastly, the proposed model and face images achieved from the camera are combined to simulate real-time application.

Keywords: face detection, facial expression recognition, CNN, LSTM, hybrid model.

ÖZET

GÖRSELLEŞTİRME SİSTEMLERİ İÇİN GERÇEK ZAMAN YÜZ DUYGU TANIMA

Özkara, Ceren

Elektrik ve Elektronik Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Pınar Oğuz Ekim

Ekim, 2022

Kamera sistemleri, teknolojideki ilerleme nedeniyle kullanıcılar için anlamlı sonuçlar verebilir hale gelmelidir. Çoğunlukla insanların veya robotların olduğu alanlarda kullanılan kamera sistemleri için destekleyici bilgi olarak yüz duygu tanıma sonuçlarının verilmesi, kamera sistemlerinde favori bir yaklaşımdır. Bu nedenle, bu tez gerçek zamanlı yüz duygu tanımaya dayalı kamera sistemlerinde en popüler derin öğrenme algoritmalarını ve performanslarını gözden geçirmeyi ve gelecekteki uygulamalar için yeni bir model önermeyi amaçlamaktadır. İlk olarak AlexNet, GoogleNet ve VGG19 gibi insan duygularını tanıyan evrişimli sinir ağı (CNN) algoritmaları performanslarına göre incelenmiştir. Ardından, geliştirme için en iyi sayısal performansa sahip CNN algoritması seçilmiştir. Seçilen CNN ve uzun kısa süreli bellek (LSTM) aracılığıyla yeni hibrit model oluşturulmuştur. Son olarak, önerilen model ve gerçek zamanlı olarak kameradan elde edilen yüz görüntüleri ile uygulama gerçekleştirilmiştir.

Anahtar Kelimeler: yüz bulma, duygu hatırlama, CNN, LSTM, hibrit model.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Assoc. Prof. Dr. Pınar Oğuz Ekim of Electrical and Electronics Engineering at the Izmir University of Economics. Without her assistance, support, and understanding over every step of the process, this paper would have never been accomplished.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN : Convolutional Neural Network

LSTM : Long Short-Term Memory

RGB : Red Green Blue

CHAPTER 1: INTRODUCTION

With the development of technology, the interest in recognizing the living thing called humans has increased powerfully. How he thinks, how he decides, how he behaves, and how the so-called feeling affects these behaviors have always been a subject of research, and many models, studies, and simulations have been made on this subject. Because when people are mentioned, many unknown and undiscovered parts come to mind. When it is desired to look in more detail, the human factor is governed and acted by hormones and emotions. In other words, emotion analysis constitutes an essential point in order to be able to define people, analyze them and predict their behaviors. At this point, technology has inevitably touched on this subject a lot. Firstly, it has been determined how human emotions are revealed, and these features are examined. These features can be listed as follows; speech, facial expressions, body language, and gaze. In this way, we can analyze a person's emotions from the tone of voice, volume, and flow of speech, as well as from eye movements, mimics, hand movements, and even the posture of his body. Thus, advanced methods such as sound and image processing are a blessing in the analysis process. The part that is wanted to be examined in this study is the visualization part of this analysis process. Thus, the importance of mimics in the analysis of human emotions will be mentioned, and the contribution of the proposed method will be discussed, taking into account the studies carried out at this point. When the literature research is done, the studies carried out in recent years will be examined in terms of the main idea and method.

1.1 Literature Research

People commonly use facial expressions to show their emotional states and make communication. Researchers that know the importance of this nonverbal information for clear communication have started to develop emotion recognition systems based on facial expressions. These systems developed with deep learning algorithms and feature extraction methods have many usage and application areas, especially security, health, and human-machine interfaces (Lu, 2022).

Recently, investigations have focused on reconstructing deep learning algorithms or combinations (Jain, Shamsolmoali, and Sehdev, 2019; Islam, Islam, and Asraf, 2020). Some researchers have proposed a model based on single deep convolutional neural networks for facial emotion recognition. The proposed model consists of 6 convolutional layers, three max-pooling layers after the convolutions layer, two deep residual learning boxes implemented after the second and fourth convolution layer, and two fully connected layers. The model performance has been tested on two public datasets: Extended Cohn–Kanade (CK+) and Japanese Female Facial Expression (JAFFE) (Jain, Shamsolmoali, and Sehdev, 2019). The searchers interested in the health sector focus on the importance of automatic health issue detection because of the increasing quick result requirement in the vast population. Because of this reason, they suggest the CNN-LSTM hybrid model instead of the CNN approach. Obtained better validation accuracy and fast training result in the hybrid model shows that researchers will focus on hybrid models in the future (Islam, Islam and Asraf, 2020). Researchers have analyzed four CNN architectures (GoogleNet, ResNet, VGGNet, and AlexNet) for facial expression recognition according to their validation accuracy on FER2013, which is one of the famous datasets. The original method for facial expression recognition is reconstructed to obtain a basic structure, and this new model is shown to improve the accuracy result (Gan, 2018).

Moreover, investigated literature shows that there are so many datasets that are related to facial emotion detection. In the human-robot interactive processes, the main point is communication; therefore, Jaiswal and Nandi focused on the applications that allow robots to understand human facial expressions. They offered a CNN-based model and tested it over eight various datasets: FER2013, CK, CK+, Chicago Face Database, JAFFE Dataset, FEI face dataset, IMFDB, and TFEID (Jaiswal and Nandi, 2019). Understanding age, gender, and emotion are vital for camera systems used in the entrances to places with alcohol prohibition or age restriction and in forensic controls. The real-time application was made by combining the results of 2 different learning algorithms with face capture for this purpose. In the research, emotions were learned by using FER2013; age and gender were learned by using the hdf5 file, which is an open-source file that comes in handy to store large amounts of data. The learning algorithms ResNet and inception v3 were used (Manasa et al., 2020). A report examining performances on Fer2013 examined how using different numbers and sizes of kernel filters would affect the results for famous CNN algorithms in the literature, such as VGGNet, AlexNet, and Inception. As a result, two CNN algorithms with an accuracy of 65% were obtained (Agrawal and Mittal, 2020).

Searchers who know feature extraction is important in deep learning used VGG-19 architecture in camera systems. Additionally, they showed that systems that applied transfer learning could extract more information than the original network while reducing the training time (Gao, Zhang, and Li, 2020). Automated analysis and results in medical imaging systems have become a popular research area with enhancements in deep learning algorithms. A model was developed to help detect diabetic retinopathy, which threatens the retinal health of diabetic patients. AlexNet and VGG-19 algorithms were investigated for the model. As a result, searchers found that the VGG-19 model is more suitable than AlexNet according to classification accuracy and elapsed time (Mateen et al., 2019).

Training accuracy and validation accuracy were used as performance criteria in a study using a two-layer deep convolutional neural network by dividing the manually collected data into five emotion classes. In addition, the adam optimizer was used to reduce the loss function, and as a result, the system was tested to have an accuracy of 78.04% (Pranav et al., 2020).

In another study, it was mentioned that human emotions consist of 6 primary emotions in psychology, and the database created based on this was initially feature extraction by applying the Gabor filter and then trained with a convolutional neural network. As a result of the training, it was suggested that the learning process accelerated thanks to the Gabor filter (Taghi Zadeh et al., 2019).

Considering the shortcomings of facial recognition systems, two essential shortcomings stand out. These deficiencies are in the form of translation in the face image and the robustness of the classifier. A clever method has been proposed to overcome these shortcomings. This method uses stationary wavelet entropy for feature extraction, and a single hidden layer feedforward neural network model is preferred as a classifier. The Jaya algorithm prevents the classifier from reaching its local optimum points. This system was applied to a database of 700 images with 20 subjects, and as a result, when examined in terms of accuracy, it was observed that the db4 wavelet gave the best result in the whole db wavelet family (Wang et al., 2018).

When facial recognition systems are examined, it is vital to properly analyze the positions of reference points and the rate of change in density. With this in mind, a study was carried out using the gradient of the image and the laplacian as features in addition to the original images in the convolutional neural network. KDEF and FERplus datasets were used in this study. As a result, the accuracy value increased by 5% compared to the original system (Pandey et al., 2019).

It is striking that convolutional neural networks are mostly used in face recognition systems. In one of the studies in which this method was used, a visual geometry group model was used to improve the results in addition to the model. Commonly used CK+, MUG, and RAFD databases were tested to evaluate this structure. As a result, it was argued that an increase in performance in face recognition was observed (Fathallah et al., 2017).

It is seen that real-time emotion detection is at the forefront of the field of health. From this point of view, it has become important to identify people with physical disabilities or autism. Based on this, it was desired to perform facial image analysis using facial markings and EEG signals. This study uses convolutional neural networks and long-short-term memory classifiers, together with an optical flow algorithm that will be successful even in irregular light and rotation. The Lucas-Kande optical flow algorithm detected the virtual markers at the positions defined on the subject's face, and the distance between the face center and the detected point was used as the feature. In addition, the significance ratio of this distance was verified using a one-way analysis of variance. On the other hand, four different EEG signals were used as features. In light of all this, the features were validated by five crossover methods and then given to the LSTM and CNN classifiers. The results of these two classifiers were compared, and it was argued that the CNN classifier resulted in higher accuracy (Hassouneh et al., 2020).

When looking at face recognition systems from another perspective, it has been observed that their usage areas are also used in areas such as identity verification, security, and criminal investigations. At this point, systems capable of real-time recognition are preferred when it comes to intelligent systems used. When we look at the studies on these systems, an example of a three-stage recognition system study in which the data processed with haar cascade detection is analyzed using the Keras convolutional neural network model. This work is a work developed in python using the Open CV library (Hussain and Salim Abdallah Al Balushi, 2020).

When the studies are generalized, it is observed that there are studies that ignore the fact that each facial expression can contain more than one emotion and that the definition of the expression is unclear. At this point, to resolve this ambiguity and reach more logical conclusions on the subject, a learning approach to label distribution and, in addition to this, a study that can perform emotion distribution by making use of local label correlations has been carried out. In order to calculate the aforementioned local tag correlation, a low-order structure is preferred. Experiments have been done on comparative mimic databases, and it has been suggested that the results are more successful (Jia et al., 2019).

In another proposed study for facial expressions, a classification method using a least squares variant of a support vector machine with a radial basis function (RBF) kernel operating in three stages has been proposed. In the first stage, wavelet transform type II was used to extract the image features, and in the second stage, the principal component analysis (PCA) + linear discriminant analysis approach was used to increase compactness. In the last step, classification is made with the radial basis function kernel and the least squares variant method of the support vector machine. The accuracy of this method was examined on the extended Cohn-Kanade(CK)+ and JAFFE datasets (Kar et al., 2019).

In order to avoid the difficulty of generalization in real-time face recognition systems, a deep neural network architecture is proposed in the widely used dataset. A model consisting of two convolutional layers, each followed by maximum pooling and four initial layers, is presented. Validations were performed on various datasets such as multiple, MMI, FERA, SFEW, CK+, DISFA, and FER2013 (Mollahosseini et al., 2016).

Another example of studies carried out in the health sector is the portable, real-time emotion identification method developed to provide autistic children with ease of communication. This method demonstrates the hardware availability of principal component analysis (PCA). In addition, the proposed method has been implemented on Virtex 7 XC7VX330T FFG1761-3 FPGA (Smitha and Vinod, 2015).

AutoML-based neural architecture search showed a different and remarkable performance in image processing studies as a different perspective to the studies on the movements of the loss function. In the study, NAS technology was used for face recognition together with the reinforcement learning strategy. Based on reinforcement learning, the NAS part was optimized and included the policy gradient algorithm for automatic scanning of the system with the cross-entropy loss value (Zhu et al., 2020).

When we consider face recognition systems in terms of sensor information, the sophisticated system is an essential part of image analysis. In a study on this subject, sensors were examined in three categories. While the first part is used for the processing of the background part, the second part analyzes the lighting variation, sess, and depth. The last part includes sensors used for filtering out unnecessary parts. Subsequently, an expression identification system using this data was designed (Duncan et al., n.d.).

Looking at the video sampling size, it was seen that eight automatic classifiers were used and tested in a study. The 937 videos with corrupted and spontaneous data were used, and results were obtained and compared among eight classifiers with accuracy values of 48% and 62%. The results of the classifiers, whose performances were compared among themselves, were compared with the results of human observers, and it was observed that they gave approximate results (Dupré et al., 2020).

In the case of humanoid robots, real-time humanoid robots have not been successfully applied to facial expression processing. Considering this issue, the designed methods generally consist of convolutional neural network models. The designed neural network model was compared with AlexNet and VGG16 architectures for performance in a study. The system consists of two stages, and 18900 data points are used to perform face recognition first. The next stage is now the mimic recognition part and consists of 5000 data points. This structure has been tested on humans in real time, and the methods have been compared among themselves. While using AlexNet, 85% to 64% accuracy rates were seen, 100% and 73% accuracy were achieved when VGG16 architecture was used. The proposed method achieved 87% and 67% accuracy rates, and the error rate was 2.52% (Dwijayanti et al., 2022).

As an example of expression recognition studies, another study proposes a mimic detection method based on edge detection in an image. In this study, it is added to the feature image after the edge information is removed to preserve the textural information. After feature extraction, size reduction is performed using the maximum barking method. In addition, in the last step, this structure is decomposed with the help of the Softmax classifier. This build has been tested with the LFW portion of the FER-2013 database (Zhang et al., 2019).

In another study, mimic identification was performed using the Kernel Extreme Learning Machine classifier, the Variational Mode Decomposition method, and the Whale Optimization method. The VMD method is used in the mode processing of the input image. It is mentioned that the reason for using this method is that it is preferred to know the edge and shape properties. Principal component analysis and Linear discrimination analysis methods were preferred to avoid the difficulty of calculating high dimensions. SVM and Least Squares SVM (LS-SVM) methods are used for face discrimination. The WO-KELM method was preferred as the classifier. This model has been compiled in the JAFFE and extended Cohn-Kanade(CK)+ database (Kar et al., 2022).

A two-part convolutional neural network model is proposed in another mimic diagnosis study using convolutional neural networks. First, the background part is separated from the images, and then feature vector extraction is performed. In addition, an expression vector was created to find regular mimics. The model has been tested with a database of 10000 images. One of the specific features of this model is that convolutional networks operate in series. The last layer is used to detect duplicates and weights. Since it differs from single-level convolutional networks at this point, the accuracy rate increases at this level (Mehendale, 2020).

In a study in which classification is done using two machine learning algorithms, in which real-time emotion recognition applications can be trained offline, face recognition is made through cascading classification with the help of Adaboost, and in addition, localized image information-based features (NDF) are extracted. Since the designed frame has a modular structure, it has a structure to be developed for N number of facial expressions. The model works regardless of gender and skin color. It has been tested and evaluated for five different data sets and has been suggested to give 13% better results for static facial expressions than reference methods (Alreshidi and Ullah, 2020).

Duncan et al. developed a convolutional neural network that classifies real-time human emotions with the help of transfer learning. The study had a unique dataset, and transfer learning was used on the connection layers of the pre-trained convolutional neural network. In addition, after the face identification was made, the emoji of the facial expression detected during the live video stream was pasted on the faces of the people in the video, making the flow continue (Duncan et al., n.d.).

Considering a study in which the transfer learning algorithm is used in emotion identification, facial accessories were developed to solve lighting inequalities and rotation problems. The study used pre-trained Resnet50, VGG19, Inception V3, and Mobile Net architectures. Trained ConvNet link layers have been discarded, and complete links added. Added links are only inserted into the training to update the weights. Experiments took place in the CK+ database (Chowdary et al., 2021).

CHAPTER 2: METHODOLOGY

The methodology chapter explains the deep learning algorithms and their usage in the project. Moreover, this chapter clarifies the requirements for implementation.

2.1 Theory

2.1.1 Convolutional Neural Network

A Convolutional Neural Network (CNN, or ConvNet) are multi-layer neural network designed to identify visual structures directly from pixel images with reprocessing. The layers of the network and their purposes are as follows (Ergin, 2022).

- *Convolutional Layer:* The convolutional layer is used for extracting high-level and low-level features by using the filter.
- *Non-Linearity Layer:* The non-linearity layer is used for introducing the non-linearity to the system. It is also called the activation layer. Sigmoid, tanh, and ReLU are the most popular functions in this layer.
- *Pooling Layer:* The pooling layer is used for down sampling network parameters and reducing the calculation. Also, the pooling layer checks the suitability of the network. Max pooling is the most popular algorithm; moreover, average pooling and L2-norm pooling are also used.
- *Flattening Layer:* The flattening layer is used for preparing the input of the classical neural network. The input is converted to a vector.
- *Fully Connected Layer:* The fully connected layer is used for learning via a neural network.

This article uses the AlexNet, GoogleNet, and VGG architectures are prevalent CNN models.

2.1.1.1 AlexNet

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton developed AlexNet architecture based on CNN, which is described in figure 1. AlexNet consists of 5 convolutional layers and three fully connected layers, as in table 1. The input of the network is 227by 227-pixel RGB images. Although standard neural networks use tanh or sigmoid, AlexNet uses ReLU (Rectified Linear Unit) as activation in non-linear parts. ReLU function has a time advantage when the model is being trained. Also, sigmoid function derivative approaches zero and causes a vanishing gradient. Therefore, it becomes difficult to update the weights in the model. AlexNet uses max-pooling in pooling layers to reduce the number of calculated parameters in the network. Approximately 60 million parameters are calculated with AlexNet. AlexNet completed the ImageNet Large Scale Visual Recognition Challenge with a 15.3% error rate (Krizhevsky, Sutskever and Hinton, 2012).

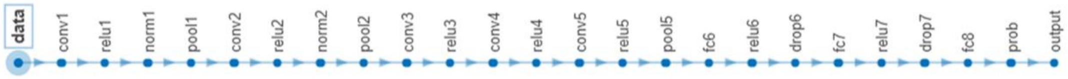


Figure 1. AlexNet Architecture

Table 1. AlexNet Layers

	Layers	
	Name	*Type*
1	data	Image Input
2	conv1	Convolution
3	relu1	ReLU
4	norm1	Normalization
5	pool1	Max Pooling
6	conv2	Convolution
7	relu2	ReLU
8	norm2	Normalization
9	pool2	Max Pooling
10	conv3	Convolution
11	relu3	ReLU
12	conv4	Convolution
13	relu4	ReLU
14	conv5	Convolution
15	relu5	ReLU
16	pool5	Max Pooling
17	fc6	Fully Connected
18	relu6	ReLU
19	drop6	Dropout
20	fc7	Fully Connected
21	relu7	ReLU
22	drop7	Dropout
23	fc8	Fully Connected
24	prob	Softmax
25	output	Classification Output

2.1.1.2 GoogleNet

GoogleNet, developed by researchers at Google, consists of 22 layers (27 layers including pooling layers), as in table 2. The nine inception modules provide feature detection through convolutions with different filters at different scales, as in figure 2. Moreover, inception module reduces the computational cost of training an extensive network through dimensional reduction. The average pooling layer takes a mean from all the feature maps produced by the last inception module. A dropout layer prevents overfitting the network and is used just before the linear layer. The dropout technique randomly reduces the number of interconnecting neurons within a neural network. The connected neurons are reduced randomly in this layer. GoogleNet architecture's last two layers are the linear layer which consists of 1000 hidden units, and the softmax layer, which uses the softmax function. The GoogleNet architecture was performed in the ILSVRC 2014 classification challenge with an error of 6.67% (Alake, 2021; Szegedy et al., 2015).

Table 2. GoogleNet Layers

	Layers	
	Name	*Type*
1	data	Image Input
2	conv1-7x7_s2	Convolution
3	conv1-relu_7x7	ReLU
4	pool1-3x3_s2	Max Pooling
5	pool1-norm1	Normalization
6	conv2-3x3_reduce	Convolution
7	conv2-relu_3x3_reduce	ReLU
8	conv2-3x3	Convolution
9	conv2-relu_3x3	ReLU
10	conv2-norm2	Normalization
11	pool2-3x3_s2	Max Pooling
12	inception_3a	Inception (3a)
13	inception_3b	Inception (3b)
14	pool3-3x3_s2	Max Pooling
15	inception_4a	Inception (4a)
16	inception_4b	Inception (4b)
17	inception_4c	Inception (4c)
18	inception_4d	Inception (4d)
19	inception_4e	Inception (4e)
20	pool4-3x3_s2	Max Pooling
21	inception_5a	Inception (5a)
22	inception_5b	Inception (5b)
23	pool5-7x7_s1	Average Pooling
24	pool5-drop_7x7_s1	Dropout
25	loss3-classifier	Fully Connected
26	prob	Softmax
27	output	Classification Output

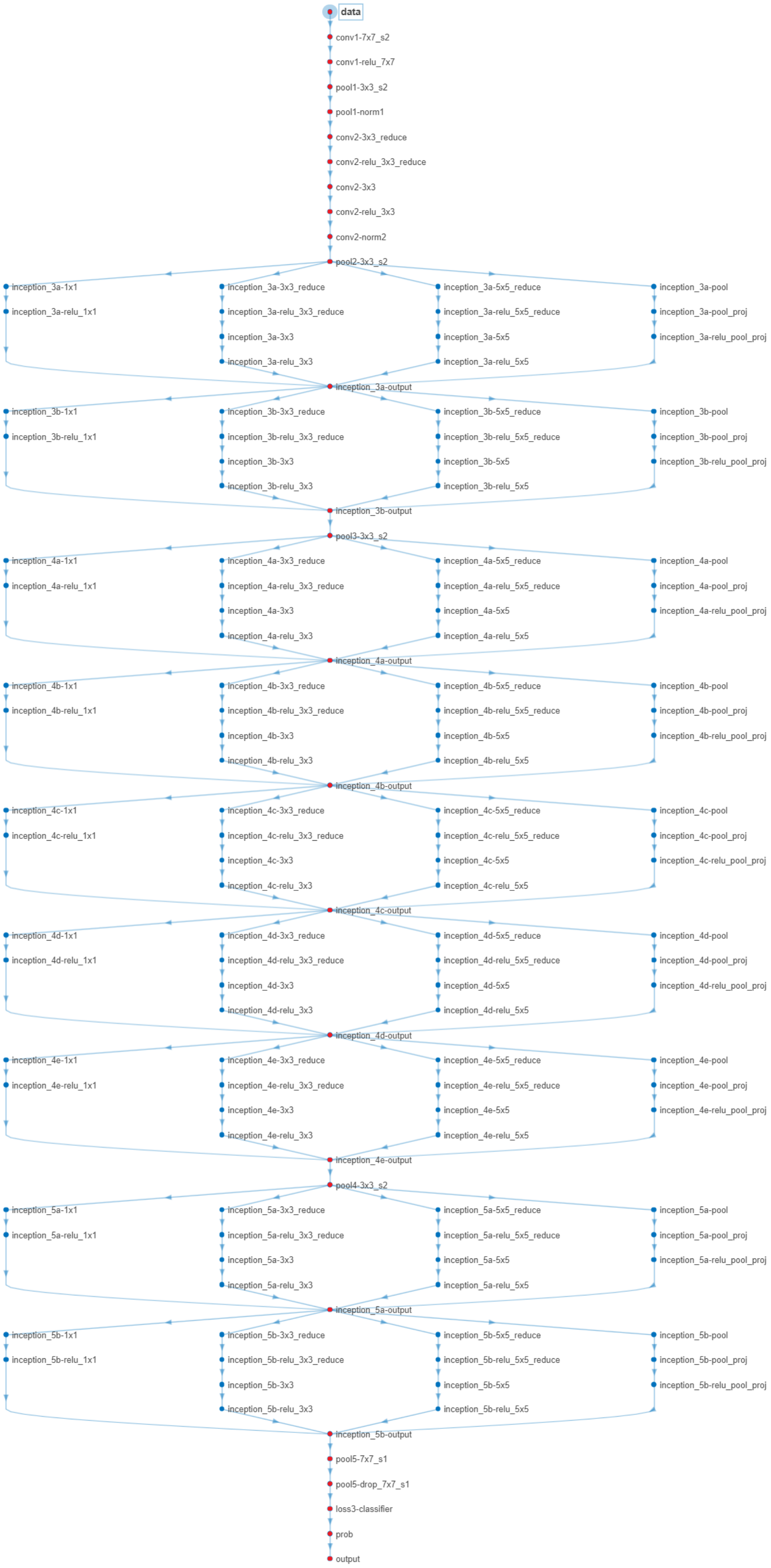


Figure 2. GoogleNet Architecture

2.1.1.3 VGG-19

Visual Geometry Group develops VGG-19 at Oxford, as in figure 3. VGG-19 architecture consists of a total of 24 main layers, which are 16 convolutional, five pooling, and three fully connected layers, as in table 3. Furthermore, it contains approximately 138 million parameters. The network has an image input size of 224-by-224 RGB (Simonyan and Zisserman, 2015).

Table 3. VGG-19 Layers

	Layers	
	Name	*Type*
1	input	Image Input
2	conv1_1	Convolution
3	relu1_1	ReLU
4	conv1_2	Convolution
5	relu1_2	ReLU
6	pool1	Max Pooling
7	conv2_1	Convolution
8	relu2_1	ReLU
9	conv2_2	Convolution
10	relu2_2	ReLU
11	pool2	Max Pooling
12	conv3_1	Convolution
13	relu3_1	ReLU
14	conv3_2	Convolution
15	relu3_2	ReLU
16	conv3_3	Convolution
17	relu3_3	ReLU
18	conv3_4	Convolution
19	relu3_4	ReLU
20	pool3	Max Pooling
21	conv4_1	Convolution
22	relu4_1	ReLU
23	conv4_2	Convolution
24	relu4_2	ReLU
25	conv4_3	Convolution
26	relu4_3	ReLU
27	conv4_4	Convolution
28	relu4_4	ReLU
29	pool4	Max Pooling
30	conv5_1	Convolution
31	relu5_1	ReLU
32	conv5_2	Convolution
33	relu5_2	ReLU
34	conv5_3	Convolution
35	relu5_3	ReLU
36	conv5_4	Convolution
37	relu5_4	ReLU
38	pool5	Max Pooling
39	fc6	Fully Connected
40	relu6	ReLU
41	drop6	Dropout
42	fc7	Fully Connected
43	relu7	ReLU
44	drop7	Dropout
45	fc8	Fully Connected
46	prob	Softmax
47	output	Classification Output

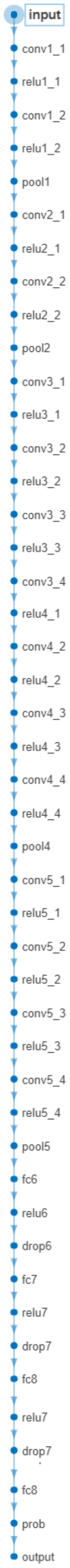


Figure 3. VGG-19 Architecture

2.1.2 Transfer Learning

The increasing dataset and complex model architecture cause difficulties in using the standard computer while the training process. Especially training process that takes days or sometimes weeks makes it impossible to work on standard computer processors. Learning from scratch, which is commonly performed for each learning process, causes vital problems in these years because of spending time in training. As a result, the method that is called transfer learning was developed. In this way, learned information from some tasks in other tasks will be possible and advantageous to use in other tasks. In other words, information such as features, and weights, obtained from previously trained models can be used for a new task (Kızrak, 2020). The working principle of transfer learning is summarized in the following figure 4 to provide a better understanding.

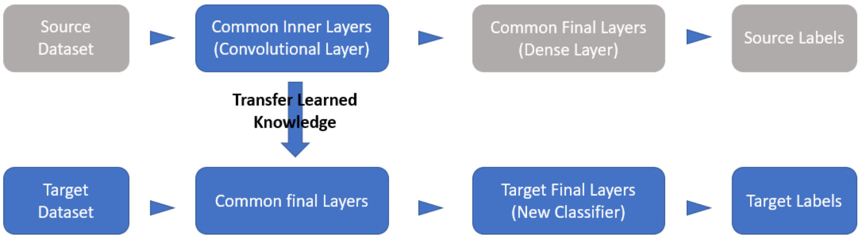

Source Dataset
Common Inner Layers (Convolutional Layer)
Common Final Layers (Dense Layer)
Source Labels
Transfer Learned Knowledge
Target Dataset
Common final Layers
Target Final Layers (New Classifier)
Target Labels

Figure 4. Transfer Learning

2.1.3 Long Short-Term Memory

LSTM is a special type of recurrent neural network (RNN) architecture developed by Hochreiter and Schmidhuber. The LSTM architecture consists of sequential blocks that repeat each other. These sequential blocks are a “forget” gate, input gate, a update gate, and an output gate. Firstly, information that will be forgotten is specified by using the inputs h_{t-1} and x_t. These are done in the forget gate by using the sigmoid activation function as in equation 1.

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f) \qquad 1$$

Secondly, information is updated in the input gate by using sigmoid as the activation function as in equation 2. Then, the new information is determined with tanh function as in equation 3.

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i) \qquad 2$$

$$g_t = tanh(W_g x_t + R_g h_{t-1} + b_g) \qquad 3$$

Then, the new information is constructed by equation 4.

$$c_t = f_t c_{t-1} + i_t g_t \qquad 4$$

The output of the system is determined by the equations 5 and 6.

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o) \qquad 5$$

$$h_t = o_t \tanh(c_t) \qquad 6$$

In the figure 5, h_t and c_t denote the hidden states and cell states at time t, respectively while the current state is represented with (t-1). Additionally, the W, R, and b signify the input weights, the recurrent weights, and the bias of each component, respectively. The described process continues iteratively to minimize the difference between the actual training values and the LSTM output values (Hochreiter and Schmidhuber, 1997; KARA, 2019).

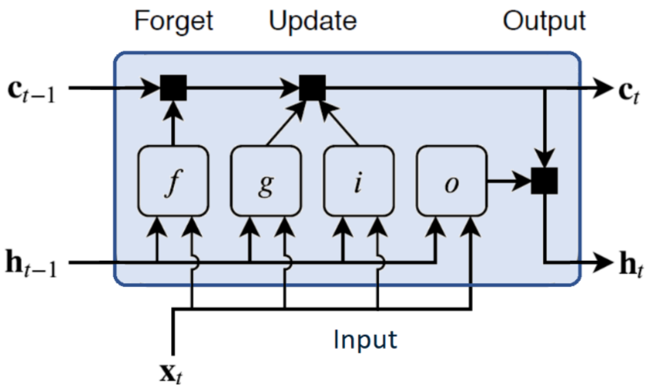


Figure 5. LSTM Architecture

2.1.4 Hybrid Model

The CNN architecture is not suitable for continuous dynamic images, and RNN has an internal memory to process dynamic data. Thus, CNN and LSTM combinations can be used to obtain a more beneficial model as in figure 7. CNN is used for deep feature extraction, and LSTM is used for a classifier for the offered hybrid model (Khamparia et al., 2019; Livieris, Pintelas and Pintelas, 2020; Li et al., 2019). Moreover, model structure is as in the figure 6.

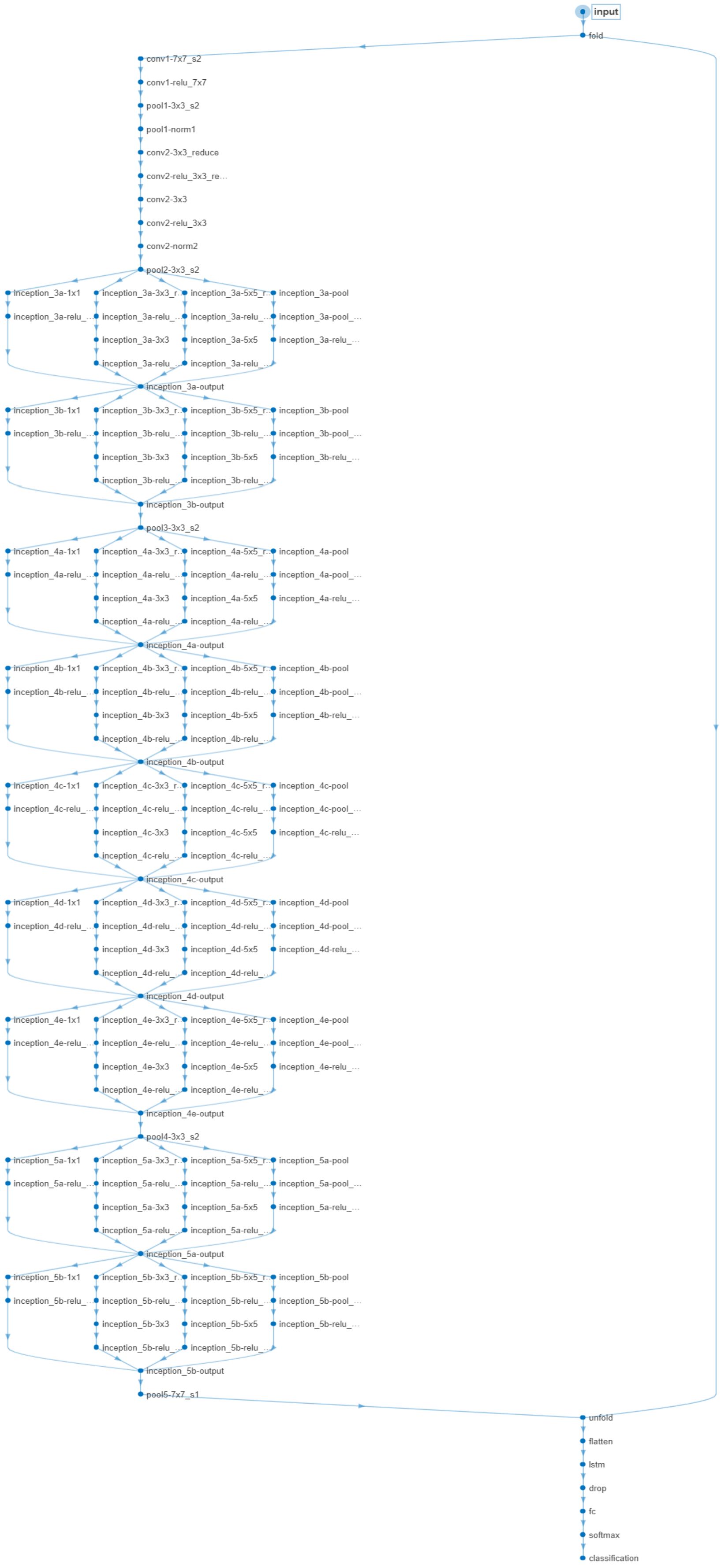


Figure 6. CNN-LSTM Architecture

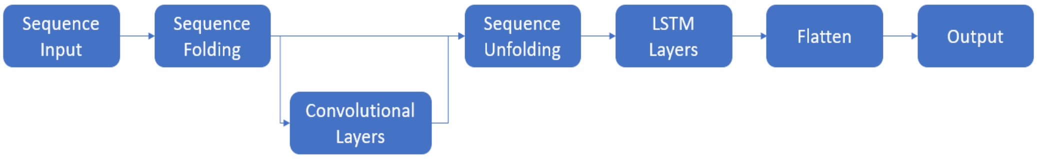


Figure 7. CNN-LSTM Model

2.2 Implementation Details

The implementation details, such as the feature of the dataset, the system model and used program, are given in section 2.2.

2.2.1 System Model

The model of the system is illustrated in figure 8. The image is obtained from the camera. The face in the image is caught by the detection algorithm. The catch image is cropped and resized. Then, this is given as an input to the learning algorithm that is trained to recognize seven human emotions such as anger, disgust, fear, happiness, neutral, sad, and surprise. Lastly, classification output is taken as a facial expression result.

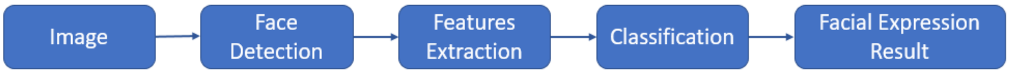


Figure 8. System Model

Viola-Jones algorithm, which is a popular detection model, is preferred as a detector in the system. It uses rectangular features to identify the particular object in the image. In this way, the system trained for face recognition is not interested in non-face areas (Viola and Jones, 2001).

2.2.2 Dataset

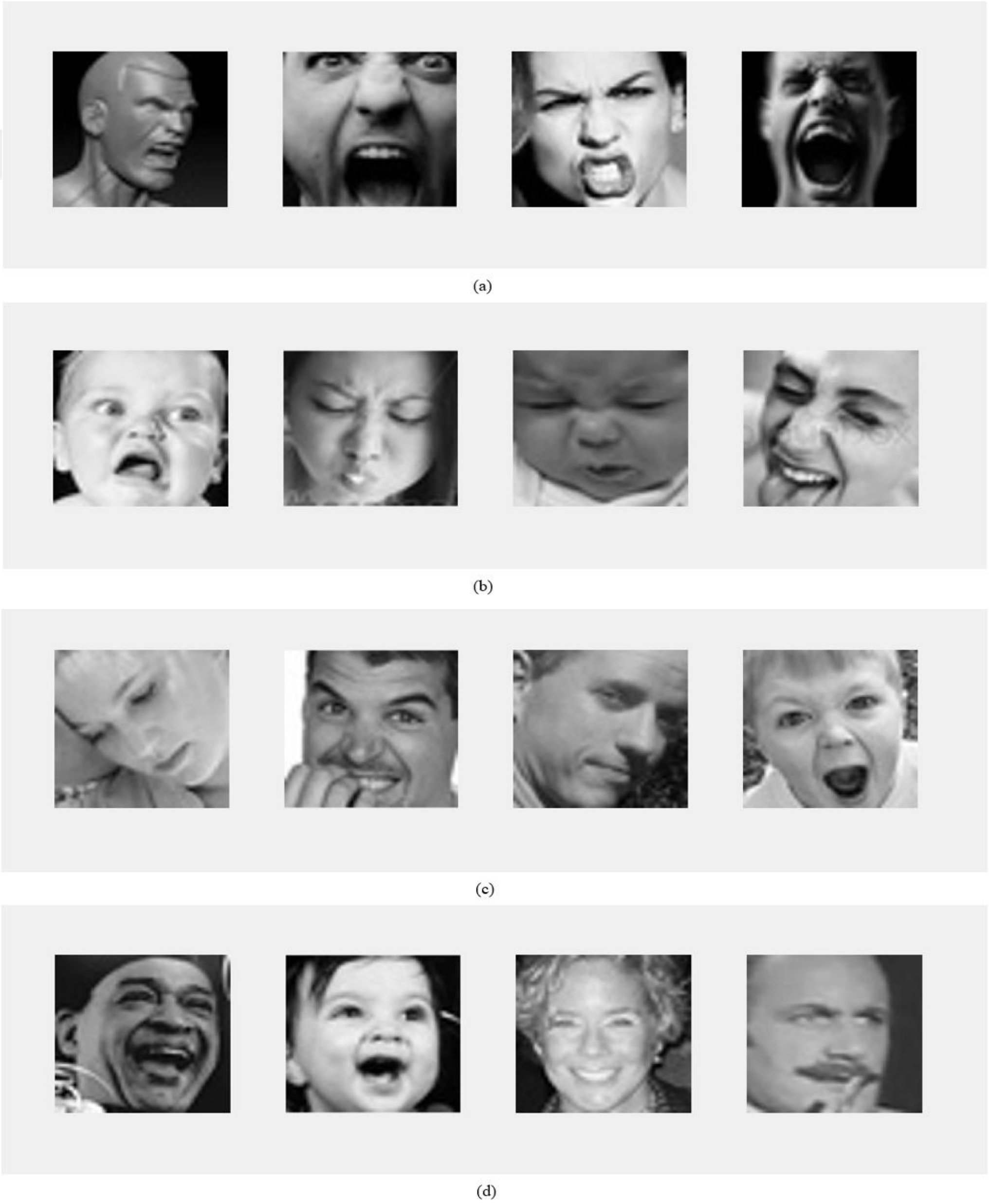
2.2.2.1 FER2013

FER2013 (FER-2013, 2022) is one of the most popular datasets for facial emotion recognition implementation. FER2013 has a considerable number of examples as seen in table 4. FER 2013 consists of 35887 grey images which are 48 by 48-pixel. The train set and test set have 28709 and 7178 images, respectively. There are seven emotions:

anger, disgust, fear, happiness, neutral, sad, and surprise. These are as sampled in figure 9.

Table 4. Number of Images in FER2013

	Angry	***Disgust***	***Fear***	***Happy***	***Neutral***	***Sad***	***Surprise***
Train	3995	436	4097	7215	4965	4830	3171
Test	958	111	1024	1774	1233	1247	831



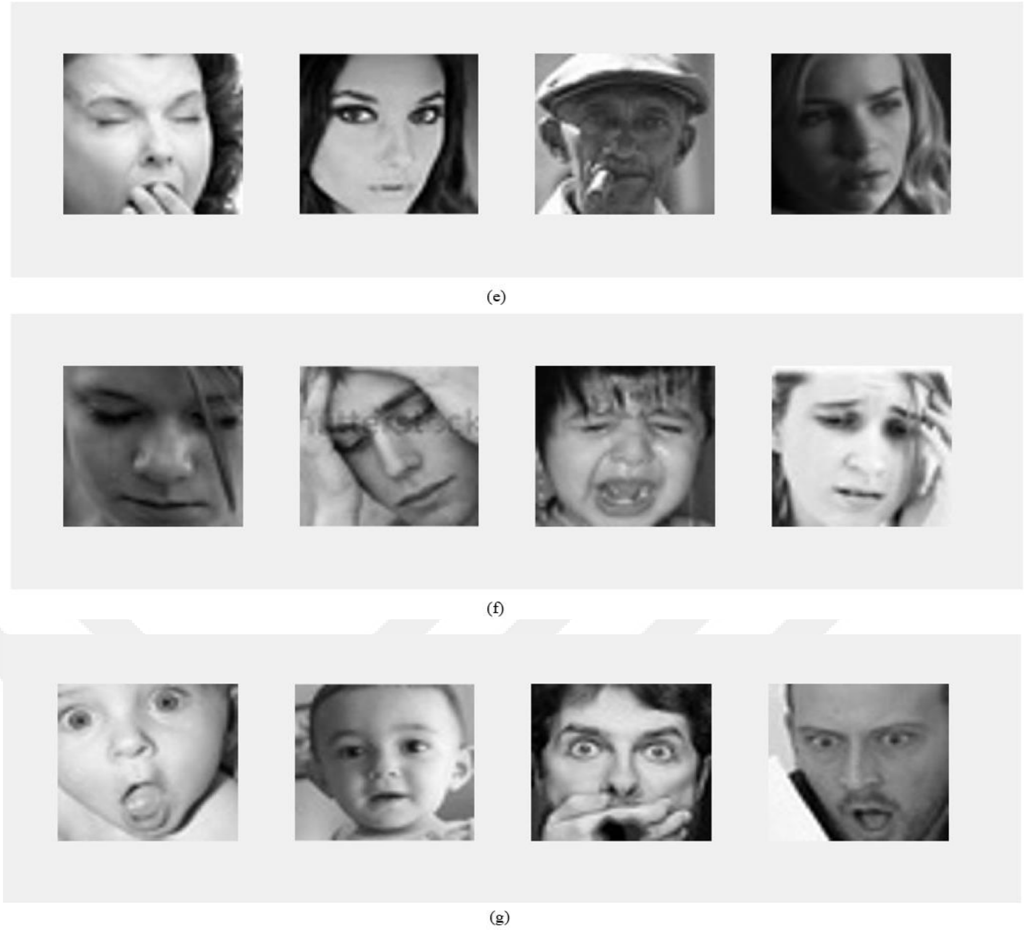
(a)

(b)

(c)

(d)

Figure 9. FER2013 Train Image Examples for Each Emotions (a) Angry (b) Disgust (c) Fear (d) Happy (e) Neutral (f) Sad (g) Surprise



(e)

(f)

(g)

Figure 9 (continued)

2.2.2.2 CK+

CK+ (CKPLUS, 2022) is one of the most popular datasets for facial emotion recognition implementation, as seen in figure 11. CK+ has very few examples, as seen in table 5. CK+ consists of 981 grey images which are 48 by 48-pixel. There are seven emotions: anger, disgust, fear, happiness, contempt, sadness, and surprise. These are as sampled in figure 10.

Table 5. Number of Images in CK+

Angry	***Disgust***	***Fear***	***Happy***	***Contempt***	***Sadness***	***Surprise***
135	177	75	207	54	84	249

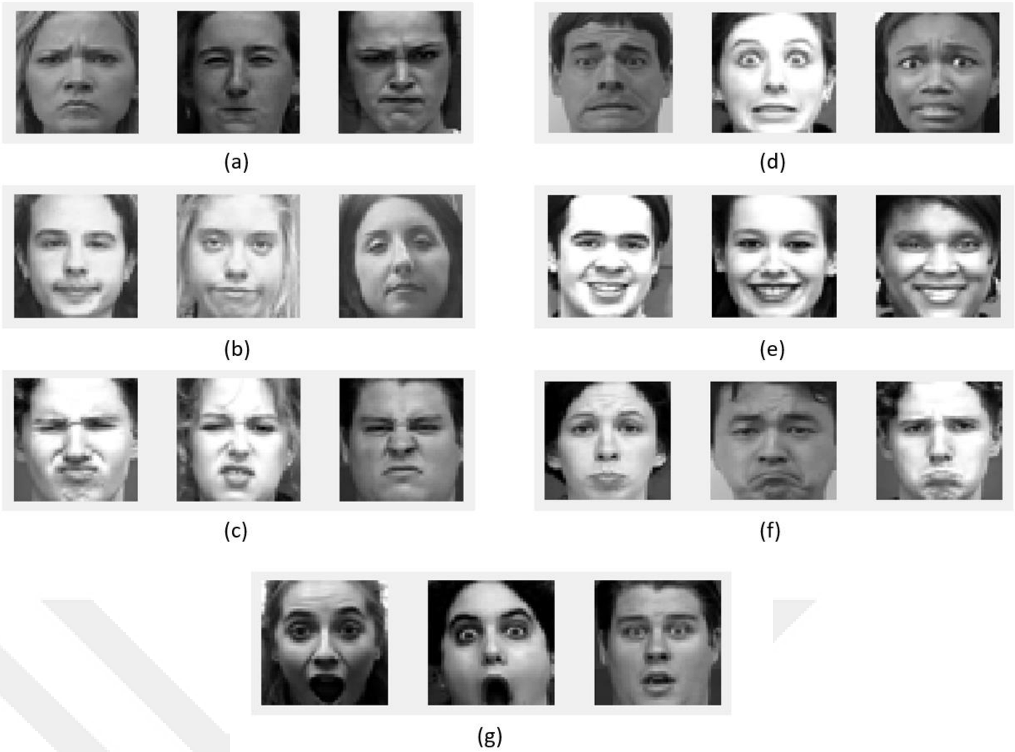


Figure 10. CK+ Image Examples for Each Emotions (a) Anger (b) Contempt (c) Disgust (d) Fear (e) Happy (f) Sadness (g) Surprise

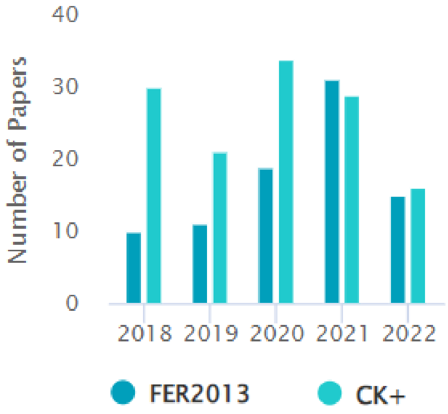


Figure 11. The popularity of Datasets in Kaggle

2.2.3 Simulation Environment

Analyses and real-time implementation were done in MATLAB2021a. The related deep learning toolboxes were initialized before the usage. Because of the high training speed of the graphics processing unit (GPU) (Gao, Zhang and Li, 2020), neural networks were trained using the NVIDIA GeForce840M.

CHAPTER 3: RESULTS AND DISCUSSIONS

The performance results for architectures AlexNet, GoogleNet, VGG-19, and CNN-LSTM hybrid model, the adequacy of the dataset for the implementation, and the applicability of the real-time systems are examined in the results and discussions chapter.

The validation accuracy used for performance criteria explains the generalization or classification ability of the model. The obtained validation accuracy results for GoogleNet and AlexNet architectures that were trained via transfer learning are in table 6. The CK+ was used as a database. The AlexNet and GoogleNet have validation accuracy results of 76.11% and 72.01%, respectively.

The elapsed time in training is another vital performance parameter. Table 6 shows that GoogleNet spent nearly 1,5 hours and AlexNet spent nearly 23 minutes in training.

Table 6. Validation accuracy for CNN algorithms for CK+

Algorithm	***Validation Accuracy***	***Elapsed Time***	***Epoch Number***
GoogleNet	72.01%	74 min 40 sec	6
AlexNet	76.11 %	23 min 58 sec	6

The learning algorithms' performances were tested via public images. The test image in figure 12 was a colored image of anger. AlexNet named this sample image disgust with 80%, and GoogleNet named it again disgust with 52%. At the end of the implementation, all learning algorithms failed to achieve the correct result for the tested

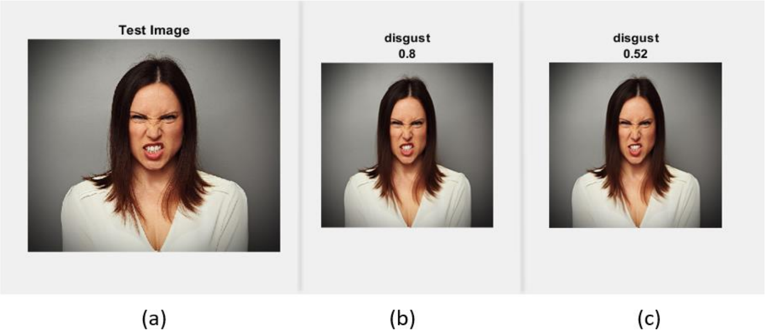


Figure 12. Algorithms results for public RGB image (a) anger test image (b) AlexNet result (c) GoogleNet result

RGB images. When the grey anger image in figure 13 was tested, AlexNet labeled the image as disgust with 74% and GoogleNet with disgust with 49%. All of them again missed the correct results for the second test image.

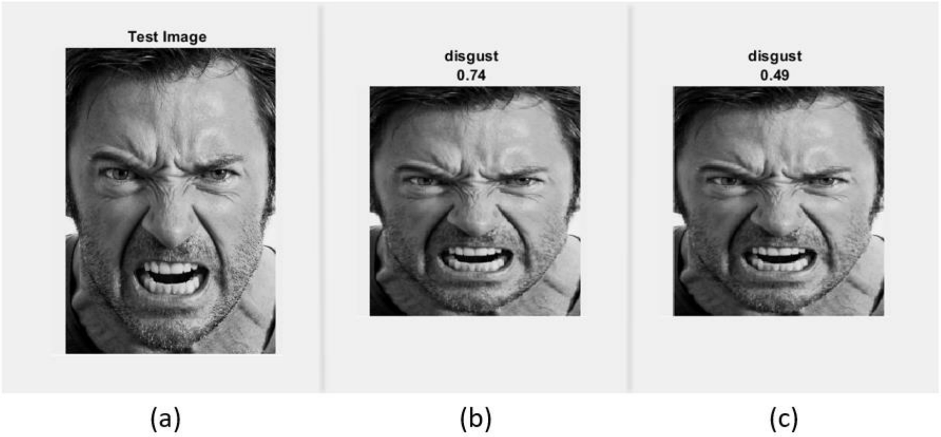


Figure 13. Algorithms results for public grey image (a) anger test image (b) AlexNet result (c) GoogleNet result

The systems tend to fail when small data is used in the deep learning process. The neural network layers are stacked on top of each other in multi-layered architectures. As a result, more complex parts of the data can be learned at every new layer. In other words, the small datasets used in the systems cause poor learning (causaLens, n.d.). Although the validation accuracies are 72.01% and 76.11% for GoogleNet and AlexNet, respectively, investigated test images that are shown in figure 12 and figure 13 cannot be labeled successfully via AlexNet and GoogleNet. Therefore, the more extensive dataset was preferred for investigation and implementation.

The obtained validation accuracy results for GoogleNet, AlexNet, and VGG-19 were in table 7 when the FER2013 was used as a source. The VGG-19 has the highest result with 62.66%, while the GoogleNet and AlexNet have 60.7% and 57.67%. As a result, the VGG-19 is the best option for implementation. Additionally, table 7 shows that GoogleNet, AlexNet, and VGG-19 spent nearly 11 hours, 6 hours, and 2.5 days in the

Table 7. Validation accuracy for CNN algorithms for FER2013

Algorithm	***Validation Accuracy***	***Elapsed Time***	***Epoch Number***
GoogleNet	60.70 %	652 min 44 sec	6
AlexNet	57.67 %	353 min 6 sec	6
VGG-19	62.66 %	3697 min	6

training, respectively. Therefore, the AlexNet is the most suitable model for applications.

The public images were used to test the performance of the learning algorithms. The test image in figure 14 was a colored image of anger. AlexNet named this sample image fear with 34%, GoogleNet named it happiness with 48%, and VGG-19 named it neutral with 23%. At the end of the implementation, all learning algorithms failed to achieve the correct result for the tested RGB image. The test image in figure 15 is a grey anger image, and AlexNet labeled this image as anger with 52%, GoogleNet as anger with 97%, and VGG-19 with 98%. For the second image tested, all algorithms caught the result correctly. In addition, the results of VGG-19 and GoogleNet were not differing much from each other.

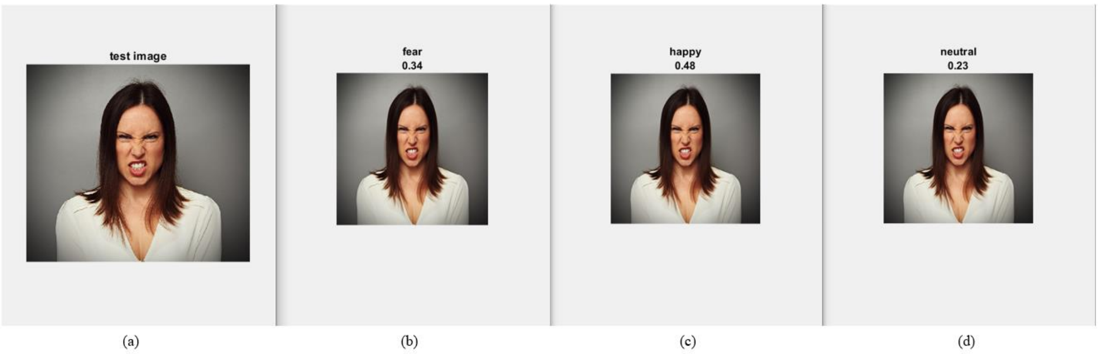


Figure 14. Algorithms results for public RGB image (a) anger test image (b) AlexNet result (c) GoogleNet result (d) VGG-19 result

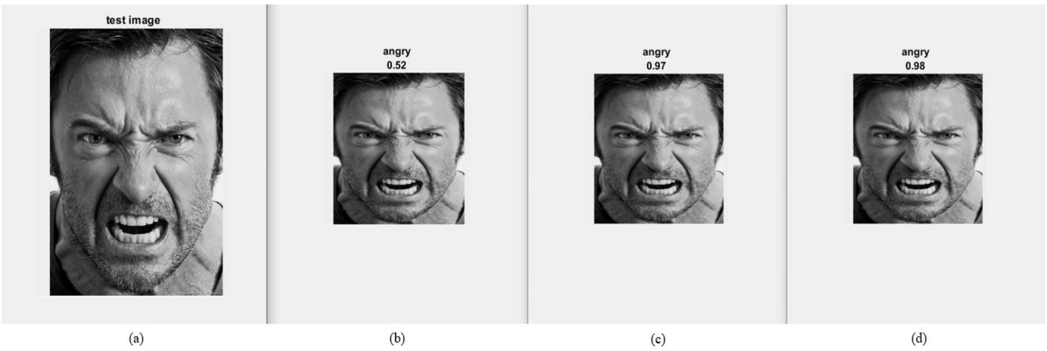


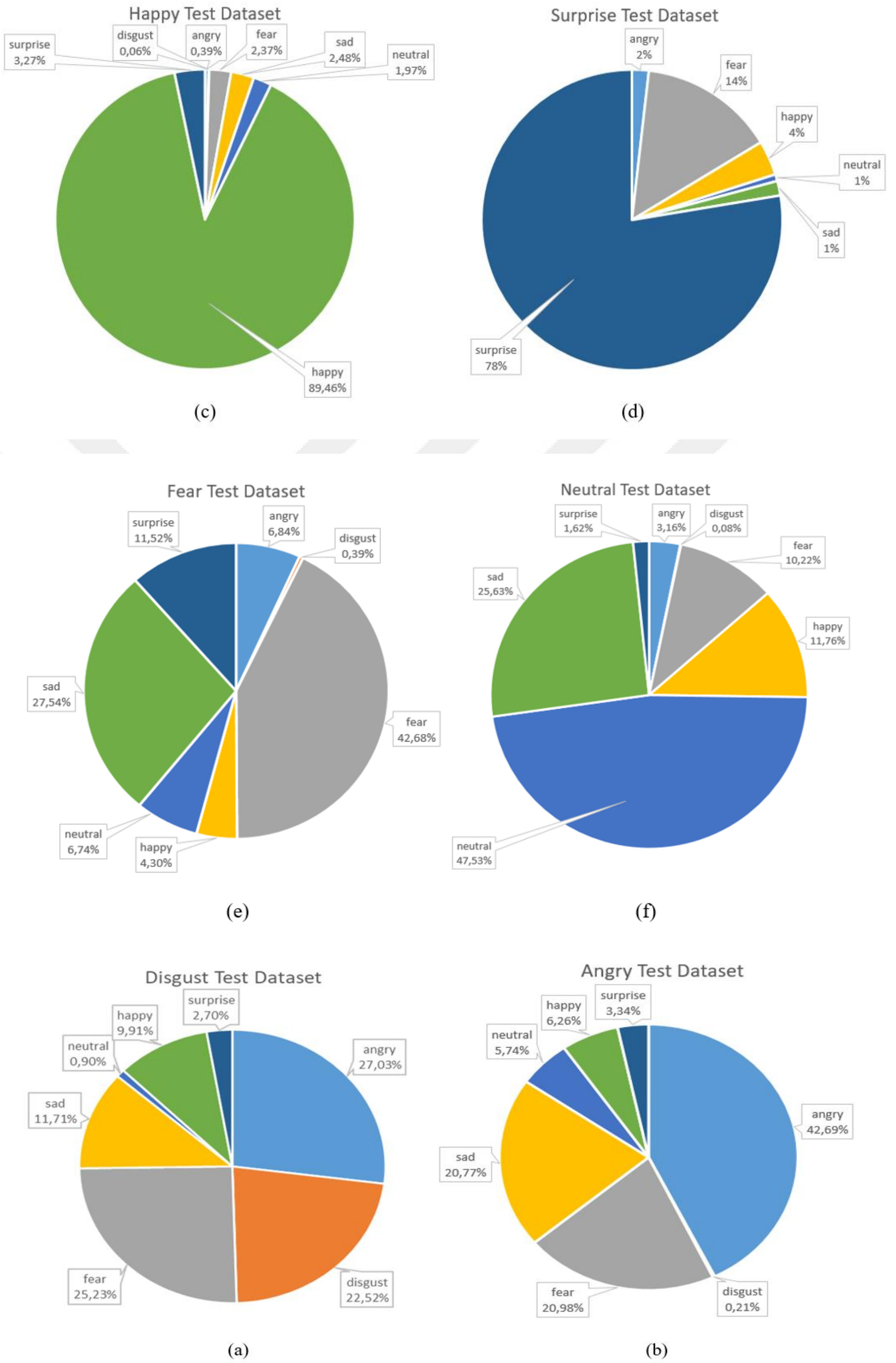
Figure 15. Algorithms results for public grey image (a) anger test image (b) AlexNet result (c) GoogleNet result (d) VGG-19 result

GoogleNet validation accuracy result is not less than AlexNet, and elapsed time in the training is not too much like VGG-19. Thus, the GoogleNet algorithm was chosen to convert the stationary to a dynamic model. The CNN-LSTM hybrid model provided this improvement. The obtained performance result for the CNN-LSTM model is in table 8. The spending time in the training increased by 11% compared with the original GoogleNet. Additionally, the reconstructed model showed a less than 1% decrease in the validation accuracy. As a result, the proposed hybrid model is transformed into the appropriate approach for real-time simulation.

Table 8. Validation accuracy for the hybrid model

Algorithm	***Validation Accuracy***	***Elapsed Time***	***Epoch Number***
CNN-LSTM	60.34 %	724 min 55 sec	6

The suitability of the FER2013 is investigated using test images inside the dataset. When the performance of the trained algorithm is examined on the test dataset, the results for each emotion individually are in the following figure 16.

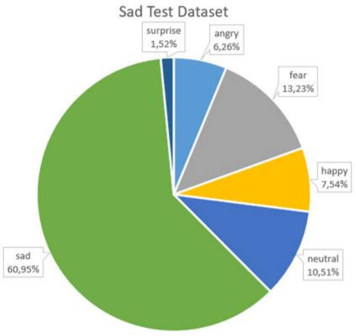


(c) (d)

(e) (f)

(a) (b)

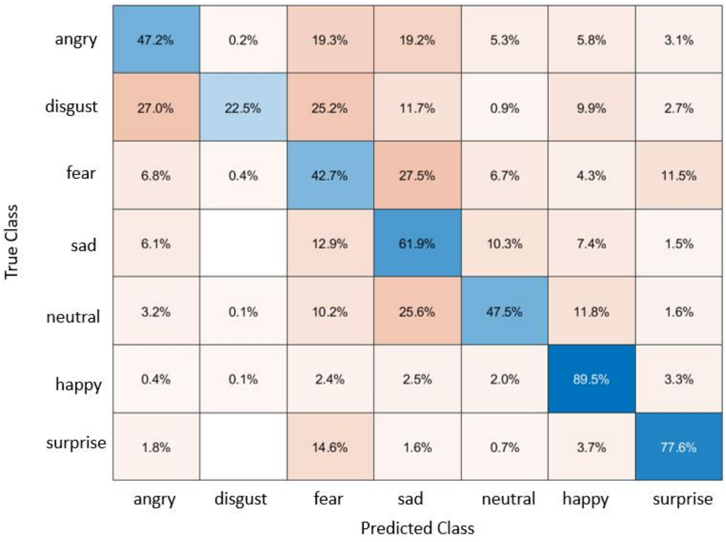
Figure 16. Test Dataset Performance Results for (a) Disgust (b) Angry (c) Happy (d) Surprise (e) Fear (f) Neutral (g) Sad



(g)

Figure 16 (continued)

In figure 16(a), the test dataset results for the emotion of disgust are 27% for anger, 25% for fear, and %22 for disgust. The model was trained with fewer samples for disgust than other emotions, as seen in table 4. Thus, the system does not catch the disgust effectively (Nguyen et al., 2019). If the emotion of anger is examined for test dataset results, the highest results are 42% for anger, and 20% for both fear and sadness, as seen in figure 16(b). In figure 16(c), 89% of the test dataset is labeled happy, and the other emotions are less than 4%. When the emotion of surprise is analyzed in figure 16(d), the highest two results of the test dataset are 78% for surprise and 14% for fear. Additionally, the test dataset results of the fear show the highest



True Class	angry	disgust	fear	sad	neutral	happy	surprise
angry	47.2%	0.2%	19.3%	19.2%	5.3%	5.8%	3.1%
disgust	27.0%	22.5%	25.2%	11.7%	0.9%	9.9%	2.7%
fear	6.8%	0.4%	42.7%	27.5%	6.7%	4.3%	11.5%
sad	6.1%		12.9%	61.9%	10.3%	7.4%	1.5%
neutral	3.2%	0.1%	10.2%	25.6%	47.5%	11.8%	1.6%
happy	0.4%	0.1%	2.4%	2.5%	2.0%	89.5%	3.3%
surprise	1.8%		14.6%	1.6%	0.7%	3.7%	77.6%

Figure 17. The confusion matrix for The CNN-LSTM hybrid model

results as 42% for fear, 27% for sad, and 11% for surprise in figure 16(e). The similarity in facial features confuses the analysis of different emotions and causes errors. As a result, fear analyses have a surprise, and surprise analyses have fear (Dandıl and Özdemir, 2019). In figure 16(f), the neutral emotion has 47% of neutral results and 25% of sad results when the test dataset is researched. Figure 16(g) shows 60% sad and 13% fear results when the test dataset of sad emotions is examined. In conclusion, the model's validation and test accuracy are directly affected by the datasets with unbalanced and unclear labeled samples. Although FER2013 is considered an enormous dataset, it is not a well-defined dataset because of these challenges (Naga, Marri, and Borreo, 2021).

The confusion matrix is the performance evaluation that allows the visualization of classification problems in machine learning. It is also called an error matrix. This matrix consists of an equal number of classification classes in two dimensions that are called actual and predicted (Sarang Narkhede, 2018). The confusion matrix in figure 17 shows that anger, disgust, fear, and neutral emotions have less than 60% accuracy performance. Also, sadness, happiness, and surprise emotions have more excellent performance than other emotions, with 61%, 89%, and 77%, respectively.

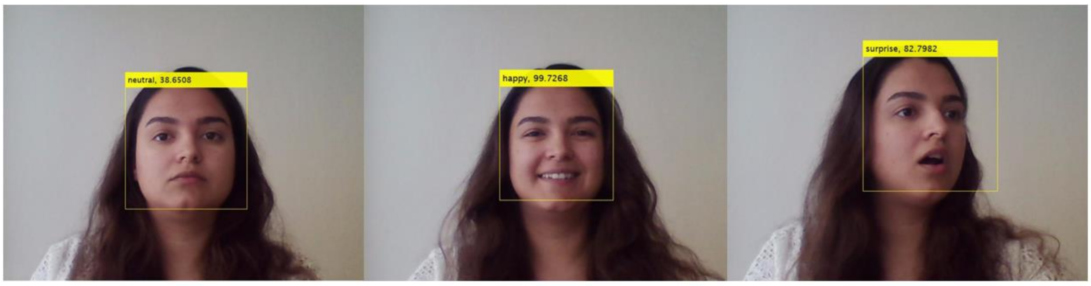


Figure 18. Simulation Results

The results are shown in figure 18 when the whole system is assembled for simulation as in the flow chart. Though FER2013 is not suitable for training and validation accuracy is 60.34 %, the proposed real-time system can give the wanted outputs. The elapsed time between the samples is nearly 1.57 seconds in real-time implementation.

CHAPTER 4: CONCLUSION

In this paper, CNN architectures, which are AlexNet, GoogleNet, and VGG-19, were investigated according to validation accuracy performance for emotion recognition. The GoogleNet, which has the average validation accuracy and spending time, was used to construct the CNN-LSTM hybrid model. The constructed model showed similar performance results to analyzed CNN algorithms according to validation accuracy (Özkara and Oğuz Ekim, 2022). The recommended model was used to implement real-time simulation. The hybrid models are the way that achieves various aims such as better performance and converting the domain from static to dynamic when the construction of the algorithms can allow combining parts of the different algorithms. In conclusion, the hybrid model approach can be used as a resource for real-time automated camera systems.

REFERENCES

Agrawal, A. and Mittal, N. (2020) *Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy*, The Visual Computer, 36(2), pp.405-412.

Alake, R. (2021) *Deep learning: Googlenet explained*, [Online]. Available at: https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765 (Accessed: 17 June 2022).

Alreshidi, A., Ullah, M. (2020) *Facial Emotion Recognition Using Hybrid Features,* Informatics [Online]. Available at: https://www.researchgate.net/publication/339359274_Facial_Emotion_Recognition_Using_Hybrid_Features (Accessed: 24 October 24 2022).

causaLens (n.d.) *Deep learning has a small data problem*, [Online] Available at: https://www.causalens.com/blog/deep-learning-has-a-small-data-problem/ (Accessed 15 October 2022).

Chowdary, M.K., Nguyen, T.N., and Hemanth, D.J. (2021) *Deep learning-based facial emotion recognition for human–computer interaction applications*, Neural Computing and Applications, 1-18.

Dandıl, E. and Özdemir, R. (2019) *Real-time facial emotion classification using deep learning*, Data Science and Applications, 2(1), pp.13-17.

Duncan, D.D., Shine, G. and English, C. (n.d.) *Facial Emotion Recognition in Real Time,* [Online] Available at: http://cs231n.stanford.edu/reports/2016/pdfs/022_Report.pdf (Accessed: 24 October 2022).

Dupré, D., Krumhuber, E.G., Küster, D. and McKeown, G.J. (2020) *A performance comparison of eight commercially available automatic classifiers for facial affect recognition,* PLOS ONE [Online]. Available at: https://www.researchgate.net/publication/340904344_A_performance_comparison_of_eight_commercially_available_automatic_classifiers_for_facial_affect_recognition (Accessed: 24 October 2022).

Dwijayanti, S., Iqbal, M. and Suprapto, B.Y. (2022) *Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network,* IEEE Access [Online]. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9864185 (Accessed: 24 October 2022).

Ergin, T. (2020) *Convolutional Neural Network (ConvNet yada CNN) nedir, nasıl çalışır?*,[Online]. Available at: https://medium.com/@tuncerergin/convolutional-neural-network-convnet-yada-cnn-nedir-nasil-calisir-97a0f5d34cad (Accessed: 29 June 2022).

Fathallah, A., Abdi, L. and Douik, A. (2017) *Facial Expression Recognition via Deep Learning*, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). Presented at the 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), IEEE, Hammamet, pp. 745–750.

Gan, Y. (2018) *Facial Expression Recognition Using Convolutional Neural Network*, In: ICVISP 2018: Proceedings of the 2nd International Conference on Vision, Image and Signal Processing.

Gao, Z., Zhang, Y. and Li, Y. (2020) *Extracting features from infrared images using convolutional neural networks and transfer learning*, [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S1350449519307431 (Accessed: 24 October 2022).

Hassouneh, A., Mutawa, A.M. and Murugappan, M. (2020) *Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods*, [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S235291482030201X (Accessed: 24 October 2022).

Hochreiter, S. and Schmidhuber, J. (1997) *Long Short-Term Memory*, Neural Computation [Online]. Available at: https://blog.xpgreat.com/file/lstm.pdf (Accessed: 24 October 2022) 9(8), pp.1735-1780.

Hussain, S.A. and Salim Abdallah Al Balushi, A. (2020) *A real time face emotion classification and recognition using deep learning model,* [Online]. Available at:

https://www.researchgate.net/publication/338431245_A_real_time_face_emotion_classification_and_recognition_using_deep_learning_model (Accessed: 24 October 2022).

Islam, M., Islam, M. and Asraf, A. (2020) *A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images*, Informatics in Medicine Unlocked [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S2352914820305621 (Accessed: 29 Juna 2022).

Jain, D., Shamsolmoali, P. and Sehdev, P. (2019) *Extended deep neural network for facial emotion recognition*, Pattern Recognition Letters [Online]. Availabe at: https://www.sciencedirect.com/science/article/pii/S016786551930008X (Accessed: 17 June 2022).

Jaiswal, S. and Nandi, G. (2019) *Robust real-time emotion detection system using CNN architecture*, Neural Computing and Applications [Online]. Available at: https://link.springer.com/article/10.1007/s00521-019-04564-4 (Accessed: 17 June 2022).

Jia, X., Zheng, X., Li, W., Zhang, C., Li, Z., 2019. Facial Emotion Distribution Learning by Exploiting Low-Rank Label Correlations Locally, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 9833–9842.

Kaggle.com. (2022) *CKPLUS*. [Online]. Available at: https://www.kaggle.com/datasets/shawon10/ckplus (Accessed: 16 October 2022).

Kaggle.com. (2022) *FER-2013*. [Online]. Available at: https://www.kaggle.com/msambare/fer2013 (Accessed: 29 June 2022).

Kar, N.B., Babu, K.S. and Bakshi, S. (2022) *Facial expression recognition system based on variational mode decomposition and whale optimized KELM*, Image Vision and Computing [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0262885622000749 (Accessed: 24 October 2022).

Kar, N.B., Babu, K.S., Sangaiah, A.K. and Bakshi, S. (2019) *Face expression recognition system based on ripplet transform type II and least square SVM*, Multimedia Tools and Applications, Vol. 78, pp. 4789–4812.

KARA, A. (2019) *Global Solar Irradiance Time Series Prediction Using Long Short-Term Memory Network*, [Online] Available at: https://dergipark.org.tr/en/download/article-file/878498 (Accessed: 17 June 2022).

Khamparia, A., Pandey, B., Tiwari, S., Gupta, D., Khanna, A. and Rodrigues, J. (2019) *An Integrated Hybrid CNN–RNN Model for Visual Description and Generation of Captions*, Circuits, Systems, and Signal Processing [Online]. Available at: https://link.springer.com/article/10.1007/s00034-019-01306-8 (Accessed: 29 June 2022).

Kızrak, A. (2020) *Derine Daha DERİNE: Evrişimli Sinir Ağları*, [Online]. Available at: https://ayyucekizrak.medium.com/deri%CC%87ne-daha-deri%CC%87ne-evri%C5%9Fimli-sinir-a%C4%9Flar%C4%B1-2813a2c8b2a9 (Accessed: 17 June 2022).

Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) *ImageNet Classification with Deep Convolutional Neural Networks*, In: Advances in Neural Information Processing Systems 25 (NIPS 2012).

Li, T., Kuo, P., Tsai, T. and Luan, P. (2019) *CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot*, IEEE Access [Online]. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8760246 (Accessed: 24 October 2022).

Livieris, I., Pintelas, E. and Pintelas, P. (2020) *A CNN–LSTM model for gold price time-series forecasting*, Neural Computing and Applications, [Online]. Available at: https://link.springer.com/article/10.1007/s00521-020-04867-x (Accessed: 24 October 2022).

Lu, X. (2022) *Deep Learning Based Emotion Recognition and Visualization of Figural Representation*, Frontiers in Psychology [Online]. Available at: https://www.researchgate.net/publication/357677255_Deep_Learning_Based_Emotion_Recognition_and_Visualization_of_Figural_Representation (Accessed: 24 October 2022).

Manasa, S. B., Abraham, J., Sharma, A., and Himapoornashree, K. S. (2020) *AGE, GENDER AND EMOTION DETECTION USING CNN*, International Journal of Advanced Research in Computer Science [Online]. Available at: https://eds.p.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=bcb53f62-e652-4a90-b009-b7dd2b50f6cb%40redis (Accessed: 24 October 2022).

Mateen, M., Wen, J., Nasrullah, Song, S. and Huang, Z. (2019) *Fundus Image Classification Using VGG-19 Architecture with PCA and SVD*, Symmetry [Online]. Available at: https://www.researchgate.net/publication/329816608_Fundus_image_classification_using_VGG-19_architecture_with_PCA_and_SVD (Accessed: 17 June 2022).

Mehendale, N. (2020) *Facial emotion recognition using convolutional neural networks (FERC),* [Online]. Available at: https://link.springer.com/article/10.1007/s42452-020-2234-1 (Accessed: 17 June 2022).

Mollahosseini, A., Chan, D., Mahoor, M.H., 2016. Going deeper in facial expression recognition using deep neural networks, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Presented at the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Lake Placid, NY, USA, pp. 1–10.

Naga, P., Marri, S. and Borreo, R. (2021) *Facial emotion recognition methods, datasets and technologies: A literature survey,* Materials Today: Proceedings [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S2214785321048987 (Accessed: 24 October 2022).

Nguyen, H., Yeom, S., Lee, G., Yang, H., Na, I. and Kim, S. (2019) *Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks*, International Journal of Pattern Recognition and Artificial Intelligence [Online]. Available at: https://www.researchgate.net/publication/329948265_Facial_Emotion_Recognition_Using_an_Ensemble_of_Multi-Level_Convolutional_Neural_Networks (Accessed: 17 June 2022).

Özkara, C. and Oğuz Ekim, P. (2022) *Real-Time Facial Emotion Recognition for Visualization Systems,* ASYU 2022 (Innovations in Intelligent Systems and Applications Conference), 7-9 September 2022, Antalya, Turkey.

Pandey, R.K., Karmakar, S., Ramakrishnan, A.G. and Saha, N. (2019) *Improving Facial Emotion Recognition Systems Using Gradient and Laplacian Images,* [Online]. Available at: https://arxiv.org/pdf/1902.05411v1.pdf (Accessed: 29 June 2022).

Pranav, E., Kamal, S., Satheesh Chandran, C. and Supriya, M.H. (2020) *Facial Emotion Recognition Using Deep Convolutional Neural Network,* in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). Presented at the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, Coimbatore, India, pp. 317–320.

Sarang Narkhede (2018) *Understanding Confusion Matrix,* [Online] Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (Accessed: 29 June 2022).

Simonyan, K. and Zisserman, A. (2015) *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION,* [Online]. Available at: https://arxiv.org/pdf/1409.1556.pdf%E3%80%82 (Accessed: 29 June 2022).

Smitha, K.G. and Vinod, A.P. (2015) *Facial emotion recognition system for autistic children: a feasible study based on FPGA implementation,* [Online]. Available at: https://link.springer.com/article/10.1007/s11517-015-1346-z (Accessed: 29 June 2022).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) *Going deeper with convolutions,* In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Taghi Zadeh, M.M., Imani, M. and Majidi, B. (2019) *Fast Facial emotion recognition Using Convolutional Neural Networks and Gabor Filters,* in: 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI). Presented at the 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), IEEE, Tehran, Iran, pp. 577–581.

Viola, P. and Jones, M. (2001) *Rapid object detection using a boosted cascade of Simple features*, In: Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. IEEE.

Wang, S.-H., Phillips, P., Dong, Z.-C. and Zhang, Y.-D. (2018) *Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm*, Neurocomputing [Online]. Available at: https://www.researchgate.net/publication/319197681_Intelligent_Facial_Emotion_Recognition_based_on_Stationary_Wavelet_Entropy_and_Jaya_algorithm (Accessed: 24 October 2022).

Zhang, H., Jolfaei, A. and Alazab, M. (2019) *A face emotion recognition method using convolutional neural network and image edge computing*, IEEE Access [Online]. Available at: https://www.researchgate.net/publication/336858850_A_Face_Emotion_Recognition_Method_Using_Convolutional_Neural_Network_and_Image_Edge_Computing (Accessed: 29 June 2022).

Zhu, N., Yu, Z. and Kou, C. (2020) *A New Deep Neural Architecture Search Pipeline for Face Recognition*, IEEE Access [Online]. Available at: https://www.researchgate.net/publication/336858850_A_Face_Emotion_Recognition_Method_Using_Convolutional_Neural_Network_and_Image_Edge_Computing (Accessed: 29 June 2022).