Facial Expression Recognition and Emotion Detection with CNN methods And SVM Classifiers

Received: 2 January 2023; Accepted: 14 March 2023 ***Research Article***

Nibras Farooq Alkhaleeli
Department of Computer Engineering
Kocaeli University
Kocaeli, Türkiye
nibrasspro@gmail.com

Yaşar Becerikli
Department of Computer Engineering
Kocaeli University
Kocaeli, Türkiye
ybecerikli@kocaeli.edu.tr
0000-0002-2951-7287

Abstract— **There are different humans in our life. With the different languages and cultures of the human, the involuntary methods of facial and body expression remain the most realistic and honest ways. In this study, we will interpret people's emotions through facial expression. A system for detecting human emotions through facial expressions is proposed, in which we first extract facial features using deep learning methods, (VGG16 and MobileNet v1 of CNN models) and then train an SVM algorithm for emotion classification. The results showed that the properties extracted and classified by SVM are superior to the SoftMax classification method in the algorithms (VGG16, MobileNet v1) are used. We see an increase in accuracy of VGG16+SVM equal 3.07 compared to using the Softmax in VGG16. And the resulting accuracy increases by MobileNet+SVM equal 2.737 compared to MobileNet+Softmax. The second part we propose to model a hybrid neural network from each VGG with MobileNet to extract the features and then classification by SVM algorithm.**

Keywords— SVM, CNN, facial expressions, deeplearning, machine learning, classification, convolutional neural network, VGG16, MobileNet

I. INTRODUCTION

Emotions are mainly reflected in the movements of the voice, hand, body and facial expressions. People can communicate their intentions and emotions through some non-verbal means, such as facial expressions, writing, speech, and other involuntary means of language.

When we feel fear, our face expresses it through involuntary movements and small muscle movements that spread all over the face. Fear can be noticed despite efforts to hide it. It expresses non-verbal ways, which largely respond to the unconscious aspects of our personality; the same thing happens with other emotions and feelings: happiness, anger, sadness, and so on.

Facial expressions may be the most useful non-verbal way for people to communicate with each other. Using emotion recognition by recognizing facial expressions and detecting emotions is one such method. Therefore, scientists have tried to develop ways to identify feelings through facial expressions. However, recognizing facial expressions is a very difficult task. You can face problems such as lighting, noise, and masking some parts of the face. To achieve this goal, computer vision and machine learning technologies must be developed. In this study, seven classifications of emotions according to the human face (normal, sad, anger, disgust, fear, happy, surprise)[16].

The research paper is divided into two parts:

The first part: discussing the traditional methods of classification by extracting the features of the image using VGG and MobileNet and using the SVM classification algorithm.

The second part we propose to model a hybrid neural network from each VGG with MobileNet to extract the features and then classification by SVM algorithm.

The results of the experiments on KDEF dataset showed the superiority of the proposed hybrid Model over the traditional model. Where we get test accuracy =94.217

In the VGG16, we get accuracy =%86.7 on the test dataset, And in the MobileNet v1, we get accuracy =%90.8. In the VGG16, we get accuracy =%86.7 on the test dataset, And in the MobileNet v1, we get accuracy =%90.8. Where in the VGG16+SVM, we get accuracy =% 89.79 on the test dataset, And in the MobileNet v1+SVM, we get accuracy =% 93.537.

II. RELATED WORK

In last years, various methods been used to analyze and identify feelings through the expressions that appear on the face. These processes are concentrated, in pre-processing in order to prepare images, and then comes a processing stage to extracting features, and the final stage after extract the features the training stage of the different classification algorithms. Chen et al. (2014), Machine learning methods for extraction features using HOG (Histogram Oriented Gradient). Where the face (for the front side of the face) has been processed using SVM for classification. Shabat (2017), in this study, emotional analysis and features extraction using LBP & LDB, with using SVM & K-NN to classification.

Dachapally (2017), in the study there have an eight layer convolutional neural network. In addition, facial expressions are also used in the field of education. Use of automatic emotion recognition; It has great potential in various smart systems (e-health, learning, recommendation for tourism, smart city, smart talk, etc.) such as digital advertising, online games, customer feedback evaluation, health services. Ruiz-Garcia et al. (2016), Deep learning has shown real promise in classification efficiency for emotion recognition problems.

Ayvaz et al. (2017), Where they proposed in the studies, using Viola-Jones method to Face Detection, for detecting facial landmarks and after then attribute extraction for creating rules for instantaneous detection Of emotional expressions and creating training dataset with SVM classification to create Statically Classification Of Emotional Expressions. Akar et al. (2022), in the study conducted with a new model developed based on VGGNet which is one type the CNN models. Hamid et al. (2022). Use suggested histogram calculation to description features. The proposed CNN was trained for the teachable matrix using Chi-squared. Dandıl et al. (2019), in their study proposed an emotion recognition system based on facial expressions from real-time video frames with the classical convolutional neural network AlexNet.

III. MATERIALS

In the study, we used the KDEF dataset which is available in open sources [16][17]. It consists of 2938 JPEG images divided into 7 categories that express the basic feelings (happy , surprise, fear, happy, neutral, angry, sad and disgust). For each group consisting of 140 people, they were taken 3 photos of the same facial expression from three different sides (front, half of the right side, half of the left side) [17]. All images are taken by adjusting keep the eyes and mouth in the horizontal position, and the three images are taken at the same time. See Figure 1. It shows the front view of the face, the right half side of the face and the left half side of the face. The dataset is divided into training dataset equal to 2644 and test dataset equal to 294.

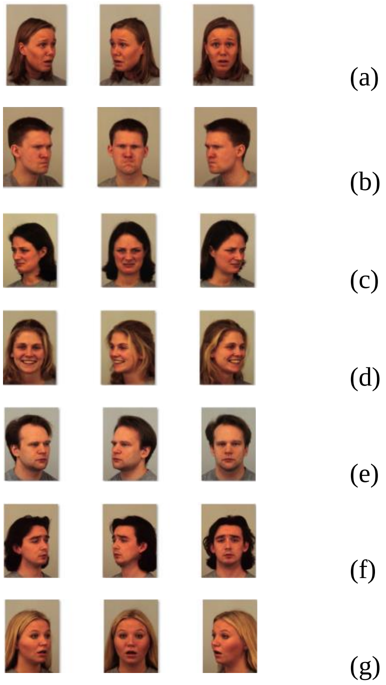


Fig. 1. Examples of face emotion in KDEF dataset (Fear(a), Angry(b), Disgust(c), Happy(d), Neutral(e), Sad(f), and Surprise(g))

IV. THE METHODS AND PROPOSED METHODS

The research paper is divided into two parts:

In this first part, this consists of two stages: first stage is Data Processing, and the second stage is Classification. The first stages: is to extract features from the dataset (images) by using one of the convolution neural network algorithms based on deep learning [2]. The first strategy: is to extract features from the dataset (images) by using one of the Convolution Neural Network algorithms based on deep learning. This first strategy has two parts [6]. The first is to train the CNN algorithm (VGG16 & MobileNet v1) on the training dataset to get trained weights. Second, using these weights trained with the CNN algorithm to extract features from the images. The second stage, after obtaining the features, it is classified them using the Support vector machines (SVM) algorithm for classification. In this second part [4–4], features are extracted by hybrid VGG&MobileNet and classified by SVM.

A. VGG16

The VGG16 algorithm 2015[3] is one of the low complexity convolutional neural network (CNN) algorithms. VGG16 features that all convolution layers consist of the same filter size (3x3) with a step of 1 (stride =1) and used same padding for all layers. Deepen and increase the number of layers with fewer parameters. And also using Maxpooling consisting of (2x2) and stride= 2 in all max pooling layers. The input layer of the image 224x224 with three channels (RGB image). Scheme of VGG16 as in Table 1.It consists of 13 convolution layers (Conv2d), 5 Maxpooling layers, 1 flattening layer, 2 Fully Connected (Dense) layers and last classification layer with activation function is softmax classification output layer. In this study, as previously described, using CNN algorithms and it one of deep learning to extract features from images [8]. This is why it doesn't need a Softmax classification layer at the feature extraction stage. It is used only once in algorithm training stage on the weights. So there is a one-time training process to obtain trained weights (which as fixed) for later use in the process of extracting features from the image. It gets 4096 features of the image after processing on the VGG16.

B. MobileNet

MobileNet [4] is uses depthwise convolution layers and Pointwise convolution to produce a deep and lightweight neural network is knows Depthwise Separable Convolution. The depthwise convolution significantly reduces the number of parameters compared other models with same depth of layers. In Table 2 the MobileNetv1 architecture that it adopted in this study. Where the input size of the image is 224 x 224, in the first layer is use Conv2D, stride =2 and filter shape (3 x 3 x 32). After each layer of Depthwise Convolution will come Conv2D layer with filter shape (1 x 1 x filter num.) and the stride value is 1, called Pointwise Convolution. In depthwise Convolution layer, stride=1 or stride=2, respectively.

After 26 Depthwise and Pointwise Convolutional layers, the get 7 x 7 x 1024 and the processed by an Average Pooling layer of size 7 x 7, the result 1 x 1 x 1024. This feature is ready for Fully Connected layer classification with Softmax. In the study, there have 7 classification categories were used.

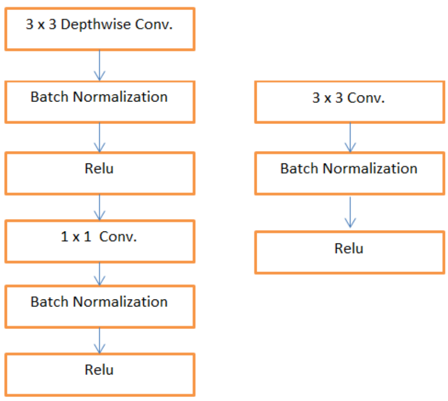


Fig. 2. The right standard convolutional layer (conv2D) with batch normalization (BN) and Activation -ReLU, Either the left Depthwise separable convolution with depthwise conv. , batch normalization, pointwise layers and activation - ReLU.

TABLE I. VGG16 ARCHITECTURE.

	Input	224 x 224 x 3
224 x 224	64 Conv1 + Relu	224 x 224 x 64
	64 Conv2 + Relu	224 x 224 x 64
	Max-pooling 1	112 x 112 x 64
112 x 112	128 Conv3 + Relu	112 x 112 x 128
	128 Conv4 + Relu	112 x 112 x 128
	Max-pooling 2	56 x 56 x 128
56 x 56	256 Conv5 + Relu	56 x 56 x 256
	256 Conv6 + Relu	56 x 56 x 256
	256 Conv7 + Relu	56 x 56 x 256
	Max-pooling 3	28 x 28 x 256
28 x 28	512 Conv8 + Relu	28 x 28 x 512
	512 Conv9 + Relu	28 x 28 x 512
	512 Conv10 + Relu	28 x 28 x 512
	Max-pooling 4	14 x 14 x 512
14 x 14	512 Conv11 + Relu	14 x 14 x 512
	512 Conv12 + Relu	14 x 14 x 512
	512 Conv13 + Relu	14 x 14 x 512
	Max-pooling 5	7 x 7
7 x 7	Flatten	1 x 25088
	Fully Connected (FC)	4096
	Fully Connected (FC)	4096
4096	FC+ Softmax	7
	output	7

TABLE II. MOBILENET ARCHITECTURE.

Input Size	Type	Stride	Filter Shape
224 x 224 x 3	Conv.	S 2	3x3x3x32
112 x 112 x 32	Conv. d w	S 1	3x3x32
112 x 112 x 32	Conv.	S 1	1x1x32x64
112 x 112 x 32	Conv. d w	S 2	3x3x64
56 x 56 x 64	Conv.	S 1	1x1x64x128
56 x 56 x 128	Conv. d w	S 1	3x3x128
56 x 56 x 128	Conv.	S 1	1x1x128x128
56 x 56 x 128	Conv. d w	S 2	3x3x128
28 x 28 x 128	Conv.	S 1	1x1x256x128
28 x 28 x 256	Conv. d w	S 1	3x3x256
28 x 28 x 256	Conv.	S 1	1x1x256
28 x 28 x 256	Conv. d w	S 2	3x3x256
14 x 14 x 256	Conv.	S 1	1x1x256x512
14 x 14 x 512	5 x Conv. d w	S 1	3x3x512
14 x 14 x 512	5 x Conv.	S 1	1x1x512x512
14 x 14 x 512	Conv. d w	S 2	3x3x512
7 x 7 x 512	Conv.	S 1	1x1x512x1024
7 x 7 x 1024	Conv. d w	S 2	3x3x1024
7 x 7 x 1024	Conv.	S 1	1x1x1024
7 x 7 x 1024	Avg. Pool	S 1	7x7
1 x 1 x 1024	Softmax	Classifier	
Output classifier 7 Category			

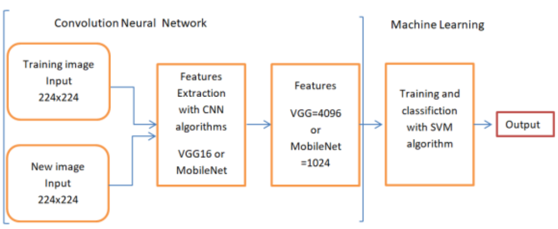


Fig. 3. Emotion Detection with CNN methods and SVM Classifiers structure

C. SVM classifier

The SVM classifier is one of Supervised Machine Learning algorithms for classification [1]. The main idea of SVM is to find the best line or decision boundary that separates classes, called a hyperplane. The SVM maximize the value of the margin, which is the distance between the supper vector and hyperplane. SVM's use of kernel functions, which gave it the ability to apply to various data structures, also gave it strength in terms of classification accuracy. In this study, the advantages of SVM are used to classify our proposed system. In Figure 3 the proposed system structure.

D. The Proposed Method

The proposed hybrid model is based on the integration of VGG and MobileNet to extract the features, and then classify them with SVA algorithm. The proposed architecture is based on 13 layers, in addition to the two FC layers of VGG. In parallel, a MobileNet operates using 10 convolutional layers and 9 deepwise convolutional layers. To produce 1024 features from each currency group (VGG, MobileNet). See the Figure 4. We perform a process of merging the features for the dataset, we divide the dataset into a training set and a test set, so that we can train the SVM algorithm for classification.

V. EXPERIMENTAL RESULTS

As mentioned in the previous chapter, in the first, the algorithm must be trained at the beginning only once on the dataset (images) that we have to form the trained weights that we will rely on in the process of extracting features from the image later. During the training stage, the goal is to reduce the error rate to a minimum and to ensure that the model (CNN to extracting features) works well on the new data (new image). In the VGG16 model, training the epoch set at 50, loss function is optimizers is {SGD(lr=0.001)} and (sparse_categorical_crossentropy). In the training stage of VGG16, we get accuracy =%86.7 on the test dataset. Also we get an error rate for the test dataset loss=0.44. See the Figure 5.

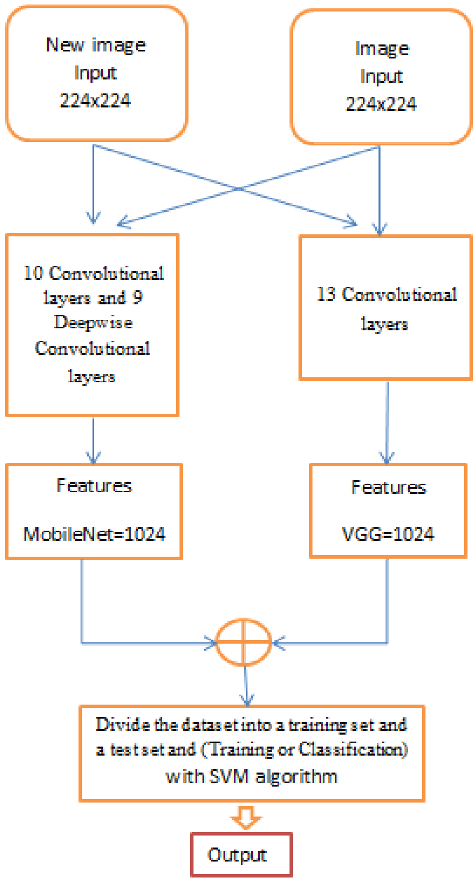
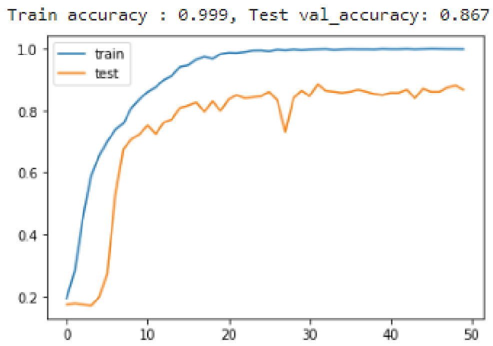
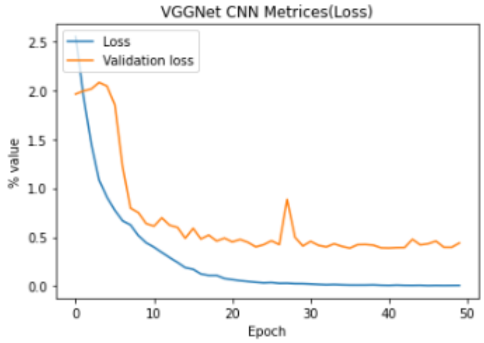


Fig. 4. Emotion Detection with hybrid CNN methods and SVM Classifiers structure



(a)

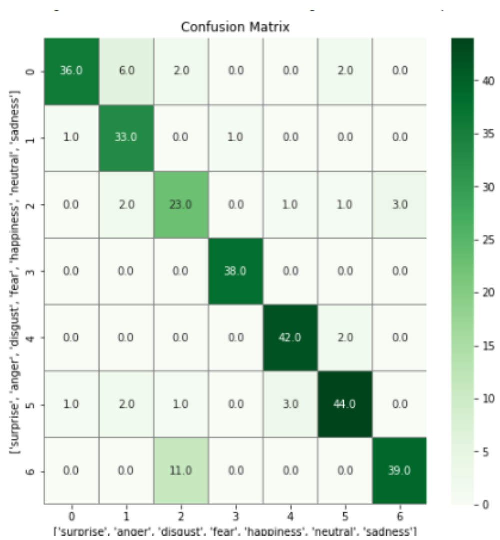


(b)

KDEF with VGG-16 CNN at epochs=50

	precision	recall	f1-score	support
0	0.95	0.78	0.86	46
1	0.77	0.94	0.85	35
2	0.62	0.77	0.69	30
3	0.97	1.00	0.99	38
4	0.91	0.95	0.93	44
5	0.90	0.86	0.88	51
6	0.93	0.78	0.85	50
accuracy			0.87	294
macro avg	0.86	0.87	0.86	294
weighted avg	0.88	0.87	0.87	294

(c)



(d)

Fig. 5. (a) VGG16 Accuracy, (b) VGG16 Loss, (d) VGG16 Confusion Matrix , (C) VGG16 Classification Report {0:"angry", 1:"disgust", 2:"fear", 3:"happy", 4:"neutral", 5:"sad", 6:"surprise"}

After the training process was completed on VGG16, the weights were trained for the model. Extract the features from the image in a dataset. This is after deleting the last layer of the model containing the FC Softmax layer. Thus, we get the features of the images according to the previously trained weights. The size of the extracted features is 4096. Now training the SVM algorithm using these features and testing them, and get accuracy = % 89.79. See the Figure 6. Now the MibileNetv1 model, training the epoch set at 50, loss function is (sparse_categorical_crossentropy) and optimizers is {Adam(lr=0.001)}. In the training stage of MobileNet v1, we get accuracy =%90.8 on the test dataset. Also we get an error rate for the test dataset loss=0.3902. See the Figure 7. After the training process for MobileNet v1 is complete, and it obtained trained weights for the model. Extract the features from the image in a dataset. This is after deleting the last layer of the model containing the FC Softmax layer. Thus, get the features of the images according to the previously trained weights. The size of the extracted features is 1024. Now training the SVM algorithm using these features and testing them, and get accuracy = % 93.537. See the Figure 8.

VI. The Proposed Method Results

After the training process for the proposed method is complete, and it obtained trained weights for the model. Extract the features from the image in a dataset. Thus, get the features of the images according to the previously trained weights. The size of the extracted features is 1024. Now training the SVM algorithm using these features and testing them, and get accuracy = % 94.217. See the Figure 9.

VII. Conclusions and Future Works

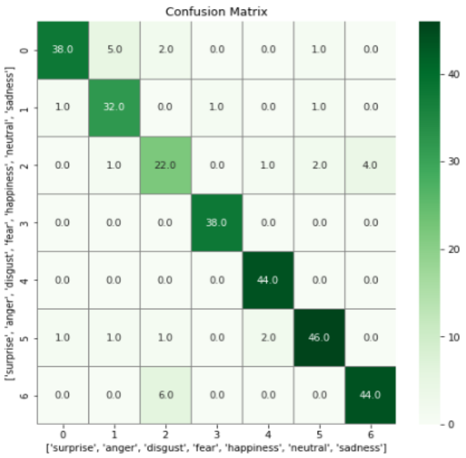
Through the result obtained, when using features extraction by deep learning for CNN VGG16 and CNN MobileNet v1, and using the SVM classifier instead of the last layer of Softmax classification in CNN model. We see an increase in accuracy of SVM equal 3.07 compared to using the Softmax in VGG16. And the resulting accuracy increases by MobileNet+SVM equal 2.737 compared to MobileNet+Softmax. See the Table 3 and Figure 10. When calculating the MobileNet+Softmax time it equal 0.609292, while the MobileNet+SVM time is 0.55083. And when calculating the VGG16+Softmax time it equal 1.430296, while the MobileNet+SVM time is 3.213790. In the Figure 9 we see the ROC curve for all classes (angry, disgust, fear, happy, neutral, sad and surprise). This experience can be benefited from and developed in the future by applying it in real time. When given new data, we only need to extract the features with pre-trained weights, and need to train the SVM algorithm on the new data for future classification.

When using the proposed method, it save the training time for the CNN algorithms when are given the new dataset for to training. Because we have previously prepared the weights and will use to extract facial expressions from the images. Where VGG16 and MobileNet v1algorithm to training is takes a long time if we want to train on new dataset images. Where when using the proposed method, we train the VGG16 and MobileNet methods on the weights, save them, and use them to extract the features from the image, then train the SVM classification algorithm on the Features extracted from the image, the new data, which makes it faster in training and prediction. That is why we recommend that you develop and test methods on other CNN methods and apply them in real time.

KDEF with VGG-16 & SVM

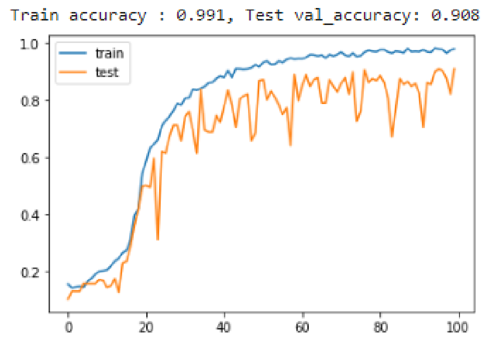
	precision	recall	f1-score	support
0	0.95	0.83	0.88	46
1	0.82	0.91	0.86	35
2	0.71	0.73	0.72	30
3	0.97	1.00	0.99	38
4	0.94	1.00	0.97	44
5	0.92	0.90	0.91	51
6	0.92	0.88	0.90	50
accuracy			0.90	294
macro avg	0.89	0.89	0.89	294
weighted avg	0.90	0.90	0.90	294

(a)

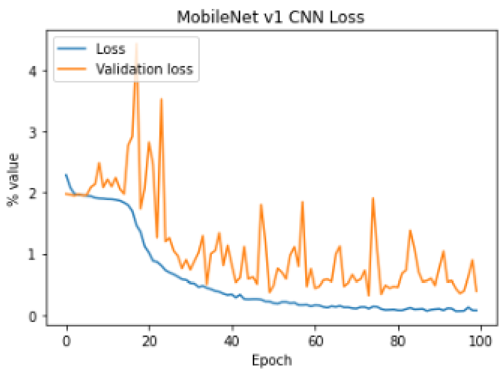


(b)

Fig. 6. (a) VGG16 & SVM Classification Report{ 0:"angry", 1:"disgust", 2:"fear", 3:"happy", 4:"neutral", 5:"sad", 6:"surprise"}. (b) VGG16 & SVM Confusion Matrix



(a)

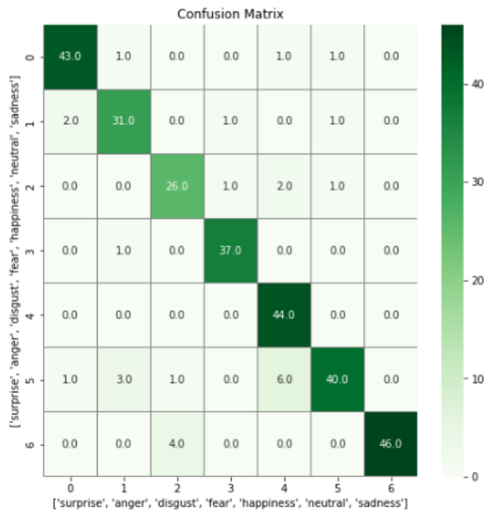


(b)

MobileNet v1 With KDEF CNN at epochs=100

	precision	recall	f1-score	support
0	0.93	0.93	0.93	46
1	0.86	0.89	0.87	35
2	0.84	0.87	0.85	30
3	0.95	0.97	0.96	38
4	0.83	1.00	0.91	44
5	0.93	0.78	0.85	51
6	1.00	0.92	0.96	50
accuracy			0.91	294
macro avg	0.91	0.91	0.91	294
weighted avg	0.91	0.91	0.91	294

(c)



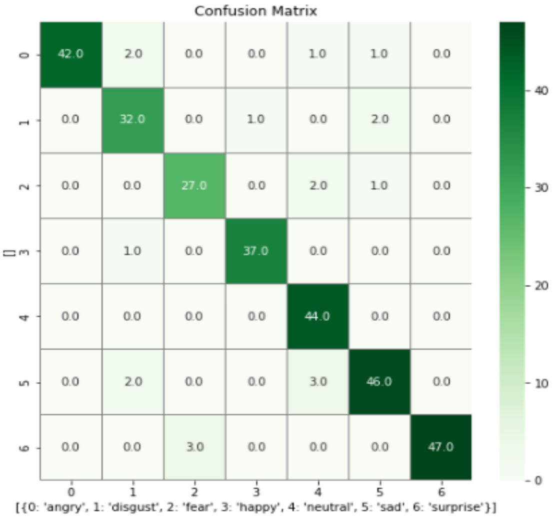
(d)

Fig. 7. (a) MobileNet v1 Accuracy. (b) MobileNet v1 Loss, (c) MobileNet v1 Classification Report {0:"angry", 1:"disgust", 2:"fear", 3:"happy", 4:"neutral", 5:"sad", 6:"surprise"} (d) MobileNet v1 Confusion Matrix.

KDEF MobileNet v1 Features Extrection with SVM

	precision	recall	f1-score	support
0	1.00	0.91	0.95	46
1	0.86	0.91	0.89	35
2	0.90	0.90	0.90	30
3	0.97	0.97	0.97	38
4	0.88	1.00	0.94	44
5	0.92	0.90	0.91	51
6	1.00	0.94	0.97	50
accuracy			0.94	294
macro avg	0.93	0.93	0.93	294
weighted avg	0.94	0.94	0.94	294

(a)



(b)

Fig. 8. (a) MobileNet v1 Classification Report, (b) MobileNet v1 Confusion Matrix

```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95        46
           1       0.87      0.94      0.90        35
           2       0.90      0.90      0.90        30
           3       0.97      0.97      0.97        38
           4       0.92      1.00      0.96        44
           5       0.92      0.92      0.92        51
           6       1.00      0.94      0.97        50

    accuracy                           0.94       294
   macro avg       0.94      0.94      0.94       294
weighted avg       0.94      0.94      0.94       294
```

Fig. 9. Hybrid CNN Methods and SVM Classifiers Report

TABLE III. ACCURACY VGG16, MOBILENET V1, VGG16+SVM MOBILENET+SVM AND HYBRID CNN METHODS

Model	Softmax classifier %	SVM classifier %
VGG16	86.7	89.79
MobileNet v1	90.8	93.537
Hybrid CNN Methods		94.217

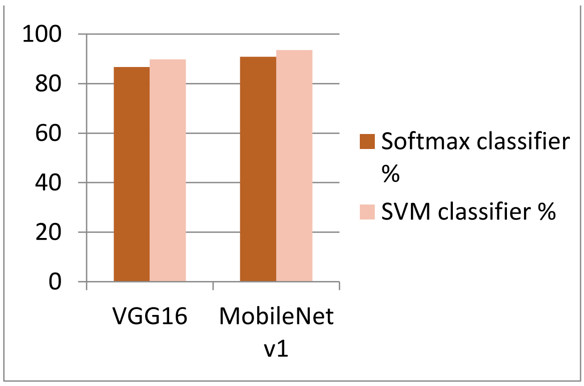
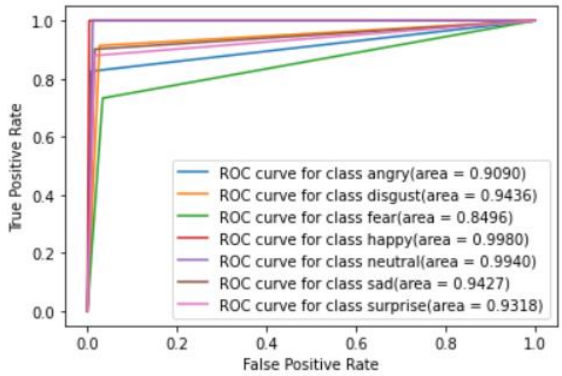
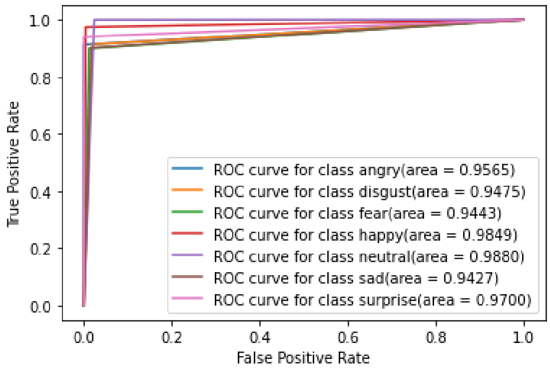


Fig. 10. Histogram Accuracy VGG16, MobileNet v1, VGG16+Svm and MobileNet+SVM



(a)



(b)

Fig. 11. The ROC curve for all classes (angry, disgust, fear, happy, neutral, sad and surprise).

REFERENCES

[1] Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5), 3.

[2] Courbariaux, M., Bengio, Y., & David, J. P. (2014). Training deep neural networks with low precision multiplications. arXiv preprint arXiv:1412.7024.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[4] Andrew, G., & Menglong, Z. (2017). Efficient convolutional neural networks for mobile vision applications. Mobilenets.

[5] SÜNNETCİ, K. M., AKBEN, S. B., KARA, M. M., & ALKAN, A. Face Mask Detection Using GoogLeNet CNN-Based SVM Classifiers. Gazi University Journal of Science, 36(2), 645-658.

[6] Çınar, A., & Tuncer, S. A. (2021). Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM. SN Applied Sciences, 3(4), 1-11.

[7] Çınar, A., & Tuncer, S. A. (2021). Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks. Computer methods in biomechanics and biomedical engineering, 24(2), 203-214.

[8] Kutlu, H., Avci, E., & Özyurt, F. (2020). White blood cells detection and classification based on regional convolutional neural networks. Medical hypotheses, 135, 109472.

[9] Brownlee, J. (2018). Better deep learning: train faster, reduce overfitting, and make better predictions. Machine Learning Mastery.

[10] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) (pp. 67-74). IEEE.

[11] Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., & Hua, G. (2017). Neural aggregation network for video face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4362-4371).

[12] Hassaballah, M., & Awad, A. I. (Eds.). (2020). Deep learning in computer vision: principles and applications. CRC Press.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[14] Pinaya, W. H. L., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Convolutional neural networks. In Machine learning (pp. 173-191). Academic Press.

[15] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.

[16] https://www.kdef.se/download-2/7Yri1UsotH.html

[17] https://www.kaggle.com/datasets/tom99763/testtt

[18] Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014, August). Facial expression recognition based on facial components detection and hog features. In International workshops on electrical and computer engineering subfields (pp. 884-888).

[19] Shabat, A. M. M. (2017). Improvements of local directional pattern for texture classification (Doctoral dissertation).

[20] Ayvaz, U., & Gürüler, H. (2017). The detection of emotional expression towards computer users. International Journal of Informatics Technologies, 10(2), 231-239.

[21] Sadeghi, H., & Raie, A. A. (2022). Histnet: Histogram-based convolutional neural network with chi-squared deep metric learning for facial expression recognition. Information Sciences, 608, 472-488.

[22] AKGÜL, İ., & Funda, A. K. A. R. (2022). Emotion Recognition from Facial Expressions by Deep Learning Model. Journal of the Institute of Science and Technology, 12(1), 69-79.

[23] Dachapally, P. R. (2017). Facial emotion detection using convolutional neural networks and representational autoencoder units. arXiv preprint arXiv:1706.01509.

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

[25] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., & Palade, V. (2016, September). Deep learning for emotion recognition in faces. In International Conference on Artificial Neural Networks (pp. 38-46). Springer, Cham.