IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 6, December, 2022, pp. 1929-1941
Journal homepage: http://iieta.org/journals/ts

A Comparative Approach for Facial Expression Recognition in Higher Education Using Hybrid-Deep Learning from Students' Facial Images

Muhammed Usame Abdullah*, Ahmet Alkan

Electrical & Electronics Engineering Department, Kahramanmaras Sutcu Imam University, Kahramanmaraş 46100, Turkey

Corresponding Author Email: 15320544581@ogr.ksu.edu.tr

https://doi.org/10.18280/ts.390605

Received: 22 August 2022
Accepted: 12 December 2022

Keywords:
Covid-19, deep learning, facial expression recognition, pre-trained networks, transfer learning

ABSTRACT

Online education has become increasingly common due to the Covid-19 pandemic. A key difference between online and face-to-face education is that instructors often cannot see their students' facial expressions online. This is problematic because facial expressions can help an instructor gauge engagement and understanding. Therefore, it would be useful to find a mechanism whereby students' facial expressions during an online lecture could be monitored. This information can be used as feedback for teachers to change a particular teaching method or maintain another method. This research presents a system that can automatically distinguish students' facial expressions. These comprise eight expressions (anger, attention, disgust, fear, happiness, neutrality, sadness, and surprise). The data for this research was collected from pictures of 70 university students' facial expressions. The data included 6720 images of students' faces distributed equally among the eight expressions mentioned above, that is, 840 images for each category. In this paper, pre-trained deep learning networks (AlexNet, MobileNetV2, GoogleNet, ResNet18, ResNet50, and VGG16) with transfer learning (TL) and K-fold validation (KFCV) were used for recognizing the facial expressions of students. The experiments were conducted using MATLAB 2021a and the best results were recorded by ResNet18 for F1-score and for AUC curve 99%, and 100% respectively.

1. INTRODUCTION

Deep learning-based facial expression recognition has potential applications in social media, marketing, public safety, and human-computer interaction. It is thus regarded as an advanced technique worthy of further development. The aim of this study is to develop techniques that enable facial recognition to be automated for eight mood classes: attention, anger, disgust, fear, happiness, neutrality, sadness, and surprise expressions to monitor students' expressions during online lectures. Most recent research in facial expression recognition has focused on seven mood classes: anger, disgust, fear, happiness, neutrality, sadness, and surprise expressions [1, 2]. Since student engagement is very important in education, "attention" has been added to emotional states in the study. A summary of research related to these techniques is given below.

Hussein et al. [1] relied on geometric aspects of facial expressions with a modest level of complexity in their study. They employed three face image databases: VDMFP, BINED, and MMI. These databases hold images of various facial expressions, captured under various imaging situations, which were first rotated around a central point, then analyzed with deep learning to extract features. Their method achieved 94% for accuracy. Sajjad et al. [2] proposed a method which can provide information on potential suspicious activity using facial expression recognition. They used Raspberry Pi to carry out this purpose, where the video is captured by the Raspberry Pi camera, and faces are detected using the Viola-Jones algorithm. The Gabor filter is also used for the pre-processing in face region and the Oriented Fast and Rotated Brief (ORB) algorithm is used for the feature extraction. SVM is used as a classifier, where SVM classified the facial expressions (anger, disgust, fear, happiness, neutrality, sad, and surprise). By the classification of the facial expressions, the emotion behind the scene was predicted. Based on these predictions, if hostility is detected steps can be taken to avert negative outcomes. They used public databases (CK+, JAFFE, MMI) for testing and training stages and their model achieved 92% for accuracy.

Leo et al. [3] suggested a system that uses a robot to calculate the robot's interaction with children with autism spectrum disorders (ASD). Researchers provide the robot with the facial expression for its imitation, then the robot asks the child to imitate it to assess his or her emotional imitation abilities and quantify the time-delay between the robot's request and the child's response. Faces are cropped and registered, and the HOG descriptor is used to create a data vector that is fed into SVM, resulting in an estimate of the observed facial expression. The researchers achieved 94% for the F1-Score. Syazana-Itqan et al. [4] developed a face recognition system with a MATLAB-based Convolution Neural Network (CNN) and a Graphical User Interface (GUI) as the user interface. To minimize neural network training time, the suggested CNN can accept new data by training the last two layers out of four layers. They used images from the AT&T database and the JAFFE database, with 40 subjects from AT&T and 10 from JAFFE database and achieved 100% for accuracy.

Mayya et al. [5] developed an approach based on a Deep Convolutional Neural Network (DCNN) to automatically

distinguish facial expressions through identification of an individual's facial expression using a single image. Using a JAFFE Dataset for a general-purpose graphic processing unit (GPGPU) significantly reduced the feature extraction time. Images from CK+ Dataset and JAFFE Dataset (Japanese Female Facial Expression) were used for several facial expressions such as anger, disgust, fear, happiness, neutral, sadness, and surprise. This method achieved 96% for F1-score. DNNs based on CNNs were employed by Fathallah et al. [6] for Automated Facial Expression Recognition. To improve results, they use the Visual Geometry Group model (VGG) to fine-tune their architecture. They test their design with a number of large general databases (CK+, MUG, and RAFD) in order to assess it. Their findings demonstrate that the CNN method is relatively effective in image expression identification on various public databases, indicating that facial expression analysis has improved. Their model achieved 93% for accuracy. Sang et al. [7] proposed DCNNs that can effectively interpret semantic information accessible in faces automatically without hand-designing feature descriptors, based on current advances in deep learning. They used the Kaggle dataset's competitor, the FERC-2013 dataset, which contains 35,887 grayscale images with a resolution of 48x48 pixels. These images have been separated into three categories by Kaggle: 28,709 training images, 3589 public test images, and 3589 private test images. Sang et al.'s method achieved 73% for accuracy.

Tarnowski et al. [8] used the KDEF dataset for emotion recognition using facial expression, focusing on seven emotional states (neutral, joy, sadness, surprise, anger, fear, disgust). The features are classified by using K-NN and MLP neural network classifiers and their model achieved 73% for accuracy. Qayyum et al. use JAFFE and CK+ dataset and MS-Kinect dataset in their study. They use Stationary Wavelet Transform (SWT) to extract features for facial expression recognition. In order for the stationary wavelet to transform properly, a combination of horizontal and vertical sub-bands are used, where these sub-bands contain information on muscle movement for the majority of the facial expressions. Their method achieved 94% for accuracy [9].

Lopes et al. [10] propose a simple approach to facial expression recognition with little data. In the method that they proposed, a combination of a CNN and specific image pre-processing steps (rotation correction, cropping, down-sampling, and intensity normalization) is used. The pre-processing steps are conducted using OpenCV and CNN library (CAFÉ). In addition, they used the public databases (CK+, JAFFE, and BU-3DFE) for training and testing stages. Their model achieved 95.75% for accuracy.

Surace et al. [11] suggest a method for group emotion recognition in the wild modelled on a combination of deep CNNs and Bayesian classifiers. They used the TensorFlow library to train CNNs and implemented the proposed approach in Python. Furthermore, OpenCV was used to perform pre-processing operations on their dataset images, where they used the Emotion Recognition in the Wild 2017 (EmotiW17) as a database. Their method achieved 67.75% for accuracy.

Gupta et al. [12] relied on the size of the face and eyes and on deep learning for facial detection. They used OpenCV for the identification of the face, as OpenCV is considered a classifier that helps in image processing. In the later stages of the processing operations, a deep belief network was used. All these operations used the Python program and their model achieved 97% for accuracy. Arya and Agrawal [13] introduced a face recognition approach based on Deep Learning and Linear Discriminant Analysis (LDA), in which typical image recognition system architectures include three primary stages. In the first stage, referred to as pre-processing, the facial images are partitioned into various face components. Prior to the LDA algorithm being applied, the feature extraction process is conducted. In the final step, the neural network is proposed for training on the extracted face classes and features and the model generated is used for face recognition. Li and Lam [14] developed a technique using a Deep Neural Network (DNN) to recognize facial expressions. The Gabor filter is used to extract facial expressions, and then kernel PCA is applied, after which the data extracted from the previous stage is processed into the deep neural network. Their model achieved 91% for accuracy.

Carcagnì et al. [15] used CK+ dataset, for facial expression recognition. One method is proposed, based on histograms of oriented gradients (HOG), where HOG is a descriptor. After obtaining the features of HOG, these are provided to inputs of a group of Support Vector Machines (SVMs), where SVM is a discriminative classifier. They achieved 92.9% for F1-Score.

Chen et al. [16] have proposed a softmax regression autoencoder network (SRDSAN) for facial expression recognition to address the dual challenges of learning efficiency and computational complexity, where a Deep Sparse Autoencoder Network (DSAN) is used to extract high-level features and learn facial emotion features and the facial emotions are classified by using softmax regression (SR). The researchers used (CK+, JAFFE) as databases and the method proposed is to combine with robots in order to analyze and understand human emotions. Their method achieved 89% for accuracy. Rao et al. [17] have proposed a method to combine the features of facial expressions and speech on the Indian Face Database and Berlin Speech Database. They used MFCC (Mel Frequency Cepstral Coefficients) for recognizing speech, and MSER (Maximally Stable Extremal Regions) for facial emotion recognition. Also, SVM classifier was as used for emotion classification and AdaBoost for gender classification. Their model achieved 92% for accuracy. Tonguç and Ozaydın Ozkara [18] analyzed pictures of facial expressions of 67 students during practical and theoretical lectures using software developed through Microsoft Emotion Recognition API and C# programming language. After analysis, they found that the feelings of disgust, sadness, happiness, fear, contempt, anger, and surprise changed during the lecture depending on the way the lecturer presented the lecture. Xu et al. [19] proposed a method for estimating the positive or negative state of focus based on the head position. They used CNN networks with two public databases CK+ & BU-4DFE and their method achieved 91.5% for accuracy. Wu et al. [20] developed a Weight Adapted Convolution Neural Network (WACNN)-based technique for facial expression recognition, where the pre-processing for the facial expression images is conducted via geometry normalization and gray normalizing. Following this, PCA is used to extract the low-level expression feature, then WACNN is used to learn representative features and recognize them, resulting in facial expression information. As databases, they used (CK+, JAFFE, and SFEW2.0) and their method achieved 94% for accuracy. Wang et al. [21] developed a system to recognize speech emotion and facial emotions, they used the CNN and RNN to distinguish the facial expressions, and they used LSTM and CNN to recognize speech emotion, they employed a variety of datasets such as RML, AFEW6.0, and eNTERFACE'05 for achieved those

purposes. Li et al. [22] proposed a method for recognizing facial expressions using a convolution neural network (CNN) to accurately separate the salient regions from a face image. Following that, the Gaussian Markov random field (GMRF) model was improved to improve texture features' ability to represent image information, and a novel feature extraction algorithm called specific angle abundance entropy (SAAE) was designed to improve shape features' ability to represent image information. The texture and shape features were then combined and trained and classified by the support vector machine (SVM) classifier.

Examining the past research that has been summarized, all the research is very recent and valuable, and addresses multiple goals. However, few of these previous studies are focused on enhancing the education in universities. Furthermore, in terms of the applicability of existing research to online education contexts, limitations can be identified. Firstly, most of the previous research focused on recognizing the facial expressions of seven emotions (anger, disgust, fear, happiness, neutrality, sadness, and surprise) without the inclusion of an expression to measure engagement. In addition, most of the research to date has used public data sets from the internet such as CK + and others. This led us to aspire to access new techniques that will enable facial expression recognition to be automated for eight mood classes, rather than seven: anger, disgust, fear, happiness, neutrality, attention, sadness, and surprise expressions. “Attention” was added to the facial expressions because this is important in an educational context as a measure of engagement. A Deep learning-based CNN and image processing techniques were employed.

The database was trained on the pre-trained networks such as Alexnet, GoogleNet, ResNet18, ResNet50, MoblieNetV2, and VGG16. A database for the current research was created from the images of the faces students at one university. These images express the facial expressions of the eight expressions, which are mentioned above.

2. MATERIALS AND DATASET

The facial images used in this study were obtained from the students of Kahramanmaraş Sütçü Imam University, Faculty of Engineering and Architecture after obtaining the necessary approvals from the ethical committee of the university. A database was created from the picture of the faces of 70 students, 60 of whom are male and 10 female. The male-to-female ratio is approximately representative of the wider student cohort in the faculty. The images of seventy students are sufficient to create an adequate data set for our study because 24 photographs were obtained from each student. At the same time, by applying augmentation to these images, the number of facial images in the data set was increased to four times This database includes students from Turkey and from Middle Eastern countries who may have the same facial expressions. The findings of this research could therefore be applicable to a large population across the Middle East. Each dataset contains the eight facial expressions of the students’ faces (anger, disgust, fear, happiness, neutrality, sadness, attention, and surprise). The number of images in our database reached 6720, distributed evenly over the eight categories shown above, where each category has 840 images. Figures 1A. and 1B show a sample of our database. Images of the students’ faces were taken in different lighting environments that simulate real lighting conditions in terms of lack of lighting in certain places or the presence of lighting in other places.

Figure 1A. Samples of eight facial expressions from the dataset for the same student

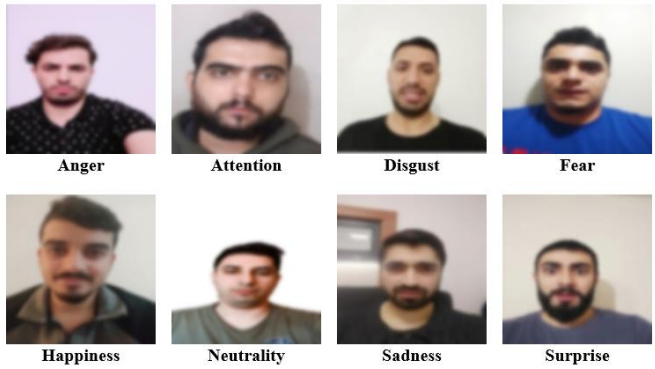

Figure 1B. Samples of eight facial expressions from the dataset from different students

3. COMPUTATIONAL SET-UP

In the current study, image processing, deep learning toolbox, and transfer learning algorithm are used in Matlab 2021A where modifications are made to the last layers in the Convolutional Neural Networks, and then these networks are trained on the new data. If the classification accuracy is sufficient, then the desired accuracy is obtained. However, if the classification accuracy is insufficient, modifications will be made to the Convolutional Neural Networks, and retraining on that data will be conducted until the desired sufficient accuracy is obtained (see Figure 2).

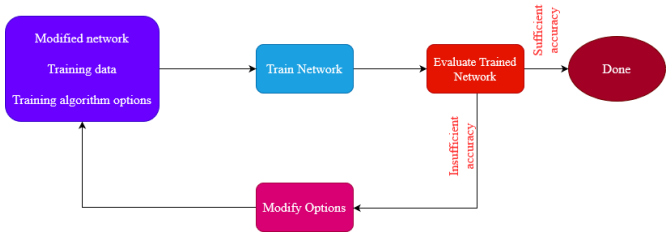

Figure 2. Workflow of transfer learning

A computer with an i7/7th CPU, 16GB RAM, and a 4GB GTX1050 was employed in this study. GPU/CUDA can be used for processing acceleration throughout the training and testing phases thanks to these features. While the NVIDIA CUDA® Deep Neural Network (cuDNN) library is a GPU-

accelerated library of deep neural network primitives, it also includes highly fine-tuned implementations of standard processes including foreground, background, convolution, pooling, normalization, and activation layers (see Figure 3).

4. METHODS

This study used six pre-trained deep convolutional neural networks. Those networks were employed with transfer learning with k-fold cross-validation (AlexNet+TL+KFCV, GoogleNet+TL+KFCV, ResNet18+TL+KFCV, ResNet50+TL+KFCV, MobileNetV2+TL+KFCV, and VGG16+TL+KFCV), where K=5.The findings of these six networks are compared in the results and discussion section regarding overall accuracy, ROC-AUC curves, and other performance measures.

4.1 AlexNet

In 2012, Alex Krizhevsky designed the network known as AlexNet and trained it on the ImageNet dataset to classify 1.2 million high-resolution images in 1000 various classes, for which he won an award: the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [23]. AlexNet is a Deep Convolutional Neural Network (DCNN) that contains eight basic layers 5 convolutional and 3 fully connected (see Figure 4).

The transfer learning algorithm can be used in this kind of Deep Convolutional Neural Network and other Deep Convolutional Neural Networks. The reason for using the transfer learning algorithm is that small or medium data can be used to train the network, thus saving training time, while training from scratch necessitates significant quantities of data to train the network and therefore takes a long time to accomplish [24] (see Figure 5).

4.2 GoogleNet

In 2014 Google researchers designed a DCNN which consists of 18 layers and is capable of sorting images into 1000 object categories. For this achievement, the researchers won the ImageNet challenge. GoogleNet's basic convolutional block is called an Inception block and GoogleNet uses a stack of a total of 9 inception blocks alongside global average pooling to generate its estimates. The dimensionality is reduced by maximum pooling between inception blocks [25] (See Figure 6).

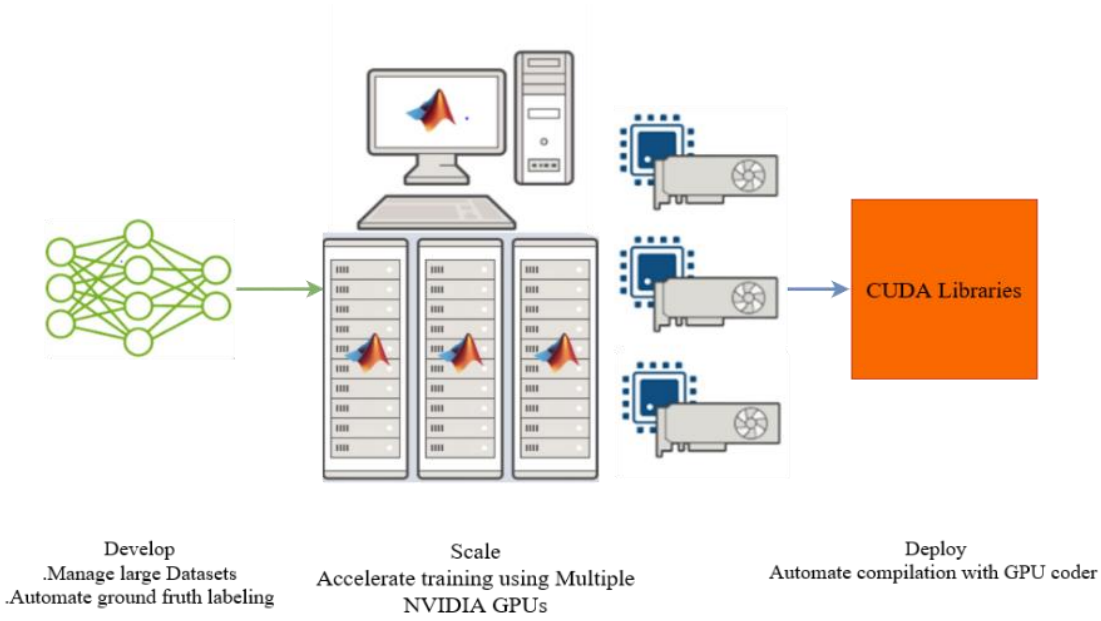

Figure 3. GPU pathway

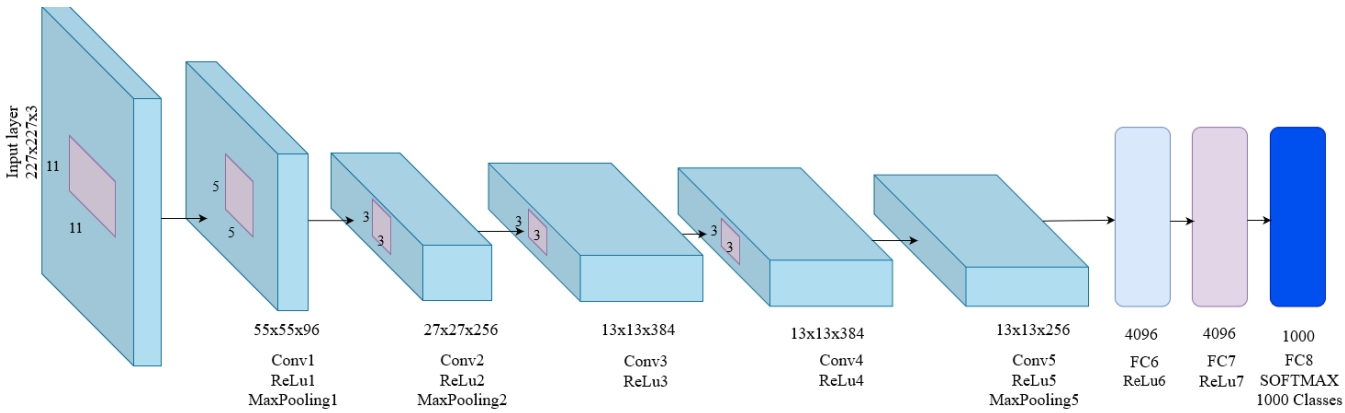

Figure 4. Architecture of AlexNet

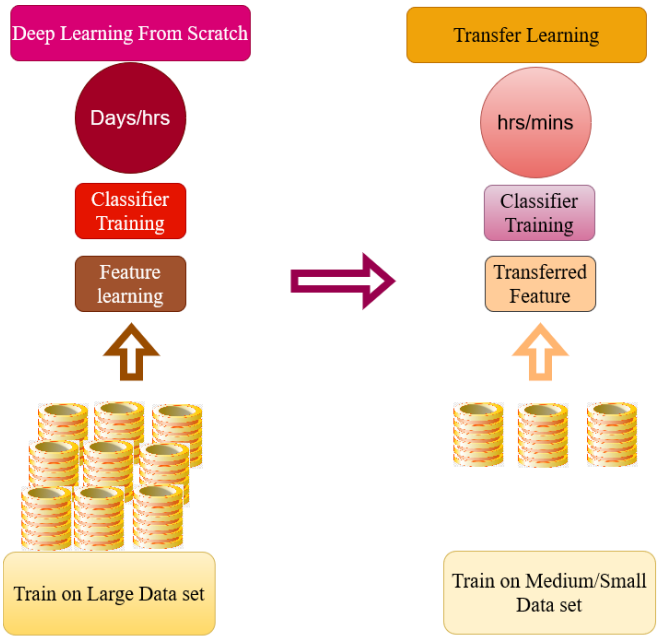

Figure 5. Comparison of network built from scratch (left) vs by transfer learning (right)

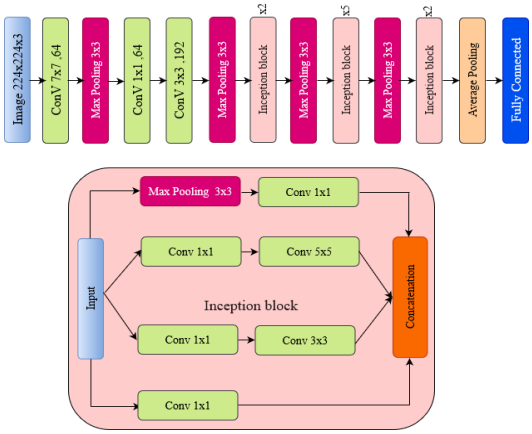

Figure 6. Architecture of GoogleNet

4.3 ResNet18

ResNet-18, the most basic model of deep residual networks, contains 18 layers and a 224x224x3 input image size. This pre-trained network is a CNN capable of classifying images into 1000 object categories after being trained on over a million images. It was designed by [26] who were the winners of the ILSVRC-2015 competition. In ResNet18, the residual block is a stack of layers configured so that the output of one layer is added to another layer deeper in the block. The nonlinearity is then applied by combining it with the output of the relevant layer in the main path. The overall design of ResNet-18 is shown in Figure 7.

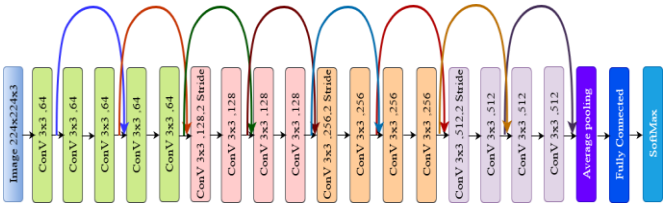

Figure 7. Architecture of ResNet18

4.4 ResNet50

A type of deep residual network (ResNet) that stacks residual blocks on top of each other to form a network, ResNet50 consists of 48 Convolution layers, 1 MaxPool, and 1 Average Pool layer. Figure 8 illustrates two modules of skips; the first is an identity block that has no convolution layer at the skips, the second is a convolution block that has a convolution layer at the skips. This technique is known as bottleneck design, and it decreases the number of parameters while maintaining network performance [26].

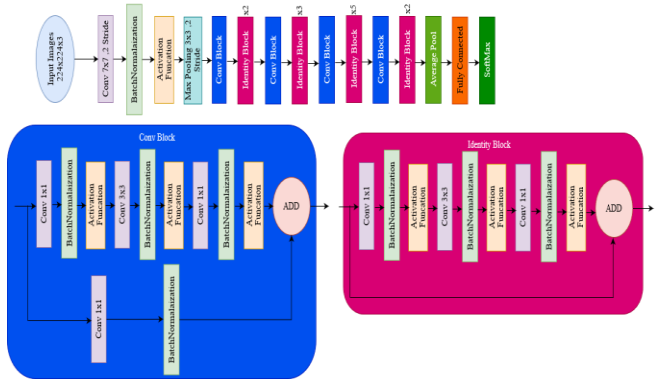

Figure 8. Architecture of ResNet50

4.5 MobileNetV2

MobileNetV2 is a 53-layer DCNN architecture designed to operate on smartphones and other low computational power devices. It employs an inverted residual structure, with residual connections forming between the bottleneck's layers. The intermediate expansion layer filters feature deep lightweight depthwise convolutions. This pre-trained network can classify images into 1000 object classes thanks to the MobileNetV2 design, which has an initial fully convolutional layer with 32 filters, followed by 19 remaining bottleneck layers. The network's image input size is 224 by 224 pixels. Figure 9 depicts the MobileNetV2 architecture [27].

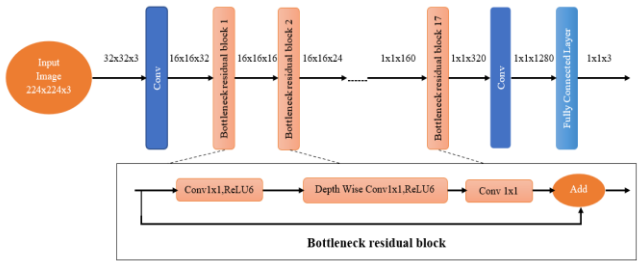

Figure 9. Architecture of MobileNetV2

4.6 VGG16

The Visual Geometry Group (VGG) Deep Convolutional Neural Network was proposed by Simonyan and Zisserman. Their model placed second in the ILSVRC-2014 competition. This network has 13 convolutional layers and three fully connected layers with five max-pooling layers in total. The first block has 64 filters, which are doubled in consecutive blocks until there are 512 total filters. This model is completed with two hidden layers which are fully connected and contain a total of 4096 neurons, and one output layer. This network

can classify photos into 1000 object types because the output layer has 1000 neurons. The overall design of VGG16 is shown in Figure 10 [28].

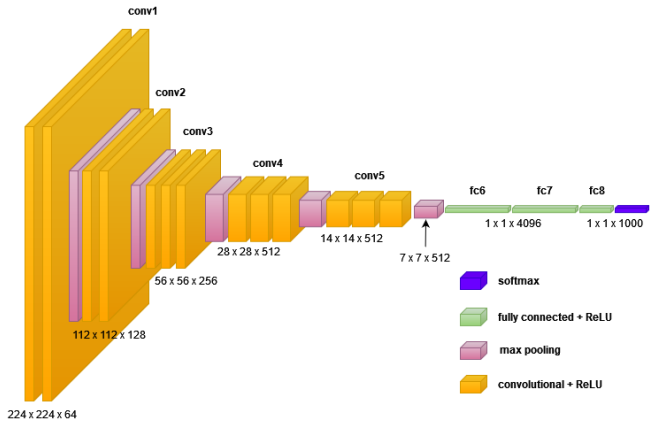

Figure 10. Architecture of VGG16

4.7 K-fold cross-validation

In classification problems, our data set is first separated into training and test sets and then a model is built on the training data set and ~~test~~ the predictions made are tested on the test data set. However, there may be some problems with the train/test separation. For example, it may not be possible to randomly separate the data set or only some data sets may have been chosen for model building, which will cause an overfitting problem. This problem can be solved with cross-validation [29]. Cross-validation will show whether the model's high performance is random. In K-Folds Cross-Validation, the data is divided into subgroups and k-1 subsets are used to train the data, leaving the final subset as test data. The mean value of the error obtained because of k experiments indicates the correctness of our model [30]. The dataset is divided into K number of sections/folds where each fold is used as a test set at the number of points. The current study took a fivefold validation scenario (K=5). Here, the data set is divided into five folds. In the first stage, the first fold is used to test the model and the remainder are used to train the model. In the second stage, the second fold is used as the test set while the others are used as the training sets. This process is repeated until each of the five folds has been used as the data set for the test (see Figure 11).

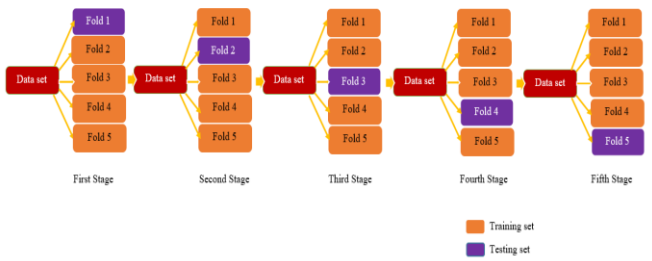

Figure 11. k-fold cross-validation

4.8 Algorithms proposed in this study

Figure 12 illustrates how k-fold cross-validation and transfer learning with CNN networks were used in this research.

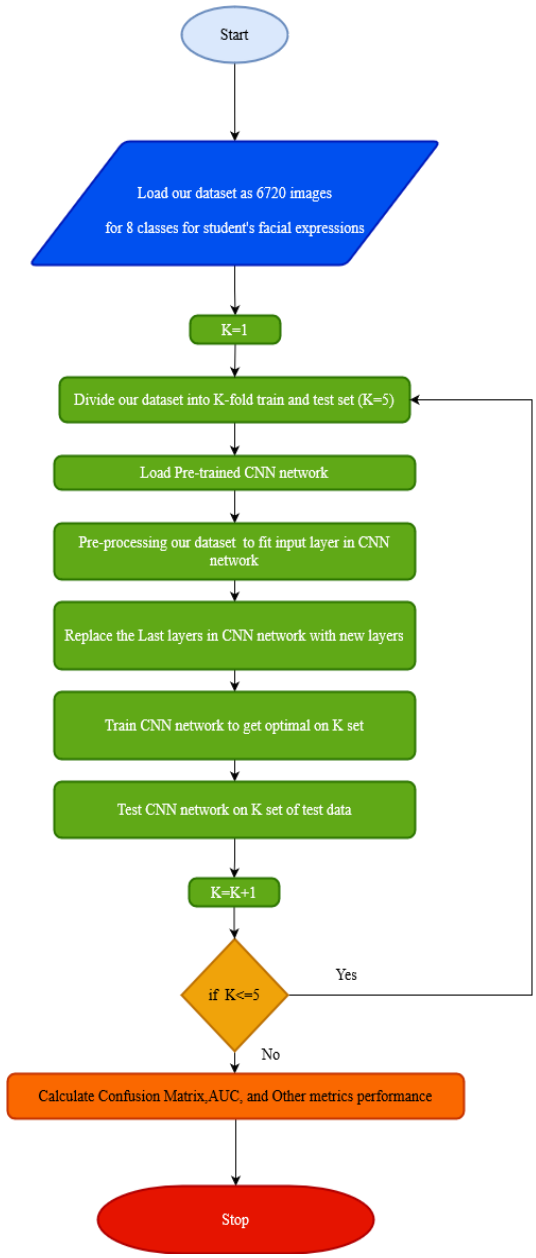

Figure 12. Flowchart for the algorithm proposed in the current study for each CNN network

The proposed algorithm can be explained by the following steps:

First step: Loading our dataset, consisting of 6720 images expressing the eight facial expressions of the students' faces.

Second step: the value of the initial counter is K=1.

Third step: Dividing the dataset into five-fold for training and testing (80% of the dataset for training and 20% for testing).

Fourth step: Loading the pre-trained CNN network.

Fifth step: The pre-processing stage in order for this data to become suitable for the input of the Deep Convolutional Neural Network (resizing images).

Sixth step: Replacing the last layers in the CNN network with new layers in this process by using transfer learning.

Seventh step: Training the CNN network on the training dataset.

Eighth step: Testing the CNN network on the Testing dataset.

Ninth step: Increasing the value of the counter by one.

Tenth step: Comparing the new value for the counter with 5

if the new value is smaller or equal to 5 repeats the steps from 3 to 8.

Else (when the counter value is greater than five), exit the loop and calculate the performance metrics (which is the average rate of the five times. Confusion matrices and ROC-AUC curves were also inferred).

5. RESULTS AND DISCUSSION

Our dataset has been applied to six pre-trained CNN networks and transfer learning and K-cross validation. Those networks are AlexNet + TL + KFCV, GoogleNet + TL + KFCV, ResNet 18 + TL + KFCV, ResNet50+TL + KFCV, MobileNetV2+ TL + KFCV, and VGG16+TL + KFCV. The transfer learning (TL) + 5-fold validation process was applied to the previous six trained deep learning architectures (AlexNet, GoogleNet, ResNet 18, ResNet 50, MobileNetV2, and VGG16). The features extracted from the training images were used to train the models to obtain the better parameterization of these architectures. By using the transfer learning algorithm, the six networks are trained on our data and thus have new parameters, and the old parameters are removed. Our dataset consists of images of the faces of university students and the number of images is 6720 pictures distributed over eight categories of facial expressions (anger, disgust, fear, happiness, neutrality, sadness, attention, and surprise) for each category of 840 images. The practical experiments were performed separately on the six networks. Table 1 shows the training and testing time for each network. It can be noted that the time of the training and testing is related to the maximum number of parameters that are distinguished by each network from the other (Table 2).

Table 1. Time of the stages of the training and testing for various CNN Networks using GPU

CNN Network+TL+KFCV (K=5)	Time for training and testing (Minutes)
AlexNet+TL+KFCV	≈147
GoogleNet+TL+KFCV	≈110
ResNet18+TL+KFCV	≈105
ResNet50+TL+KFCV	≈300
MobileNetV2+TL+KFCV	≈250
VGG16+TL+KFCV	≈660

Table 2. The maximum number of parameters for various CNN networks

CNN Network	Max Size of Data	Parameters (Millions)	Image input Size	Depth of layers *
AlexNet [22]	227-by-227	61.0	227 MB	8
GoogleNet [23]	224-by-224	7.0	27 MB	22
ResNet18 [24]	224-by-224	11.7	44 MB	18
ResNet50 [24]	224-by-224	25.6	96 MB	50
MobileNetV2 [25]	224-by-224	3.5	13 MB	53
VGG16 [26]	224-by-224	138	515 MB	16

*On a path from the input layer to the output layer, the network depth is defined as the greatest number of sequential convolutional or fully connected layers.

5.1 Performance metrics

It can be inferred that many values help in evaluating the performance of the model since these values are inferred from the confusion matrix which is the cornerstone of evaluating a classification model (that is, classifier). This square matrix displays the performance of a learning algorithm by reporting the counts of the True positive (TP), True negative (TN), False positive (FP), and False negative (FN) predictions of a classifier. In the current study, confusion matrices for the models are shown (AlexNet + TL + KFCV, GoogleNet + TL + KFCV, ResNet 18 + TL + KFCV, ResNet50+TL + KFCV, MobileNetV2+ TL + KFCV,and VGG16+TL + KFCV) on which our datasets were trained (see Figures from 13 to 18).

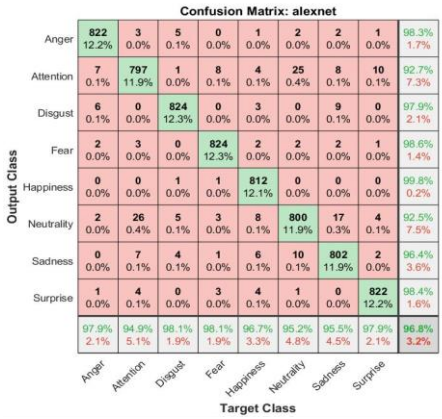

Figure 13. Confusion Matrix for AlexNet+TL+KFCV

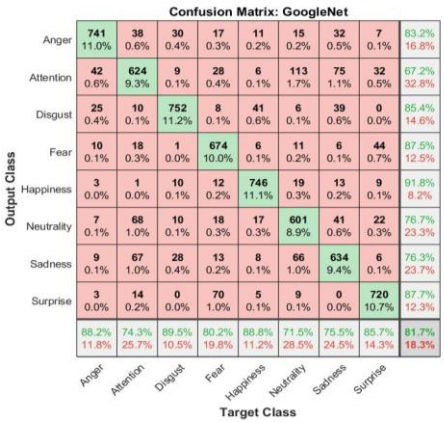

Figure 14. Confusion Matrix for GoogleNet+TL+KFCV

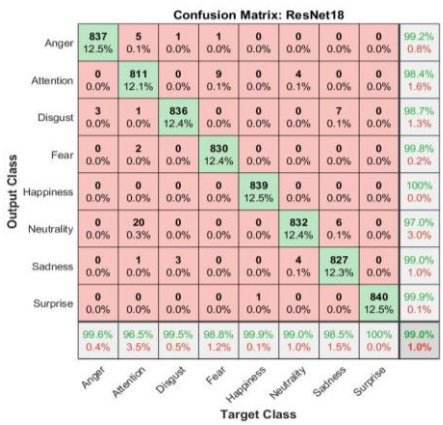

Figure 15. Confusion Matrix for ResNet18+TL+KFCV

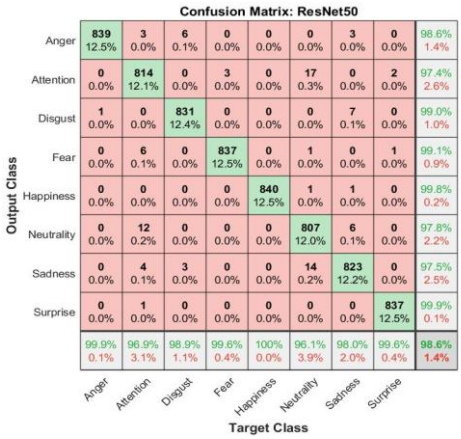

Figure 16. Confusion Matrix for ResNet50+TL+KFCV

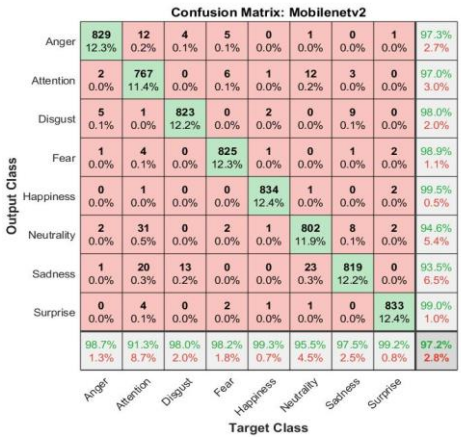

Figure 17. Confusion Matrix for MobileNetV2+TL+KFCV

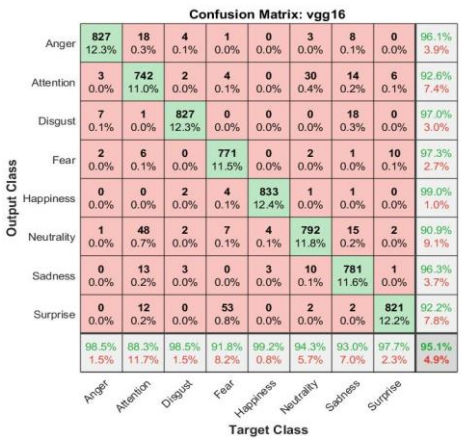

Figure 18. Confusion Matrix for VGG16+TL+KFCV

From the confusion matrix, several parameters of performance metrics can be deduced which are given as follows:

1. Accuracy (ACC) measures how many times a classifier makes an accurate prediction, and is calculated using the following equation:

$$ACC= (TP+TN)/(TP+TN+FP+FN)$$

2. Recall (REC) is also called sensitivity (SNS) or true positive rate (TPR). It estimates the proportions of true positives from all the positive values observed for a target:

$$SNS=TP/(TP+FN)$$

3. Precision (PRC) estimates the proportions of true positives that were precisely defined, and is calculated as following:

$$PRC=TP/(TP+FP)$$

4. Specificity (SPC) estimates the proportions of true negatives from all negative values observed for a target, also known as true negative rate (TNR).

$$SPC=TN/(TN+FP)$$

5. F1 Score shows the harmonic mean of Precision and Recall values. F1 score is also a sound measure to estimate an imbalanced classifier.

$$F1\text{-}Score=2xPRCxSNS/(PRC+SNS)$$

Now, based on the confusion matrices, it is possible to deduce performance metrics (ACC, SNS, PRC, SPC, F1-Score) for the six networks that are used in this research (AlexNet + TL + KFCV, GoogleNet + TL + KFCV, ResNet 18 + TL + KFCV, ResNet50+TL + KFCV, MobileNetV2+ TL + KFCV, and VGG16+TL + KFCV). Tables have been organized showing the values of performance metrics Tables 3-8).

In order to compare the results between the six studied pre-trained networks, the average values of the evaluation parameters for each network were separated. These values can be represented in a chart (Figure 19) in order to facilitate the comparison of the results.

Table 3. The values of the performance metrics for AlexNet

AlexNet	ACC	SNS	PRC	SPC	F1_score
Anger	0.995	0.983	0.978	0.996	0.980
Attention	0.984	0.926	0.948	0.992	0.937
Disgust	0.994	0.978	0.980	0.997	0.979
Fear	0.995	0.985	0.980	0.997	0.983
Happiness	0.995	0.997	0.966	0.995	0.981
Neutrality	0.984	0.924	0.952	0.993	0.938
Sadness	0.989	0.963	0.954	0.993	0.959
Surprise	0.995	0.984	0.978	0.996	0.981
Average	0.991	0.968	0.967	0.995	0.967

Table 4. The values of the performance metrics for GoogleNet

GoogleNet	ACC	SNS	PRC	SPC	F1_score
Anger	0.962	0.831	0.882	0.983	0.856
Attention	0.922	0.671	0.742	0.962	0.705
Disgust	0.967	0.853	0.895	0.984	0.873
Fear	0.961	0.875	0.802	0.972	0.837
Happiness	0.976	0.917	0.888	0.984	0.902
Neutrality	0.937	0.766	0.715	0.959	0.740
Sadness	0.940	0.762	0.754	0.965	0.758
Surprise	0.967	0.876	0.857	0.979	0.866
Average	0.954	0.819	0.817	0.973	0.817

Table 5. The values of the performance metrics for ResNet18

ResNet18	ACC	SNS	PRC	SPC	F1_score
Anger	0.998	0.991	0.996	0.999	0.994
Attention	0.993	0.984	0.965	0.995	0.974
Disgust	0.997	0.987	0.995	0.999	0.991
Fear	0.998	0.997596	0.988	0.998	0.992
Happiness	0.999	1	0.998	0.999	0.999
Neutrality	0.994	0.969	0.990	0.998	0.979
Sadness	0.996	0.990	0.984	0.997	0.987
Surprise	0.999	0.998	1	1	0.999
Average	0.997	0.989	0.989	0.998	0.989

Table 6. The values of the performance metrics for ResNet50

ResNet50	ACC	SNS	PRC	SPC	F1_score
Anger	0.998	0.985	0.998	0.999	0.992
Attention	0.992	0.973	0.969	0.995	0.971
Disgust	0.997	0.990	0.989	0.998	0.989
Fear	0.998	0.990	0.996	0.999	0.993
Happiness	0.999	0.997	1	1	0.998
Neutrality	0.992	0.978	0.960	0.994	0.969
Sadness	0.994	0.975	0.979	0.997	0.977
Surprise	0.999	0.998	0.996	0.999	0.997
Average	0.996	0.986	0.986	0.998	0.986

Table 7. The values of the performance metrics for MobileNetV2

MobileNetV2	ACC	SNS	PRC	SPC	F1_score
Anger	0.994	0.973	0.986	0.998	0.979
Attention	0.985	0.969	0.913	0.987	0.940
Disgust	0.994	0.979	0.979	0.997	0.979
Fear	0.996	0.989	0.982	0.997	0.985
Happiness	0.998	0.995	0.992	0.998	0.994
Neutrality	0.987	0.945	0.954	0.993	0.950
Sadness	0.988	0.934	0.975	0.996	0.954
Surprise	0.997	0.990	0.991	0.998	0.991
Average	0.993	0.972	0.972	0.996	0.971

Table 8. The values of the performance metrics for VGG16

VGG16	ACC	SNS	PRC	SPC	F1_score
Anger	0.993	0.960	0.984	0.997	0.972
Attention	0.976	0.926	0.883	0.983	0.904
Disgust	0.994	0.969	0.984	0.997	0.976
Fear	0.986	0.973	0.917	0.988	0.944
Happiness	0.997	0.990	0.991	0.998	0.991
Neutrality	0.981	0.909	0.942	0.991	0.925
Sadness	0.986	0.963	0.929	0.990	0.946
Surprise	0.986	0.922	0.977	0.996	0.949
Average	0.987	0.951	0.951	0.993	0.951

Because of their importance, AUC-ROC curves are included in this research (Figure 20). The AUC-ROC curve helps visualization of the performance of the CNN networks classifiers used in this research, and the ROC curve summarizes the performance of the classifier across all possible thresholds. The graph of the ROC curve is plotted with the true positive rate (TPR) in the y-axis and the false positive rate (FPR) in the x-axis for all possible thresholds.

AUC: AUC is the area under the ROC curve. If the classifier is great, the true positive rate will increase, and the area under the curve will be close to 1.

By examining the confusion matrices (Figure 13 to 18, Tables 3 to 8, and Figures 19 and 20), it is clear that the best results were recorded by ResNet18 closely followed by ResNet50. The average values of performance metrics (ACC, SNS, PRC, SPC,F1-Score) for ResNet18+TL+KFCV are 0.997, 0.989, 0.989, 0.998, 0.989 respectively while for ResNet50+TL+KFCV are 0.996,0.986,0.986,0.998,0.986 respectively. Next came the results for MobileNetV2 + TL + KFCV, AlexNet + TL + KFCV, and VGG16 + TL + KFCV, respectively, while the lowest results were recorded by GoogleNet+TL+KFCV.

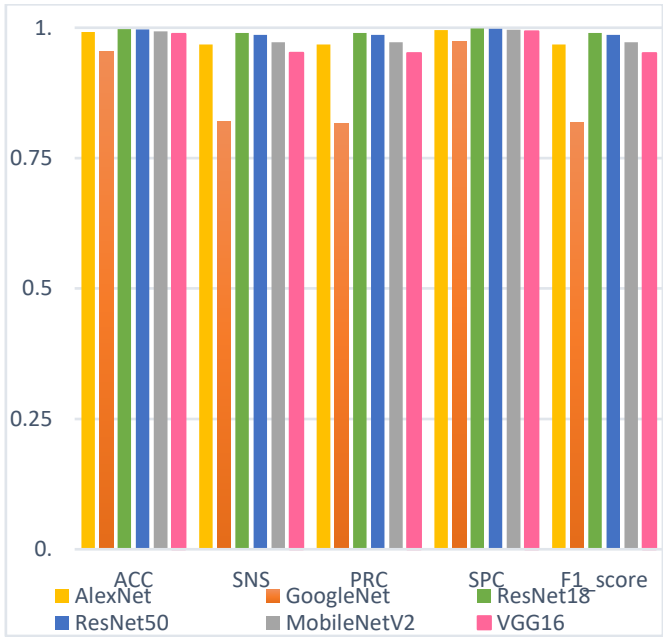

Figure 19. The comparison of averages the performance metrics for six deep learning methods that were used in this study

In addition, by examining ROC curves and AUC values (Figure 20) the best results were recorded when the pre-trained network of ResNet18+TL+KFCV was used, and the same result was obtained for ResNet50+TL+KFCV. This means that they scored the best results for automatic recognition, of the facial expressions of anger, disgust, fear, happiness, neutrality, sadness, attention, and surprise.

From the tables AlexNet, MobileNetV2, and VGG16 (Tables 3, 7 and 8), it can be noted that values of F1-Score for facial expressions of anger, disgust, fear, happiness, and surprise were high. However, for the facial expressions of neutrality, attention, and sadness, they decreased by approximately 5% to 7% from the highest value recorded.

As for GoogleNet (Table 4), it can be noted that values of F1-Score for facial expressions of anger, disgust, fear, happiness, and surprise were approximately 85%, but for facial expressions of neutrality, attention, and sadness, they decreased by approximately 20% from the highest value recorded. If we look at the values of F1-Score (Tables 5 and 6) for ResNet18 and ResNet50, the scores for facial expressions of anger, disgust, fear, happiness, and surprise were very high, but they decreased by approximately 2% for facial expressions of neutrality, attention, and sadness.

It is possible that the facial expressions of neutrality, attention, and sadness may differ from one person to another, and it may be difficult to recognize the state of attention and the neutrality state of a person, so the low F1- Score values for these facial expressions can be noted.

From all the above, it is evident that the best results were recorded by ResNet18 and ResNet50. By comparing ResNet18 and ResNet50 for the time of the stages of the

training and testing examination It can be seen that less time was recorded by ResNet18 (see Table 1). It may be the recommended approach that automatically tracking the students' emotional states throughout the lecture, gives the lecturer immediate feedback and helps to improve the educational experience.

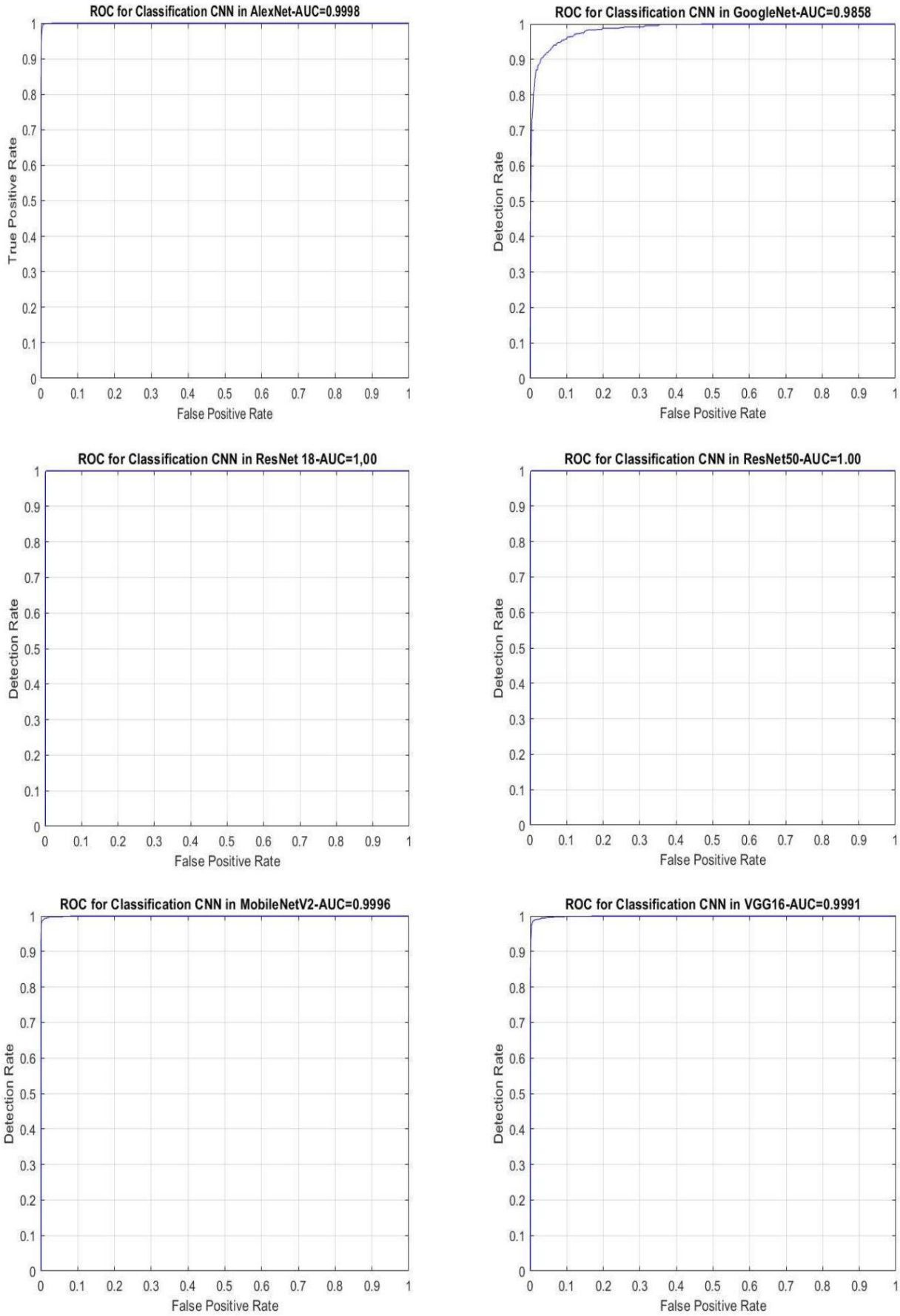

Figure 20. ROC-AUC Curves for classification in the six CNN models

5.2 Comparison of the current study with other studies

In this study, the database of our own was created from the images of the faces of 70 students from the faculty of Engineering expressing eight facial expressions (anger, disgust, fear, happiness, neutrality, sadness, attention, and surprise). The number of images of the students' faces was 6720, distributed equally over the eight facial expressions mentioned above, that is, for each category 840 images. Most of the existing studies used public databases and did not have private data such as that used in our research. Most of the existing studies focused on facial expressions in six or seven categories, while our study focused on facial expressions in eight categories. In addition, deep learning in our research was used with transfer learning with k-fold cross-validation. The results obtained were compared with similar studies on facial expression recognition and a summary of this comparison is shown in Table 9. From the table, it is clear that the results of the current study give high accuracy and AUC values. Furthermore, the obtained results in the current study for the performance metrics (Sensitivity, Specificity, Precision, and F1-Score value) are high. From this comparison, it can be seen that this study can contribute to the field of recognition of facial expressions for students for educational purposes.

Table 9. Comparison of the current study with other studies which used various deep learning methods for facial expression recognition

Study	Facial expressions	Type of data set	Method used	Overall accuracy	Average of Sensitivity (Recall)	Average of Precision	Average of Specificity	Average of F1-Score
[6]	Angry Disgust Happy Neutral Sad Surprise	CK+, MUG, and RAFD (Public dataset)	CNN VGG	93.33%	-	-	-	-
[7]	Angry Disgust Fear Happy Sad Surprise Neutral	FERC-2013(Public dataset)	CNN VGG	71.9%	-	-	-	-
[9]	Angry Disgust Fear Happy Sad Surprise	JAFFE and CK+ (Public dataset)	ANN +Stationary Wavelet Transform Features	98.8% 96.6%	-	-	-	-
[15]	Angry Disgust Fear Happy Sad Surprise Neutral	CK+, JAFFE, MMI	ORB+SVM	99%	-	-	-	-
[16]	Angry Disgust Fear Happy Sad Surprise Neutral	JAFFE and CK+ (Public dataset)	Softmax regression-based deep sparse autoencoder network (SRDSAN)	89.51%	-	-	-	-
[17]	Angry Disgust Fear Happy Sad Surprise	Indian Face Database (Public dataset)	MSER (Maximally Stable Extremal Regions) +SVM+Adaboost	78%	-	-	-	-
[19]	Angry Disgust Fear Happy Sad Surprise	CK+ & BU-4DFE (Public dataset)	VGG16	91.5%	-	-	-	-
[20]	Angry Disgust Fear Happy Sad Surprise Neutral	CK+, JAFFE, SFEW2.0 (Public dataset)	Weight-Adapted Convolution Neural Network (WACNN)	94%	-	-	-	-
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	AlexNet+TL+KFCV	99%	97.1%	97.7%	99.5%	96.8%
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	GoogleNet+TL+KFCV	95.4%	81.9%	81.7%	97.4%	81.8%

Study	Facial expressions	Type of data set	Method used	Overall accuracy	Average of Sensitivity (Recall)	Average of Precision	Average of Specificity	Average of F1-Score
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	ResNet18 +TL+KF CV	99.8%	99%	99%	99.9%	99%
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	ResNet50 +TL+KF CV	99.7%	98.6%	98.6%	99.8%	98.6%
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	MobilNet V2+TL+K FCV	99.3%	97.2%	97.2%	99.6%	97.2%
This study	Happiness Sadness Anger Fear Disgust Attention Neutrality Surprise	Our dataset	VGG16+ TL+KFC V	98.8%	95.2%	95.2%	99.3%	95.1%

6. CONCLUSION AND FUTURE WORK

The facial expression recognition system using deep learning is one of the most prominent areas of artificial intelligence. An experimental system based on deep learning to automatically recognize students' facial expressions (anger, disgust, fear, happiness, neutrality, sadness, attention, and surprise) has been developed employing six various deep learning networks+TL+KFCV. The highest results were obtained by ResNet18+TL+KFCV with the rate of 99.7% accuracy. The findings of this research have important implications for online education because most educational institutions in different countries of the world, tend to make this type of education an alternative option to face-to-face education for students. This system can automatically and rapidly recognize students' facial expressions, making our proposed system useful for lecturers and researchers in the field. Students' facial expressions in an educational context can give instant feedback on the students' experiences in lectures. Techniques that allow for automated online facial recognition can have important practical applications in online education contexts. Such techniques could be used by lecturers to gather feedback about engagement and understanding at a class -level. This feedback could be used by lecturers to adapt their teaching at the local level. At a larger scale, information on students' facial expressions could be used by higher education ministries, university administrators and decision makers in higher education to inform planning at a more strategic level. The current research has some limitations because of the hardware available, there were certain constraints in processing the data; however, when compared to other studies, the acquired findings for the performance metrics (Sensitivity, Specificity, Precision, F1-Score, accuracy, and AUC) are excellent.

For future research, we aim to include the facial expressions of all university students in the institution, to help provide feedback to the lecturers about the students' situations during the lecture, and we hope to work on integrating facial expressions recognition with online learning educational platforms.

ACKNOWLEDGMENT

This study was supported by the individual research project entitled Detection of Emotional States from Facial Expressions of University Students Using Deep Learning and Image Processing Techniques (Derin Öğrenme ve Görüntü İşleme Teknikleri Kullanarak Üniversite Öğrencilerinin Yüz İfadelerinden Duygu Durumlarının Tespiti), BAP Project (2022/4-25M) of Kahramanmaraş Sütçü Imam University.

Ethics Committee Permission for the data set used in the study was obtained from Kahramanmaraş Sütçü Imam University Science and Engineering Sciences Ethics Committee Decision (decision no. 2019/07, dated 15.06.2021).

REFERENCES

[1] Hussein, H., Angelini, F., Naqvi, M., Chambers, J.A. (2018). Deep-learning based facial expression recognition system evaluated on three spontaneous databases. In 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC), pp. 270-275. https://doi.org/10.1109/isivc.2018.8709224

[2] Sajjad, M., Nasir, M., Ullah, F.U.M., Muhammad, K., Sangaiah, A.K., Baik, S.W. (2019). Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. Information Sciences, 479: 416-431. https://doi.org/10.1016/j.ins.2018.07.027

[3] Leo, M., Del Coco, M., Carcagni, P., Distante, C., Bernava, M., Pioggia, G., Palestra, G. (2015). Automatic emotion recognition in robot-children interaction for ASD treatment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 145-153. https://doi.org/10.1109/iccvw.2015.76

[4] Syazana-Itqan, K., Syafeeza, A.R., Saad, N.M. (2016). A MATLAB-based convolutional neural network approach for face recognition system. J. Bioinf. Proteomics Rev, 2(1): 1-5. https://doi.org/10.15436/2381-0793.16.009

[5] Mayya, V., Pai, R.M., Pai, M.M. (2016). Automatic facial expression recognition using DCNN. Procedia

Computer Science, 93: 453-461. https://doi.org/10.1016/j.procs.2016.07.233
[6] Fathallah, A., Abdi, L., Douik, A. (2017). Facial expression recognition via deep learning. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 745-750. https://doi.org/10.1109/aiccsa.2017.124
[7] Sang, D.V., Van Dat, N. (2017). Facial expression recognition using deep convolutional neural networks. In 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 130-135. https://doi.org/10.1109/kse.2017.8119447
[8] Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J. (2017). Emotion recognition using facial expressions. Procedia Computer Science, 108: 1175-1184. https://doi.org/10.1016/j.procs.2017.05.025
[9] Qayyum, H., Majid, M., Anwar, S.M., Khan, B. (2017). Facial expression recognition using stationary wavelet transform features. Mathematical Problems in Engineering, 2017: 1-9. https://doi.org/10.1155/2017/9854050
[10] Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognition, 61: 610-628. https://doi.org/10.1016/j.patcog.2016.07.026
[11] Surace, L., Patacchiola, M., Battini Sönmez, E., Spataro, W., Cangelosi, A. (2017). Emotion recognition in the wild using deep neural networks and Bayesian classifiers. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 593-597. https://doi.org/10.1145/3136755.3143015
[12] Gupta, P., Saxena, N., Sharma, M., Tripathi, J. (2018). Deep neural network for human face recognition. International Journal of Engineering and Manufacturing (IJEM), 8(1): 63-71. https://doi.org/10.5815/ijem.2018.01.06
[13] Arya, S., Agrawal, A. (2018). Face recognition with partial face recognition and convolutional neural network. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 7: 91-94.
[14] Li, J., Lam, E.Y. (2015). Facial expression recognition using deep neural networks. In 2015 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1-6. https://doi.org/10.1109/ist.2015.7294547
[15] Carcagnì, P., Del Coco, M., Leo, M., Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus, 4(1): 1-25. https://doi.org/10.1186/s40064-015-1427-3
[16] Chen, L., Zhou, M., Su, W., Wu, M., She, J., Hirota, K. (2018). Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Information Sciences, 428: 49-61. https://doi.org/10.1016/j.ins.2017.10.044
[17] Rao, K.P., Rao, M.C.S., Chowdary, N.H. (2019). An integrated approach to emotion recognition and gender classification. Journal of Visual Communication and Image Representation, 60: 339-345. https://doi.org/10.1016/j.jvcir.2019.03.002
[18] Tonguç, G., Ozaydın Ozkara, B. (2020). Automatic recognition of student emotions from facial expressions during a lecture. Computers & Education, 148: 103797. https://doi.org/10.1016/j.compedu.2019.103797
[19] Xu, R., Chen, J., Han, J., Tan, L., Xu, L. (2020). Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. Computing, 102(3): 765-780. https://doi.org/10.1007/s00607-019-00722-7
[20] Wu, M., Su, W., Chen, L., Liu, Z., Cao, W., Hirota, K. (2019). Weight-adapted convolution neural network for facial expression recognition in human–robot interaction. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(3): 1473-1484. https://doi.org/10.1109/tsmc.2019.2897330
[21] Wang, X., Chen, X., Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication, 84: 115831. https://doi.org/10.1016/j.image.2020.115831
[22] Li, A.H., An, L., Che, Z.H. (2020). A Facial expression recognition model based on texture and shape features. Traitement du Signal, 37(4): 627-632. https://doi.org/10.18280/ts.370411
[23] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386
[24] Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. Proceedings of ICML Workshop on Unsupervised and Transfer Learning, in Proceedings of Machine Learning Research. 27: 17-36 https://proceedings.mlr.press/v27/bengio12a.html
[25] Guo, Z., Chen, Q., Wu, G., Xu, Y., Shibasaki, R., Shao, X. (2017). Village building identification based on ensemble convolutional neural networks. Sensors, 17(11): 2487. https://doi.org/10.3390/s17112487
[26] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. https://doi.org/10.1109/cvpr.2016.90
[27] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. https://doi.org/10.1109/cvpr.2018.00474
[28] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint. https://doi.org/10.48550/arXiv.1409.1556.
[29] Hussain, L., Huang, P., Nguyen, T., Lone, K.J., Ali, A., Khan, M.S., Duong, T.Q. (2021). Machine learning classification of texture features of MRI breast tumor and peri-tumor of combined pre-and early treatment predicts pathologic complete response. BioMedical Engineering OnLine, 20(1): 1-23. https://doi.org/10.1186/s12938-021-00899-z
[30] Yadav, S., Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In 2016 IEEE 6th International conference on advanced computing (IACC), pp. 78-83. https://doi.org/10.1109/iacc.2016.25