Explainable Model Selection of a Convolutional Neural Network for Driver's Facial Emotion Identification

Chapter *in* Lecture Notes in Computer Science · February 2021

DOI: 10.1007/978-3-030-68780-9_53

3 authors, including:

Amany Kandeel
Queen's University
3 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Hossam S. Hassanein
Queen's University
726 PUBLICATIONS **10,718** CITATIONS
SEE PROFILE

Explainable Model Selection of a Convolutional Neural Network for Driver's Facial Emotion Identification

Amany A. Kandeel[1](✉), Hazem M. Abbas[2], and Hossam S. Hassanein[1]

[1] School of Computing, Queen's University, Kingston, Canada
19aak6@queensu.ca, hossam@cs.queensu.ca

[2] Department Computer and Systems Engineering, Ain Shams University, Cairo, Egypt
hazem.abbas@eng.asu.edu.eg

Abstract. Road accidents have a significant impact on increasing death rates. In addition to weather, roads, and vehicles, human error constitutes these accidents' main reason. So, driver-safety technology is one of the common research areas, whether in academia or industry. The driver's behavior is influenced by his feelings such as anger or sadness, as well as the physical distraction factors such as using mobile or drinking. Recognition of the driver's emotions is crucial in expecting the driver's behavior and dealing with it. In this work, the Convolutional Neural Network (CNN) model is employed to implement a Facial Expression Recognition (FER) approach to identify the driver's emotions. The proposed CNN model has achieved considerable performance in prediction and classification tasks. However, it is similar to other deep learning approaches that have a lack of transparency and interpretability. We use Explainable Artificial Intelligence (XAI) techniques that generate interpretations for decisions and provide human-explainable representations to address this shortage. We utilize two visualization methods of XAI approaches to support our decision of choosing the architecture of the proposed FER model. Our model achieves accuracies of 92.85%, 99.28%, 88.88%, and 100% for the JAFFE, CK+, KDEF, and KMU-FED datasets, respectively.

Keywords: Driver-safety · Facial Expression Recognition · Convolutional Neural Network · Explainable Artificial Intelligence

1 Introduction

As indicated by the worldwide status report on road safety, the World Health Organization (WHO) announced that the number of annual road traffic deaths is approximately 1.3 million [32]. The reasons can be a result of the road conditions, weather, or vehicle, but the main reasons for this tremendous number of incidents can be attributed to the driver's behavior. The motorist behavior is affected by

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12666, pp. 699–713, 2021.
https://doi.org/10.1007/978-3-030-68780-9_53

drinking alcohol, using mobile, fatigue, and other physical factors, but emotions can also be one of the significant factors that affect driver behavior and increase accident probability. According to Virginia Tech research, [10] about studying the risk of a crash, claims that the risk of accidents is influenced by various motivations such as: drinking, drug, and alcohol, mobile phone dialing, reading and writing, emotions (i.e., anger, sadness, and crying), reaching for an object, drowsiness, fatigue, etc.; all these factors are organized concerning its effect on a driver's distraction in Fig. 1.

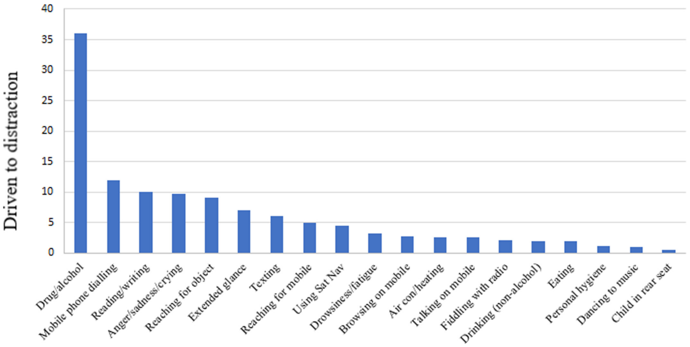


Fig. 1. Different motivations effects of crash risk [10].

As supposed by this study, driving with sadness, anger, or agitation feelings increases the danger of an accident by almost ten times. While unexpectedly, drowsiness or fatigue makes a crash three times more likely. At the same time, eating or talking on mobile-only doubles the risk. Previously, numerous highways show signs that caution drivers to continue driving if they feel tired, and they should pull over. However, new researches propose that a healthy emotional state is more significant for safe driving [10]. Emotions can be considered as a procedure that elevates adjustment to the environment and prepares the individual for adaptive action [25]. This process is accompanied by expressions such as face or voice, in addition to some physiological alterations such as an increment in heart rate. Mesken [20] explained the three functions of emotion concerning examples about driving. The first function of emotion is the adaptive function, where emotions are adaptive, and this means they support behavior that is suitable for the environment. For instance, when a vehicle's driver is up to another vehicle unexpectedly coming from a side street, the dread or surprise reaction makes the driver brake in a flash. This response is much faster than a more cognitive response. A second function is a communicative or social function. We can predict others' behaviors by watching their emotions, but it is difficult to do that in driving issues, so the drivers can represent their feelings in other ways such as hand gestures or using horns. The last function of emotions depends on planning. People may complete their goals by using their emotions. A driver can use a hand gesture to express his apologies if he prevents another driver's progress during parking to avert a hostile response. Detecting a driver's emotion is quite crucial to predicting his behavior; for instance, detecting aggressive driver emotion can help keep the driver safe. This detection can be performed by using biomedical means such as Electrocardiogram (ECG) or Electroencephalogram (EEG) sensors [3]. However, these sensors are not comfortable for the driver. Others use a self-reporting questionnaire, and this is also unpractical because it needs a lot of effort and time in recording the symptoms [22]. Some approaches depend on evaluating drivers' behavior by analyzing data from sensors at different vehicle components, such as the steering wheel and pedals. Facial expression is considered a behavioral approach that monitors the face, eye, mouth, or any other face component to evaluate the driver's behavior; it is more reliable, which indicates the driver state immediately in real-time [6].

Facial Expression Recognition (FER) early approaches used manual feature engineering to build emotion classifiers. Both feature extraction and classification methods are chosen to cope with the datasets [15] such as Support Vector Machines (SVMs), Modified Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) [21]. Recently, deep learning approaches are trained to learn emotion features from input images automatically. Deep Neural Networks (DNNs) have been successfully applied to computer vision tasks such as object detection, image segmentation, and classification [27]. This advancement is because of the development of Convolutional Neural Networks (CNNs) [9]. Despite the great accomplishment of deep learning approaches in detection and classification, most of these systems' decisions and behavior are less understandable for humans. Recently, researchers provided some explainable methods to illustrate Artificial Intelligence (AI) models to be more interpretable.

This paper has two main goals: first, to identify the driver's psychological state utilizing the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) to facilitate safe driving. We implement a FER model using CNN because of the high accuracy it can achieve. Secondly, find the best architecture to achieve the highest accuracy with the lowest possible complexity using some explainability visualizing methods to illustrate the reasons for selecting a particular model architecture.

The rest of our paper is structured as follows. Section 2 provides the related work. The proposed explainable model selection approach is introduced in Sect. 3. The experimental results are presented in Sect. 4. Finally, we conclude our work in Sect. 5.

2 Related Work

Although some companies intend to produce fully autonomous vehicles during the next five years, it is considered as a future step. Furthermore, most companies depend on semi-autonomous systems such as the Advanced Driver Assistance System (ADAS) or Driver Monitoring System (DMS), which have seen significant improvements in recent years. These systems can include automatic parking systems, lane assistance, or driver concentration control that help with a way or another in increasing road safety [12].

The system can identify aggression, fatigue, or drunk driving by recognizing the driver's facial expression. This identification aids to improve traffic situations and reduces accidents occurrence [31]. Some researchers use classical machine learning approaches such as SVMs, K-Nearest Neighbors (K-NN), Hidden Markov Models (HMMs), and Random Forests in FER systems [21]. These approaches are suitable for small datasets and do not need a massive amount of computation time, speed, or memory [2]. Jeong et al. [8] presented a FER method based on geometric features and the hierarchical Weighted Random Forest (WRF). First, they reduced the number of landmarks used for generating geometric features. Afterward, they differentiated the facial expressions by using a hierarchical WRF classifier. Additionally, they developed a new benchmark dataset, Keimyung University Facial Expression of Drivers (KMU-FED), which considers the real driving environment. Their model achieved accuracies of 94.7%, 92.6%, and 76.7% for the KMU-FED, CK+, and MMI datasets, respectively.

On the other hand, deep learning-based approaches have been shown to achieve high accuracy, and thus many researchers employed them in FER solutions. Du et al. [4] presented a deep learning framework called Convolution Bidirectional Long Short-Term Memory Neural Network (CBLNN). They used facial features and heart rate to predict the driver's emotions using CNNs and Bidirectional Long Short-Term Memory (Bi-LSTM), respectively. The output of the Bi-LSTM was used as input to the CNN module. CBLNN classified the extracted information into five common emotions: anger, happiness, neutrality, fear, and sadness. To evaluate their model, they used a simulated driving environment. They proved that the combination of using facial features and heart rate increased the accuracy of emotion identification. Lee et al. [14] used a CNN model for aggressive emotion detection. They used near-infrared (NIR) light and thermal camera sensors to capture the driver's images. They evaluated the model on their dataset. Zhang et al. [34] proposed a model to detect stress by recognizing related facial expressions, anger, fear, and sadness. They used four layers of the connected convolutional network, which combined low-level features with high-level features to train the deep network to identify facial expressions. They used three datasets to evaluate their model that achieved an accuracy of 93.6% for CK+, 81.3% for Oulu-CASIA, and 97.3% for the KMU-FED dataset.

3 The Proposed Explainable Model Selection Approach

This paper's main objective is to build a simplified FER model with high accuracy to recognize the driver's facial emotions. To achieve high performance, we experimented with different CNN architectures. The architecture of the proposed CNN model results from experiments with a different number of layers with different sizes of convolution kernels. Before applying the CNN model on datasets images, we used face detection to define face areas, crop these faces from images, and remove irrelevant information. After the face extraction, the images are fed to the CNN model for training.

3.1 Face Extraction

The image's background may contain unimportant details, which can reduce the expression's recognition accuracy. Therefore, the first step of our approach is detecting the face area, followed by extracting it. We use the *Haar Cascades* algorithm [30] in detecting the faces in the images. *Haar Cascades* algorithm utilizes the *Adaboost* learning mechanism [13] which picks a few significant highlights from a huge set to give a productive consequence of classifiers. *Haar* features are composed of either two or three rectangles. Face candidates are scanned and looked for *Haar* features of the current stage. The weights are constants produced by the learning algorithm. There is a variety of forms of features, as can be seen below in Fig. 2. The *Haar* feature classifier multiplies the weight of each rectangle by its area, and the results are added together [13].

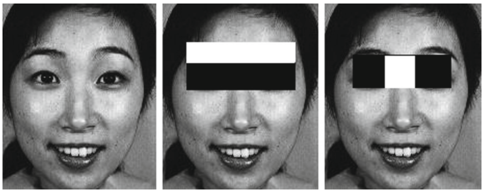


Fig. 2. Examples of Haar features.

3.2 CNN Model

CNN has dramatically decreased the dependence on physics-based models by allowing "end-to-end" learning directly from input images. Therefore, CNN achieves a significant performance in various applications such as image recognition and classification [11]. CNN includes two main stages; the first stage is the feature extraction layer(s), responsible for features detection. It contains a convolutional layer(s) and a pooling layer. The second stage is the classifying layer, which includes a fully connected layer(s) and the final output layer.

Convolutional Layer: This layer extracts different features from the input image by using a series of convolutional kernels. Each kernel has a particular task such as sharpening or edge and line detection [5]. In this layer, ReLU is utilized as an activation function, and it is one of the most popular activation functions. It does not cause the vanishing gradient that may be caused by other activation functions such as *Sigmoid* function.

Pooling Layer: Pooling converts the spatial information into features and helps in decreasing memory consumption in deeper layers. There are three types of pooling, *max pooling*, *average pooling*, and *sum pooling*. The pooling layer does not have any parameters, and the depth that represents the number of output features is always the same as the input.

Fully Connected Layers: This layer is responsible for classification, so all nodes are fully connected to the nodes in the following layer. The output of the convolution layer before entering the fully connected is usually flattened into a vector. In the output, the *Softmax* function is employed to produce a value between 0 and 1, based on the model's confidence. The output class is one of the six emotions [5]. We endeavor to implement a FER model that fulfills both goals of simplicity and performance. To specify the optimal architecture that achieves the highest accuracy, we tried different layers and various filter numbers and compared each architecture's performance. We also compared the model complexity using the number of its parameters. Section 4.2 explains this comparison in detail and analyzes the results. Furthermore, to deeply understand the model's behavior, we employed some interpretation tools to verify different architectures' performance and explained the reasons for selecting the proposed model's particular design; these tools are described in the next section.

3.3 Explainable Model

Although the deep learning approaches outperform other approaches, it is considered a black box in its nature, and this is the primary obstruction to use it. Researchers may undertake many experiments to obtain the perfect model that achieves the best results specific to their task. However, they can hardly explain the real reasons for providing a particular model. Indeed, they may not have a complete conception of illustrating the results of their model. To address this issue, a new approach, named Explainable Artificial Intelligence (XAI) has appeared to make the machine learning algorithms more understandable; XAI does not only give an appropriate explanation to AI-based systems but also improves the transparency and trust of them [1]. Many interpretability techniques are used to discuss explainability in Machine Learning (ML) algorithms. We used two of the most frequent interpretation methods in deep learning, the Saliency map [28], and Gradient-Weighted Class Activation Map (Grad-CAM) [26].

Saliency Map: It is one of the local interpretable methods which give the reason for a specific decision and explain it [1]. This map is used to clarify the importance of some areas in the input image that might be involved in generating the class prediction. Saliency is calculated at a particular position by the degree of the difference between that position and its surrounding [7]. Simonyan et al. [28] tried to categorize the pixels of image I_0 dependent on their impact on the class score function $S_c(I_0)$ of class c. The first step to calculate the class saliency map M_{ij} is calculating w as the derivative of the class score S_c with respect to image I, S_c is obtained from the relation, $S_c(I) = w_c^T I + b_c$, where w_c, b_c are the weight vector and the bias of the model, respectively [28],

$$w = \frac{\partial S_c}{\partial I}\Big|_{I_0} \tag{1}$$

$$M_{ij} = max_c \left| w_{h(i,j,c)} \right| \tag{2}$$

where $h(i, j, c)$ is the index at i-th row and j-th column of the element w of class c.

Grad-CAM: This map specifies each neuron's important values for a specified decision of interest; it utilizes the gradient information flowing into CNN's last convolutional layer. It distinguishes among different classes, wherever it is exclusively highlighting the class regions for the class to be visualized [26]. In other words, it shows which regions in the image are related to this class. With this tool, two features from previous methods are combined; the class-discriminative, which is achieved in the Class Activation Map (CAM) approach [35], and the high-resolution which is provided by Guided Backpropagation [29]. To specify the significant values of each neuron for a specified choice, Selvaraju et al. use gradient information at the last convolutional layer. First, they determine neuron importance weights α_k^c [26],

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

where Z is the number of pixels in the feature map, $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient of the score y^c for class c with respect to the feature map activation A^k. Afterwards, to determine the class-discriminative localization map Grad-CAM $L^c_{Grad-CAM}$, they apply the ReLU function on the weighted combination of forward activation maps,

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha_k^c A^k) \tag{4}$$

They use ReLU to highlight only the features that have a positive effect on the class of interest.

After comparing different CNN architectures, which is discussed in detail in Sect. 4.2, one can determine the optimal CNN model architecture, which is highly accurate with the least number of parameters. The model is shown in Fig. 3. It is composed of three convolutional layers with a 3 × 3 convolution kernel and a stride value of 1, followed by a max-pooling layer with a 2 × 2 kernel. The number of filters is 32, 64, 128 for the three layers, respectively. The first layer has 256 nodes at the fully connected layers, and the second layer has 128 nodes, with a dropout ratio of 0.5. At the final output layer, there are six nodes for classifying the six basic emotions.

4 Experimental Results

To evaluate our model, we used four different datasets. These datasets vary from using static images or sequences of frames.

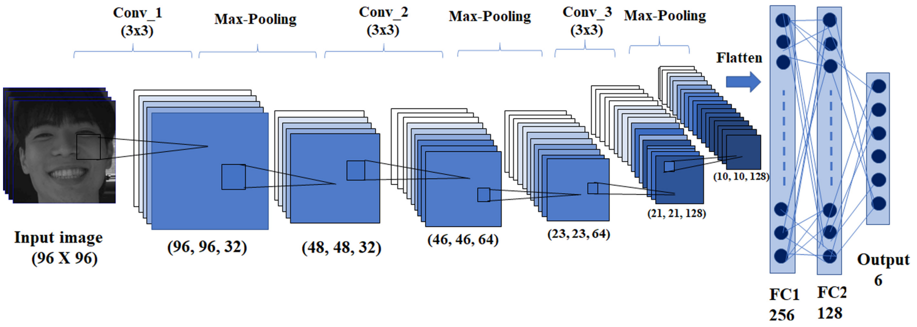


Fig. 3. The proposed CNN architecture

4.1 Datasets

Driver's emotions datasets are scarce, so many researchers used the general FER datasets such as CK+, Oulu-CASIA, and MMI [8,14]. Others use a simulated driving environment [4]. Some researchers built their own dataset with images in a real driving environment, such as the KMU-FED dataset [8]. In this work, we used both the general FER datasets, JAFFE, CK+, and KDEF, in addition to a real driving dataset, KMU-FED.

CK+ Dataset: The Extensive Cohn–Kanade (CK+) [17] has 327 labeled sequences of frames with two different sizes (640 × 480 and 640 × 490 pixels). The emotion labels are categorized into seven emotions; anger, contempt, disgust, fear, happiness, sadness, and surprise. The image sequences have various durations (10 to 60 frames) which contain different lighting levels. The emotion sequence starts from neutral representation to peak formation of the facial emotion. Therefore, we utilized the last three frames, which are the most expressive frames to represent the emotion [16]. To do a fair comparison with the recent approaches [8,16,33,34] in our experiments, we used only the six basic expressions without the contempt emotion. Therefore, the total number of frames is 930 frames.

JAFFE Dataset: The Japanese Female Facial Expression (JAFFE) database [19] is the most commonly used and oldest FER datasets. It contains 213 images of 7 facial expressions (neutral as well as the six main expressions) that 10 Japanese female models took. Images are digitized into 256×256 pixels. We used the six main emotions, to perform a reasonable correlation with other approaches [8,16,33,34]. So the total number of images is 183 images.

KDEF Dataset: The Karolinska Directed Emotional Faces (KDEF) [18] was provided by the Karolinska Institute in 1998. It has seven different emotions (afraid, angry, disgust, happy, neutral, sad, and surprise) with an equal number of male and female expressions. This dataset has frontal and non-frontal faces, where each expression is viewed from five different angles. We used just the six basic expressions with the frontal faces profile, in order to compare equitably with the ongoing methodologies [8,16,33,34]. Thus the total number is 2504 samples with digitization 562 × 762 pixels.

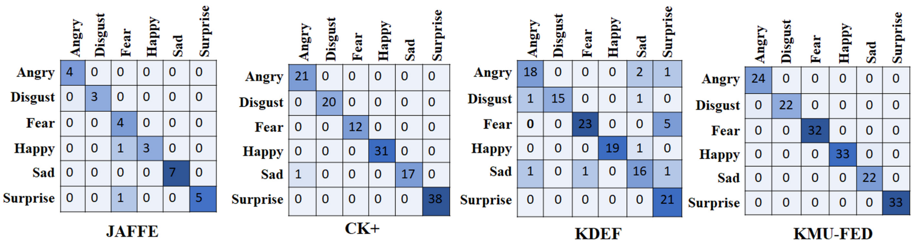
KMU-FED Dataset: Keimyung University Facial Expression of Drivers (KMU-FED) [8] contains images in a real driving environment. It consists of six basic expressions captured using a Near Infrared (NIR) camera installed on the dashboard or steering wheel. It includes 55 image sequences from 12 subjects with a total of 1106 frames. The image sequences vary in duration (10 to 26 frames) that include various changes in illumination (front, back, right and left light) and partial occlusions caused by sunglasses or hair. The images have pixel resolutions of 1600 × 1200 pixels.

4.2 Discussion and Analysis

The model was evaluated using the previous four datasets. It was trained with 70% of the dataset samples, validated with 15%, and tested with 15%. The number of correctly predicted emotions is provided in each dataset's confusion matrix, as shown in Fig. 4. After extracting the face region from images, we applied the proposed CNN model on the input images with different sizes to determine which image resolution achieves the best results. Each dataset has the highest test accuracy with particular sizes. Figure 5 shows that the JAFFE dataset's highest accuracy is 92.85% with image size 96×96. Regarding the CK+ dataset, its most extraordinary performance occurs at size 48 × 48 pixels, and it is 99.28%. We reached the best accuracy for the KDEF dataset, 88.88% at both sizes 112 × 112 and 160 × 160 pixels. Finally, the model achieved the highest performance of 100% for the KMU-FED dataset at three different resolutions; 96 × 96, 112 × 112, and 144 × 144 pixels.

To demonstrate the effect of image resolution on the accuracy, we used the proposed CNN architecture, as shown in Fig. 3. The image sizes vary from 32×32 to 208×208. As shown in Fig. 5, each dataset has the best accuracy at particular resolutions. Both CK+ and KMU-FED datasets are not affected greatly by the change in the image size. On the contrary, the JAFFE dataset is enormously influenced by image resolution variations. This influence can be attributed to the low resolution of the original images in this dataset. Besides, this dataset has few samples and is the main cause of its weak performance that can be deduced from Fig. 5. The impact of the resolution variation on the KDEF dataset is moderate.

Also, to show the effect of different CNN model architectures with a different number of layers and filters, we chose a specified size, 48 × 48, to reduce the number of parameters, change in the number of layers, and the number of filters



JAFFE

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	4	0	0	0	0	0
Disgust	0	3	0	0	0	0
Fear	0	0	4	0	0	0
Happy	0	0	1	3	0	0
Sad	0	0	0	0	7	0
Surprise	0	0	1	0	0	5

CK+

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	21	0	0	0	0	0
Disgust	0	20	0	0	0	0
Fear	0	0	12	0	0	0
Happy	0	0	0	31	0	0
Sad	1	0	0	0	17	0
Surprise	0	0	0	0	0	38

KDEF

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	18	0	0	0	2	1
Disgust	1	15	0	0	1	0
Fear	0	0	23	0	0	5
Happy	0	0	0	19	1	0
Sad	1	0	1	0	16	1
Surprise	0	0	0	0	0	21

KMU-FED

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	24	0	0	0	0	0
Disgust	0	22	0	0	0	0
Fear	0	0	32	0	0	0
Happy	0	0	0	33	0	0
Sad	0	0	0	0	22	0
Surprise	0	0	0	0	0	33

Fig. 4. The confusion matrix of each dataset

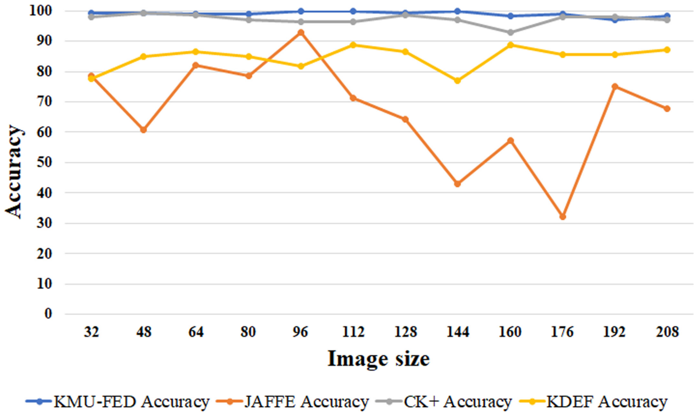


Fig. 5. Dataset accuracy versus different image sizes

in each layer as shown in Fig. 6. For each architecture, there are two varying values, the number of parameters and the model accuracy. Clearly, the number of parameters does not depend only on the number of filters in each layer, but it also relies on the number of flattened values before fully connected layers. The number of parameters P in a convolution layer can be calculated as

$$P = (((m \cdot n \cdot d) + 1) \cdot k) \qquad (5)$$

Where m is the filter width, n is the filter height, d is the number of filters in the previous layer, and k is the number of filters in the current layer. For instance, the highest number of model parameters occurs when only two layers are employed. On the other hand, when we used four layers, we obtained the smallest value. The main reason for that, in the first case, the last convolutional layer before the fully connected layers has a high spatial resolution, 11 × 11, and then the number of parameters is 4,075,014. However, in the case of four layers, the image information passes through four layers, so the last convolutional layer size is 1 × 1, and consequently, the number of parameters is only 164,646. One can conclude that, in addition to the number of filters, the model's complexity depends strongly on the values that are fed to the dense layers. Our main aim is

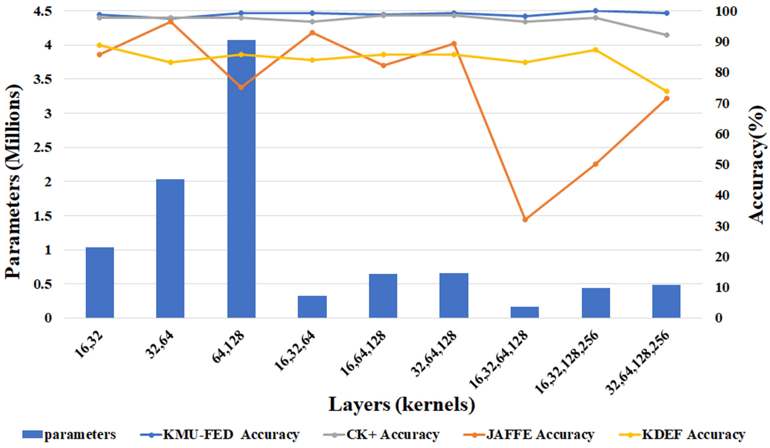


Fig. 6. Datasets accuracy versus different CNN architectures

to recognize facial expressions with high accuracy and simplicity. The two goals might seem to contradict, so we tried to balance them using a CNN model with a small number of parameters and provide the best performance. Thus, in light of the relationship analysis between the number of layers and both the accuracy and the number of parameters, we may select the architecture of three layers, which achieves both goals with the four datasets. However, to verify this model selection, we used explainable methods to clarify the model's behavior more thoroughly. Due to the internal complexity and nonlinear structure of deep neural networks, it is obscure and perplexing to interpret the reasons for designing a particular architecture. To state the reasons for using this specific architecture, we used some visualization tools such as the Saliency map [28], and Gradient-Weighted Class Activation Maps (Grad-CAM) [26].

Saliency Map: We employed the saliency map with the three different architectures, two, three, and four layers. We can conclude from Fig. 7 that the saliency map is accurate and precise for three layers' architecture. The saliency map of the two layers does not contain any details. Although the four layers of CNN architecture's saliency map has a dense view, it is ambiguous for understanding.

Grad-CAM: We applied this method to the input image with three different architectures. The Grad-CAM of the two layers' architecture highlights some facial features such as eyebrows, nose, and mouth. The Grad-CAM of the three layers is more accurate because it spotlights the facial features that affect the emotion shape, such as in angry emotion, highlighting the area of the mouth, the cheeks, and head front as shown in Fig. 7. On the other hand, the Grad-CAM of the model with four layers is not clear.

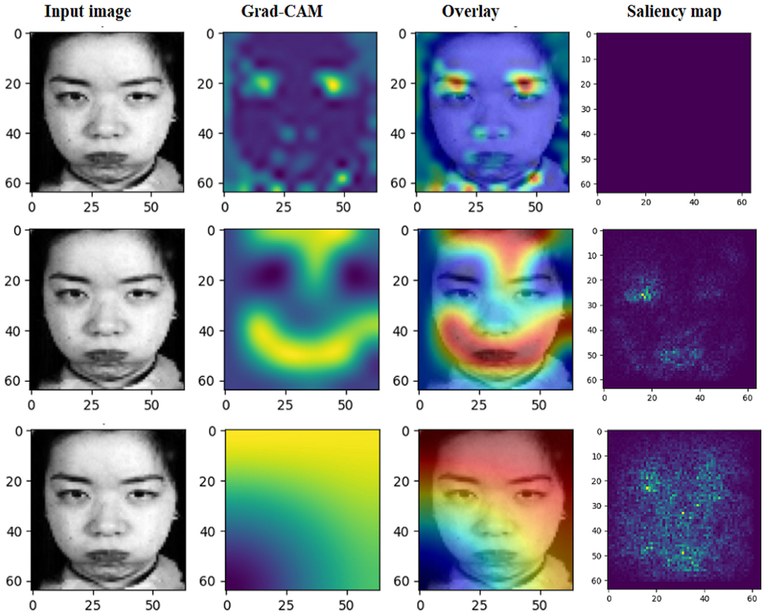


Fig. 7. Grade-CAM and Saliency maps for different CNN architectures (16, 32), (32, 64, 128), (16, 32, 64, 128) layers for angry emotion

We compared our models' results to some other state-of-the-art FER approaches in Table 1, our results are comparable to these results. Some of them used general FER datasets, such as Lopes et al. [16] who used data augmentation techniques to increase the size of datasets in order to improve the accuracy. However, their model's accuracies are 53.44% and 96.76% for JAFFE and CK+ datasets, respectively. Although Yang et al. [33] utilized a more complex model, the VGG16 model, which consists of 16 layers with 138 million parameters, their model achieved accuracies of 92.21% and 97.02% for the JAFFE and CK+ datasets, respectively. Pandey et al. [23] provided a combination of the Laplacian of the input image and the gradient with the original input into a CNN model; they evaluated their model on the KDEF dataset and provided an accuracy of 83.33%. Puthanidam et al. [24] used data augmentation techniques to overcome the small size dataset's challenge, which led to improved performance. The accuracies were 100% and 89.58% for JAFFE and KDEF datasets, respectively. We also compared with other researchers who used a combination of the real driving environment dataset and general FER datasets. Jeong et al. [8] generated a realistic driving environment, KMU-FED, and used it to evaluate their model in identifying the driver's emotion with the CK+ dataset. Their model provided accuracies of 92.6% for the CK+ dataset and 94.70% for the KMU-FED dataset. Also, Zhang et al. [34] utilized the same datasets, and their model's accuracies were 93.6% and 97.3% for CK+ and KMU-FED, respectively.

Table 1. Accuracy (%) comparisons between our approach and the state-of-the-art FER approaches.

Method	JAFFE	CK+	KDEF	KMU-FED
Jeong et al. [8]	–	92.60	–	94.70
Zhang et al. [34]	–	93.60	–	97.30
Lopes et al. [16]	53.44	96.76	–	–
Yang et al. [33]	92.21	97.02	–	–
Puthanidam et al. [24]	100	–	89.58	–
Pandey et al. [23]	–	–	83.33	–
Proposed	**92.85**	**99.28**	**88.88**	**100**

5 Conclusion

In this paper, we implemented a CNN model to recognize the driver's emotions to expect his/her behavior for safe driving. We evaluated the model on four different FER datasets, three general FER datasets, and one real driving environment. It succeeds in recognizing frontal face datasets, whether static images or sequences of frames, and achieves accuracies of 92.85%, 99.28%, 88.88%, and 100% for the JAFFE, CK+, KDEF, and KMU-FED datasets, respectively. Because of the shortage of deep learning's understandability, we used two interpretability methods to illustrate the proposed model behavior with its particular architecture. In this paper, we applied our model to frontal faces. So we plan to apply it on non-frontal samples and evaluate its performance. In our future work, we will identify the driver's distraction in real-time with more datasets.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Dagher, I., Dahdah, E., Al Shakik, M.: Facial expression recognition using three-stage support vector machines. Visual Comput. Ind. Biomedi. Art **2**(1), 1–9 (2019). https://doi.org/10.1186/s42492-019-0034-5
3. Dong, Y., Hu, Z., Uchimura, K., Murayama, N.: Driver inattention monitoring system for intelligent vehicles: a review. IEEE Trans. Intell. Transp. Syst. **12**(2), 596–614 (2010)
4. Du, G., Wang, Z., Gao, B., Mumtaz, S., Abualnaja, K.M., Du, C.: A convolution bidirectional long short-term memory neural network for driver emotion recognition. IEEE Trans. Intell. Transp. Syst. 1–9 (2020)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
6. Guettas, A., Ayad, S., Kazar, O.: Driver state monitoring system: a review. In: Proceedings of the 4th International Conference on Big Data and Internet of Things. BDIoT 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3372938.3372966

7. Harel, J., Koch, C., Perona, P.: Saliency map tutorial (2012). https://www.techylib.com/en/view/skillfulwolverine/saliency_map_tutorial
8. Jeong, M., Ko, B.C.: Driver's facial expression recognition in real-time for safe driving. Sensors **18**(12), 4270 (2018)
9. Kandeel, A.A., Rahmanian, M., Zulkernine, F.H., Abbas, H., Hassanein, H.S.: Facial expression recognition using a simplified convolutional neural network model. In: 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA 2020)(Accepted). Sharjah, United Arab Emirates (2020)
10. Knapton, S.: Which emotion raises the risk of a car crash by nearly 10 times? February 2016. https://www.telegraph.co.uk/news/science/science-news/12168472/Which-emotion-raises-the-risk-of-a-car-crash-by-nearly-10-times.html?
11. Ko, B.C.: A brief review of facial emotion recognition based on visual information. Sensors **18**(2), 401 (2018)
12. Kowalczuk, Z., Czubenko, M., Merta, T.: Emotion monitoring system for drivers. IFAC-PapersOnLine **52**(8), 200–205 (2019)
13. Krishna, M.G., Srinivasulu, A.: Face detection system on Adaboost algorithm using HAAR classifiers. Int. J. Mod. Eng. Res. **2**(5), 3556–3560 (2012)
14. Lee, K.W., Yoon, H.S., Song, J.M., Park, K.R.: Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. Sensors **18**(4), 957 (2018)
15. Li, K., Jin, Y., Akram, M.W., Han, R., Chen, J.: Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Visual Comput. **36**(2), 391–404 (2019). https://doi.org/10.1007/s00371-019-01627-4
16. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
17. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101 (2010)
18. Lundqvist, D., Flykt, A., Öhman, A.: The karolinska directed emotional faces (kdef). CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet 91, 630 (1998)
19. Lyons, M., Kamachi, M., Gyoba, J.: The Japanese Female Facial Expression (JAFFE) Database, April 1998. https://doi.org/10.5281/zenodo.3451524
20. Mesken, J.: Determinants and consequences of drivers' emotions. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV (2006)
21. Nonis, F., Dagnes, N., Marcolin, F., Vezzetti, E.: 3d approaches and challenges in facial expression recognition algorithms–a literature review. Appl. Sci. **9**(18), 3904 (2019)
22. Nübling, M., Stößel, U., Hasselhorn, H.M., Michaelis, M., Hofmann, F.: Measuring psychological stress and strain at work-evaluation of the COPSOQ questionnaire in Germany. GMS Psycho-Social Medicine **3** (2006)
23. Pandey, R.K., Karmakar, S., Ramakrishnan, A., Saha, N.: Improving facial emotion recognition systems using gradient and Laplacian images. arXiv preprint arXiv:1902.05411 (2019)
24. Puthanidam, R.V., Moh, T.S.: A hybrid approach for facial expression recognition. In: Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, pp. 1–8 (2018)

25. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford (2001)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
27. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 60 (2019)
28. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
29. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
30. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 1, pp. I-I (2001)
31. Wang, Q., Jia, K., Liu, P.: Design and implementation of remote facial expression recognition surveillance system based on PCA and KNN algorithms. In: 2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 314–317 (2015)
32. World Health Organization: Global status report on road safety 2018: Summary. World Health Organization, Technical report (2018)
33. Yang, B., Cao, J., Ni, R., Zhang, Y.: Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access **6**, 4630–4640 (2017)
34. Zhang, J., Mei, X., Liu, H., Yuan, S., Qian, T.: Detecting negative emotional stress based on facial expression in real time. In: 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), pp. 430–434 (2019)
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)