Education and Information Technologies (2023) 28:4069–4092
https://doi.org/10.1007/s10639-022-11370-4

Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning

Imane Lasri[1] · **Anouar Riadsolh**[1] · **Mourad Elbelkacemi**[1]

Received: 27 May 2022 / Accepted: 22 September 2022 / Published online: 11 October 2022

Abstract

Nowadays, facial expression recognition (FER) has drawn considerable attention from the research community in various application domains due to the recent advancement of deep learning. In the education field, facial expression recognition has the potential to evaluate students' engagement in a classroom environment, especially for deaf and hard-of-hearing students. Several works have been conducted on detecting students' engagement from facial expressions using traditional machine learning or convolutional neural network (CNN) with only a few layers. However, measuring deaf and hard-of-hearing students' engagement is yet an unexplored area for experimental research. Therefore, we propose in this study a novel approach for detecting the engagement level ('highly engaged', 'nominally engaged', and 'not engaged') from the facial emotions of deaf and hard-of-hearing students using a deep CNN (DCNN) model and transfer learning (TL) technique. A pre-trained VGG-16 model is employed and fine-tuned on the Japanese female facial expression (JAFFE) dataset and the Karolinska directed emotional faces (KDEF) dataset. Then, the performance of the proposed model is compared to seven different pre-trained DCNN models (VGG-19, Inception v3, DenseNet-121, DenseNet-169, MobileNet, ResNet-50, and Xception). On the 10-fold cross-validation case, the best-achieved test accuracies with VGG-16 are 98% and 99% on JAFFE and KDEF datasets, respectively. According to the obtained results, the proposed approach outperformed other state-of-the-art methods.

Keywords Facial emotion recognition · Deep convolutional neural networks · Transfer learning · Deafness · Student engagement

Imane Lasri, Anouar Riadsolh and Mourad Elbelkacemi contributed equally to this work.

✉ Imane Lasri
imane_lasri@um5.ac.ma

Extended author information available on the last page of the article

1 Introduction

Facial expressions are one of the most important means for humans to express emotions and intentions without saying a word and are a form of nonverbal communication, especially for people in the deaf community, as they are used in sign language to express grammatical functions and emotions. Therefore, recognizing emotion from facial expressions has become a well-researched area. In psychology, Ekman and Friesen (1971) identified six universal emotions (happiness, sadness, disgust, fear, surprise, and anger), each with its unique facial expression that can be recognized automatically through computer vision algorithms.

Recently, automatic facial emotion recognition (AFER) has drawn the attention of the research community for its numerous applications in various fields including, medicine (Jin et al., 2020; Leo et al., 2020), security systems (Yin et al., 2017), and education. In the education field, facial emotion recognition can be used to monitor students' engagement in the classroom. Lasri et al. (2019) proposed a CNN architecture to recognize students' facial emotions in a classroom environment. ELLaban et al. (2017) also proposed a real-time system for students' facial expression recognition in the e-learning environment. Thomas and Jayagopi (2017) classified the level of students' engagement by analyzing behavioral cues from their facial expressions, head movements, and gaze behavior. Aslan et al. (2019) presented a real-time student engagement system that improves instructors' classroom practices.

One of the main research questions focused on educational data mining (EDM) is how deaf and hard-of-hearing students are engaged in a classroom. The question of engagement is significant and can affect teaching and student achievement in different learning environments, such as classrooms and massively open online courses (MOOCs). The lack of deaf and hard-of-hearing students' engagement can be caused by various reasons, including students who rely on lip-reading may not have time to process the preceding subject information when teachers don't make a pause before passing to a new subject. And teachers who don't know sign language can experience difficulties when communicating with deaf or hard-of-hearing students. Moreover, some universities are not able to provide their deaf or hard-of-hearing students with assistive technology. To improve the effectiveness of the learning process, teachers can keep track of the engagement level of each student.

Deaf and hard-of-hearing students' engagement can be evaluated using questionnaires or automated systems based on eye movement and facial emotion recognition (FER) for a better teaching pedagogy and learning experience. Extracting facial expression features from facial images and recognizing different facial expressions with a trained classifier is the major task of facial emotion recognition (FER). The classical FER techniques consist of three main steps: image preprocessing, feature extraction, and emotion recognition. In the preprocessing step, the face region is detected and then cropped from the input image. Subsequently, reducing to eliminating noise, scaling, resizing, and normalization are performed on the face image. The feature extraction step from the processed image is a significant stage, which consists of finding various spatial and temporal features from the facial components. Finally, traditional machine learning (ML) methods and deep learning (DL) methods classify the input image using the extracted features to understand emotions. The traditional machine learning (ML) methods aim to detect the face region in the image and extract features, then classify the input image using the extracted features. While the deep learning (DL) methods, especially convolutional neural networks (CNNs) and deep convolutional neural networks (DCNN), perform the FER task by combining feature extraction and classification steps in its single composite operational process. Other deep learning approaches include pre-trained DCNN networks, such as VGG-16 (Simonyan et al., 2015), VGG-19, Inception v3 (Szegedy et al., 2015), Xception (Chollet et al., 2017), Resnet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), DenseNet-169, and MobileNet (Howard et al., 2017) which reduce the long training process by using pre-trained weights.

To the best of our knowledge, no studies have been conducted, to date, on detecting the deaf and hard-of-hearing students engagement using machine learning or deep learning algorithms. In the present study, we attempt to address this challenge, for the first time, by proposing an automatic system that evaluates the deaf and hard-of-hearing students engagement from their facial expressions based on a deep convolutional neural network (DCNN) and transfer learning (TL). The facial images of students are obtained using a camera in the classroom. This system can help teachers observe the reaction of deaf or hard-of-hearing students on a particular topic during a lecture, adjust the teaching methodology according to students' comprehension, identify deaf or hard-of-hearing students who are not engaged and need academic support sessions in order to improve classroom management and save time and resources. An ImageNet pre-trained VGG-16 model was employed in the proposed FER model by replacing its upper layers with other dense layers, dropout layers, and batch normalization layers. Next, we fine-tuned the VGG-16 model on two facial image datasets: the Japanese female facial expression (JAFFE) (Lucey et al., 2010) and the Karolinska directed emotional faces (KDEF) (Calvo & Lundqvist, 2000). Then, the performance of the proposed model is evaluated and compared to seven different pre-trained DCNN models (VGG-19, Inception v3, DenseNet-121, DenseNet-161, MobileNet, ResNet-50, Xception) on JAFFE and KDEF datasets.

The overall objectives of this study can be outlined as follows:

- We propose a novel system that detects the engagement of deaf and hard-of-hearing students from their facial emotions based on deep convolutional neural networks (DCNN) and transfer learning (TL).
- To evaluate the performance of the facial emotion model, we tested different model optimizers and different popular pre-trained DCNN models on JAFFE and KDEF datasets.
- To monitor deaf and hard-of-hearing students' engagement, a result visualization is provided by our system in real-time.

The rest of this article is described as follows: In Section 2 we present the related works. Section 3 contains an overview of CNN, VGG-16 model, and transfer learning followed by a description of the proposed method. Section 4 discuss the experimental results. Finally, Section 5 presents the conclusion and the future extensions of our work.

2 Related work

2.1 Facial expression recognition

Several techniques have been carried out on applying machine learning and deep learning methods to analyze human facial expressions in the last few decades. Earlier works on facial emotion recognition relied on traditional machine learning methods such as support vector machines (SVM), k-Nearest Neighbor (KNN), and neural networks (NN) with differents features extraction techniques. Lee et al. (2012) used contourlet transform (CT) for feature extraction and regularized discriminant analysis-based boosting algorithm (RDAB) for classification. Their proposed approach was evaluated using the JAFFE dataset. Liew and Yairi (2015) examined five feature descriptors, including Gabor, Haar, local binary pattern (LBP), histogram of oriented gradients (HOG), and binary robust independent elementary features (BRIEF), by using several classifications algorithms such as SVM, KNN, linear discriminant analysis (LDA) and adaptive boosting (AdaBoost) on extended Cohn-Kanade (CK+) (Lucey et al., 2010), multimedia understanding group (MUG) (Aifanti et al., 2010), JAFFE, and frontal image from the KDEF dataset. The authors identified HOG as the best feature descriptor and SVM as the best classifier. HOG and SVM have been also used by Eng et al. (2019). They employed the JAFFE and the whole KDEF dataset to evaluate their method. Holder and Tapamo (2017) used the Scharr gradient operator, dimensionality reduction, and facial component extraction to add improvements to the gradient local ternary patterns (GLTP), which has been used for feature extraction. Then they used SVM for feature classification on the CK+ and JAFFE datasets.

With the great success of deep learning for image classification, many researchers employed deep learning-based methods for facial emotion recognition (FER). Zhao et al. (2015) used a deep belief network (DBN) for feature learning and extraction from facial expression images. Then, a multi-layer perceptron (MLP) model is used for emotion classification on JAFFE and CK+ datasets. A boosted deep belief network (BDBN) framework was proposed by Liu et al. (2014) that combines feature learning, feature selection, and emotion classification. The BDBN framework was evaluated using the CK+ and JAFFE datasets. Sari et al. (2021) proposed a standard CNN architecture with two convolutional-pooling layers for facial emotion recognition on CK+, JAFFE, and KDEF datasets. Hamester et al. (2015) considered a multi-channel convolutional neural network (MCCNN) architecture evaluated on the JAFFE dataset. The first channel is composed of a standard CNN. Then, the second channel uses pre-trained parameters obtained by a convolutional autoencoder (CAE), which learns Gabor-like filters. The two channels are connected with a fully-connected layer, followed by a logistic regression classifier. A hybrid CNN-RNN approach is employed by Jain et al. (2018) for facial emotion recognition. A CNN model with six convolutional layers and two fully connected layers was used primarily for feature extraction. Then, RNN was introduced to classify facial emotion using JAFFE datasets. However, to the best of our knowledge, no studies have been conducted, to date, on analyzing the facial expressions of deaf and hard-of-hearing students using machine learning or deep learning algorithms.

2.2 Engagement detection

Measuring students' engagement in the classroom is a major concern for teachers, as it positively affects the quality of education and learning. Numerous methods have been applied to detect student engagement from educational data, facial emotion recognition, and head and eye movement. Ayouni et al. (2021) proposed a system that predicts students' engagement levels (actively engaged, passively engaged, and not engaged) using support vector machine (SVM), artificial neural network, and decision tree on recorded students' activities. The system can alert the instructor when a student has a low engagement level via course messages or e-mail. Shen et al. (2022) developed a framework for assessing the students' engagement level (great, not bad, and not so well) from their facial expressions in the e-learning environment. The authors used an attentional convolutional network model for facial expression recognition.

3 Methodology

To provide a better understanding of the proposed methodology, we present in the subsections that follow, an overview of convolutional neural network (CNN), VGG-16 model, transfer learning (TL), and the proposed approach for detecting the engagement of deaf and hard-of-hearing students from their facial expressions.

3.1 Overview of convolutional neural network (CNN), deep CNN models, and transfer learning (TL)

3.1.1 Convolutional neural network (CNN)

A convolutional neural network (CNN) is a deep learning neural network, which is most commonly applied to recognize visual patterns in the input image with minimal pre-processing compared to traditional image classification algorithms. The generic CNN architecture, as shown in Fig. 1, includes different layers such as convolution layers, activation layers, pooling layers, and fully connected layers.

The convolution layer derives its name from the convolution process that preserves the spatial relationship between pixels by using small squares of input data to learn image features. It uses a kernel that moves over the input image and computes a dot product with the overlap local region aggregating the result in a feature map. Equation 1 below represents the convolution formula of a 2D image h with a 2D kernel x :

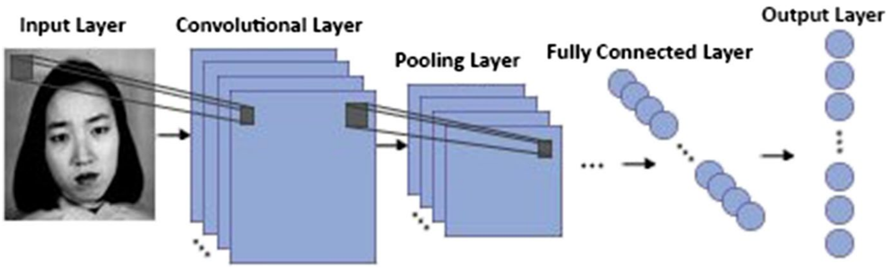


Fig. 1 Generic architecture of a convolutional neural network (CNN)

$$\begin{aligned} y[m,n] &= h[m,n] * x[m,n] \\ &= \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} h[i,j]x[m-i,n-j] \end{aligned} \tag{1}$$

where *m* and *n* are the indexes of rows and columns of the result matrix.

Each convolutional layer is followed by a nonlinear activation layer where rectified linear unit (ReLU) function, defined in (2), is the most widely used.

$$R(x) = \begin{cases} max(0,x) & ,x >= 0 \\ 0 & ,x < 0 \end{cases} \tag{2}$$

Then, the pooling layer executes a downsampling operation on the feature maps obtained after applying the ReLU activation function to reduce their dimension while keeping the important information. In particular, there are different types of pooling operations like max pooling, sum pooling, and average pooling. Finally, the pooled feature maps are flattened into a single column then a fully connected layer is used to classify the images.

3.1.2 Visual geometry group 16 (VGG-16)

With the development of deep learning, deep convolutional neural networks (DCNN) is constructed by modulating the basic CNN architecture in more depth. DCNN is considered a powerful deep learning algorithm in computer vision tasks, as it allows the automatic extraction of features of large datasets and can achieve better performance than standard neural networks. VGG 16 is a popular convolution neural network (CNN) architecture proposed by Simonyan et al. (2015) from Oxford University. To date, it is considered to be one of the best vision model architecture and was the winning model of the 2014 ILSVR (ImageNet) competition. The VGG-16 architecture, as shown in Fig. 2, consists of 13 convolutional layers of a (3x3) filter with stride 1, five max-pooling layers with stride 2, two fully connected layers with 4096 channels each followed by another fully connected layer with 1000 channels, and the final layer is the softmax layer.

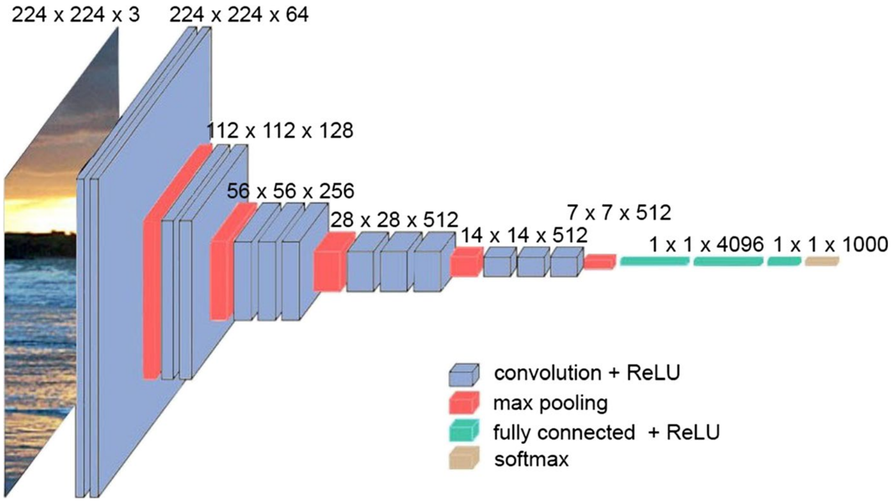


Fig. 2 VGG-16 network architecture

3.1.3 Transfer learning (TL)

Transfer learning (TL) is a machine learning method where a pre-trained model is reused on a new problem. The original trained model usually needs a high generalization to adapt to unseen data. Transfer learning means that training won't need to be restarted from scratch for every new task, it simply applies a pre-trained model that is usually trained on a huge dataset like ImageNet and the obtained weights from this model can be employed for any other task. Training deep convolutional neural network models may take weeks on very large datasets as they have many parameters to tune. Thus, transfer learning can be very useful to solve this issue because it can save training time and resources, which is the main motivation behind this research.

3.2 Proposed system architecture for emotion recognition and engagement detection

In this subsection, we introduce a novel system for real-time engagement detection from facial expressions of the deaf and hard-of-hearing students using deep convolutional neural network (DCNN) and transfer learning (TL), as illustrated in Fig. 3. The system can be used in the classroom environment to assist the teachers for understanding the attention and engagement of the deaf and hard-of-hearing students with the learning material. The students' images are automatically analysed by the system to evaluate their state of concentration from facial expressions using a web camera.

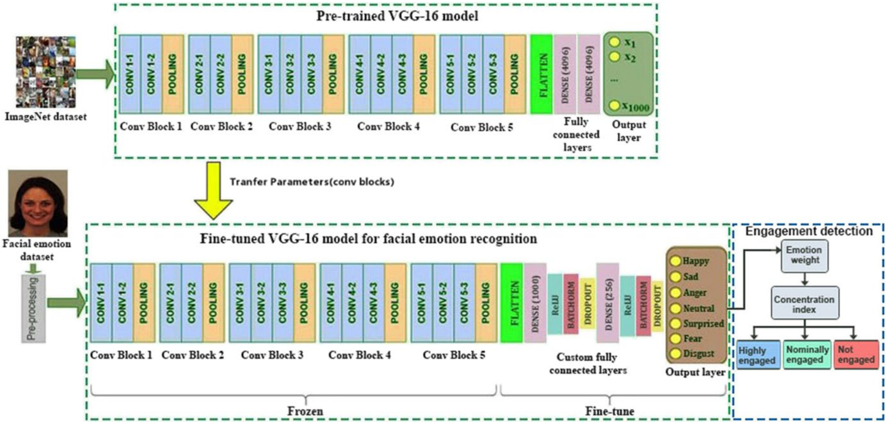


Fig. 3 Overall architecture of the proposed system for emotion recognition and engagement detection of deaf and hard-of-hearing students

3.2.1 Facial emotion recognition

In the facial emotion recognition phase, the pre-trained VGG-16 model described in Section 3.1.2 with transfer learning is used to identify the dominant emotion expressed by the deaf and hard-of-hearing students' faces at each moment. Fine-tuning is a commonly used technique for transfer learning where the weights and learning of the pre-trained model are used as an initializer for a new task. This technique is much faster than training the whole model from scratch and can help reduce the risk of overfitting. There are three most used ways to fine-tune a model: train the entire model, freeze some layers and train the others, and freeze the convolution base.

First, we froze the Conv blocks of the pre-trained VGG-16 model so that their weights don't get updated in each epoch. Then, we replaced the last dense layers with new dense layers to classify a facial image into one of seven basic emotion classes (fear, anger, happy, surprised, sad, neutral, and disgust). The channel size of the new dense layers is 1000 and 256, respectively. The rectified linear activation function (RELU) is included after the added fully connected layers, followed by batch normalization layers and dropout layers with a probability of p = 0.5 to avoid overfitting. Further, the output layer uses the Softmax activation function and the categorical cross-entropy as a loss function for multi-class classification as shown in (3) and (4).

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{3}$$

$$CE = -\sum_{i}^{C} t_i \log\left(f(s)_i\right) \tag{4}$$

where C represents the number of different classes, the subscript i denotes the i^{th} class, t_i is the truth label, S_i is the Softmax probability for the i^{th} class, and z_i is the predicted score for the i^{th} class.

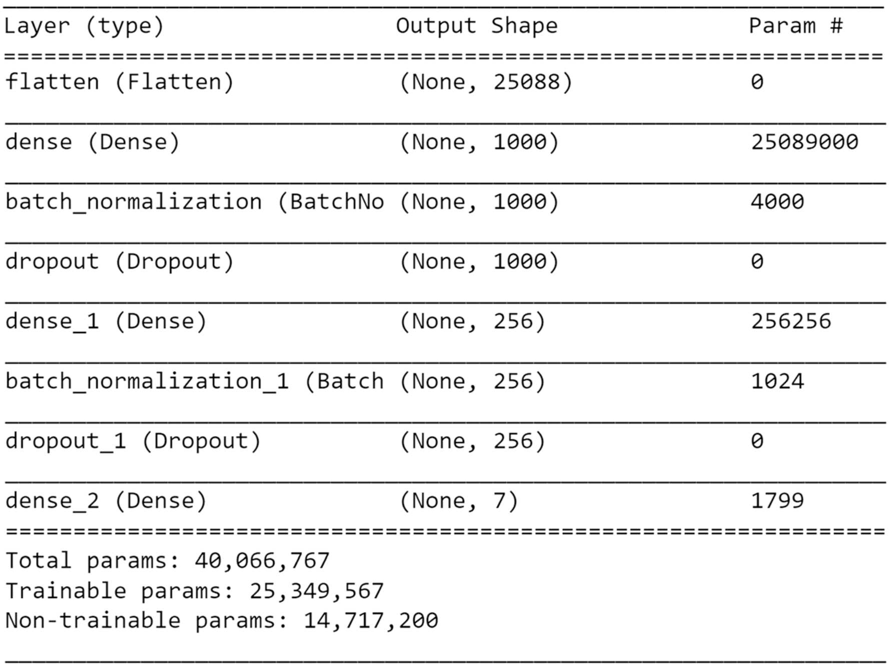
Stochastic gradient descent (SGD) with Nesterov's momentum, defined in (5), has been used as the model optimizer with learning rate 0.01 and Nesterov's momentum 0.9 to generate better performance and faster results. We set the batch size to 16 and the number of epochs to 150. Figure 4 shows a summary of our VGG-16 model fine-tuning, and Table 1 describes the details of the model's hyperparameters.

$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1}) \\ \theta &= \theta - v_t \end{aligned} \tag{5}$$

where v is the velocity and γ is usually set to 0.9.

3.2.2 Engagement detection

The recognized facial emotions are used to detect the concentration level of deaf and hard-of-hearing students. The resulting concentration index (CI) is defined by multiplying the probability of dominant emotions probability (DEP) by the related emotion weights (EW), as shown in (6).



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
dense (Dense)                (None, 1000)              25089000
_________________________________________________________________
batch_normalization (BatchNo (None, 1000)              4000
_________________________________________________________________
dropout (Dropout)            (None, 1000)              0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               256256
_________________________________________________________________
batch_normalization_1 (Batch (None, 256)               1024
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 1799
=================================================================
Total params: 40,066,767
Trainable params: 25,349,567
Non-trainable params: 14,717,200
_________________________________________________________________
```

Fig. 4 Model summary of the proposed VGG-16 fine-tuning

Table 1 VGG-16 model's hyperparameters

Parameter	Value
Input shape	(224 × 224 x 3)
Optimizer	SGD with Nesterov momentum
Momentum	0.9
Learning rate	0.01
Batch size	16
Number of epochs	150
Loss function	Categorical cross-entropy
Classifier	Softmax
Dropout rate	0.5
Bach normalization	Yes
Activation function	ReLU

$$CI = DEP \times EW \tag{6}$$

Emotion weight is a value between 0 and 1 that determines the concentration degree of a facial emotion at a given time, as shown in Table 2.

According to the obtained concentration index, the deaf or hard-of-hearing student's level of engagement (highly engaged, nominally engaged, or not engaged) is evaluated by Table 3. It can be used by teachers to adjust the lesson accordingly. Teachers can also have a real-time engagement report of each deaf or hard-of-hearing student, which can help them understand the classroom knowledge more pertinently.

4 Experimental results and discussion

To evaluate the effectiveness of the proposed FER approach on two datasets. First, a description of these datasets and image pre-processing and augmentation are presented. Then the experimental setup and the results of the proposed model, followed by a results comparison with prior studies.

Table 2 Emotion weight

Emotion	Neutral	Happy	Surprised	Sad	Fear	Anger	Disgust
Emotion weight (EW)	0.9	0.6	0.6	0.3	0.3	0.25	0.2

Table 3 Engagement detection from concentration index (CI)

Engagement type	Concentration index (CI)
Highly engaged	> 50%
Nominally engaged	20-50%
Not engaged	< 20%

4.1 Facial expression datasets

Two benchmark facial expression datasets were used to evaluate our proposed approach: the Japanese Female Facial (JAFFE) and the Karolinska Directed Emotional Faces (KDEF) datasets. Images of the datasets are labeled into seven basic emotion classes: happiness, fear, sadness, neutral, disgust, anger, and surprise. The brief description and selection reasons of the datasets used are given below.

4.1.1 Japanese Female Facial Expression (JAFFE) dataset

The JAFFE (Lyons et al., 1998) dataset includes 213 grayscale facial expression images of 10 different Japanese female models that were taken at the psychology department at Kyushu University. Each model performed seven basic emotions (30 angry, 29 disgust, 33 fear, 30 happiness, 31 sad, 30 surprises, and 30 neutral) in which each expression contains 3 to 4 images per model, as shown in Fig. 5. The images are in .tiff format with a resolution of 256×256 pixels. We chose this dataset to prove the possibility of training a small dataset using deep convolutional neural network (DCNN) models.

4.1.2 Karolinska Directed Emotional Faces (KDEF) dataset

The KDEF (Calvo and Lundqvist, 2008) dataset is created by Karolinska institute, department of clinical neuroscience, section of psychology, Stockholm, Sweden. The dataset is a collection of 4900 images of 70 individuals (35 females and 35 males) showing seven emotional states photographed twice from 5 different angles (full-left profile, half-left profile, straight, half-right profile, and full-right profile), as shown in Fig. 6. The images are in RGB format with a resolution of 562×762 pixels. Different criteria were applied for the actors' selection and the picture-taking procedure, such as the ages between 20 and 30 years, the absence of facial hair, earrings or eyeglasses, and visible make-up during the photo session. Facial expression recognition on the KDEF dataset is challenging for profile views, especially for full-left or full-right profile views, as only one side of the face with one ear and eye is visible. Hence, we examined the whole dataset in the present study to evaluate the performance of the proposed method for these challenging cases.

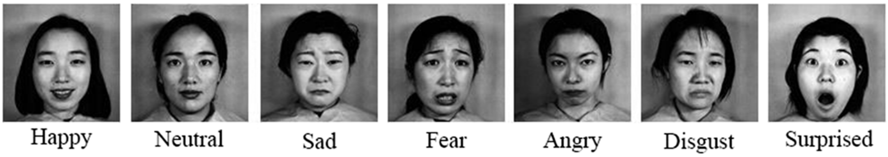


Fig. 5 Sample facial expression images from JAFFE dataset

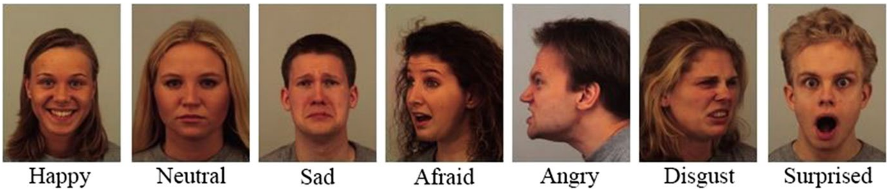


Fig. 6 Sample facial expression images from KDEF dataset

4.2 Image pre-processing and augmentation

The following image pre-processing steps are considered in our work to format images before they are used by DCNN models. First, the Haar feature-based cascade classifier (Viola & Jones, 2001) from OpenCV (Bradski, 2000) is used to detect and crop faces from each image. It is an effective machine learning based-approach, in which a cascade function is trained using a lot of positive and negative images to recognize the face region in the image. Second, the face regions were resized into 224 × 224 pixels, which is the default input dimension of pre-trained DCNN models. Finally, we applied some image augmentation techniques to avoid overfitting and classify the unseen data very accurately. Each image is horizontally flipped and rotated with an angle of (− 10° to 10°). No data augmentation was made to the images in the test set.

4.3 Experimental setup

The facial emotion recognition of deaf and hard-of-hearing students model has been written in the python programming language using Keras (Chollet, 2015) with TensorFlow backend (Abadi et al., 2016) for the image classification and OpenCV for image pre-processing. The experiments were performed on a PC with Nvidia GTX 1070, Intel Core i7, 16 GB RAM, CPU of 3.20 GHz in a 64-bit Windows 10 environment.

In this study, we used two different modes to split the JAFFE and KDEF datasets into training and testing: (i) 80% of images are used for training, and 20% of images are used as a test set. (ii) a 10-Fold Cross-Validation (CV), in which the whole datasets are randomly partitioned into ten parts, nine parts are used for training the model, and the rest is used for testing.

4.4 Experimental analysis and comparison

In this section, the performance of our proposed model is evaluated using different types of optimization algorithms on two comprehensive benchmark facial expression datasets: JAFFE and KDEF. Then, the obtained optimized model performance is compared with different pre-trained DCNN models. The optimizers like Stochastic gradient descent (SGD) (Robbins & Monro, 1951), SGD with momentum (Qian, 1999), SGD with Nesterov's momentum (Nesterov, 1983), Adam (Kingma & Ba, 2014), Adagrad (Duchi et al., 2011), and Adadelta (Zeiler, 2012) are used in the present work to obtain optimized model performance. Figures 7 and 8 show a comparison of the train loss evolution with various optimization algorithms through the epochs on the JAFFE and KDEF datasets with a batch size of 16 and categorical cross-entropy as loss function.

Stochastic gradient descent (SGD) with Nesterov momentum has shown the best recognition accuracy of 97.7% and 86.33% on JAFFE and KDEF datasets, respectively. Moreover, it is observed from Figs. 7 and 8 that the SGD with Nesterov momentum has the lowest training loss among all optimization algorithms whereas Adadelta provides the highest training losses. Table 4 presents the recognition accuracy of our proposed model with different optimizers using an 80–20% split validation scheme.

Figures 9 and 10 illustrate the accuracy and loss graphs of the JAFFE and KDEF dataset training and testing phases. Then, Figs. 11 and 12 show the confusion matrices for JAFFE and KDEF datasets validated with an 80–20% split validation sch eme.

A specific set of performance metrics were considered to provide additional analysis of our approach: precision, accuracy, recall, and F1-score. Corresponding formulas regarding each of these metrics are defined in (7), (8), (9), and (10), where TP (resp. TN) stands for true positive (resp. negative) and FP (resp. FN) for false positive (resp. negative).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

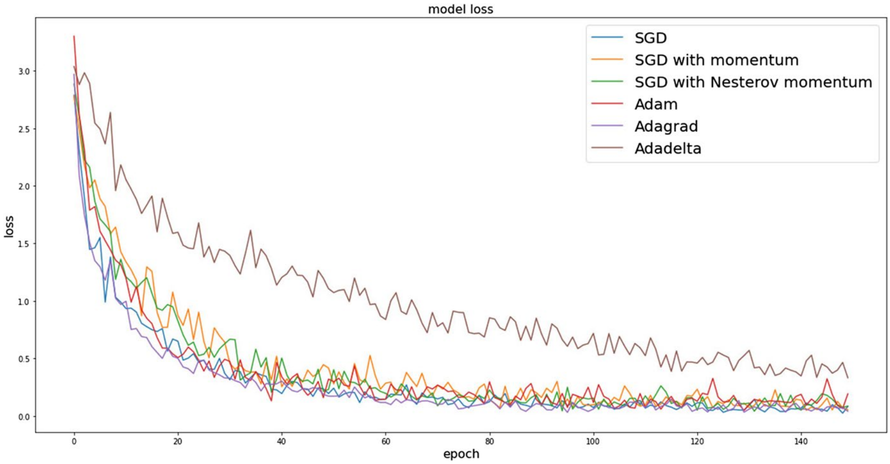


Fig. 7 Training loss comparison of all optimizers using the fine-tuned VGG-16 model on JAFFE dataset

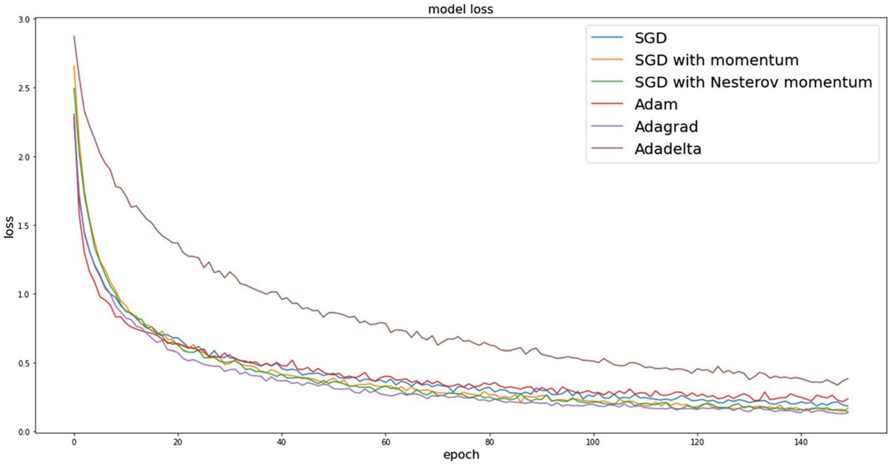


Fig. 8 Training loss comparison of all optimizers using the fine-tuned VGG-16 model on KDEF dataset

Table 4 Recognition accuracy of our proposed model on JAFFE and KDEF datasets using different optimizers

	Recognition accuracy	
Optimizer	JAFFE	KDEF
SGD	95.35%	85.50%
SGD with momentum	95.35%	85.90%
SGD with Nesterov’s momentum	97.70%	86.33%
Adam	95.35%	83.70%
Adagrad	95.35%	85.80%
Adadelta	90.70%	82.75%

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = 2 \times \frac{(P\times R)}{(P+R)} \tag{10}$$

Table 5 gives the overall performance of the proposed model on the JAFEE and KDEF datasets. The test accuracies of the proposed model using schemes of split (80–20%) and 10-fold CV on JAFFE are 97.7% and 98%, respectively. Similarly, on the KDEF dataset, the test accuracies are 86.33% and 99% using

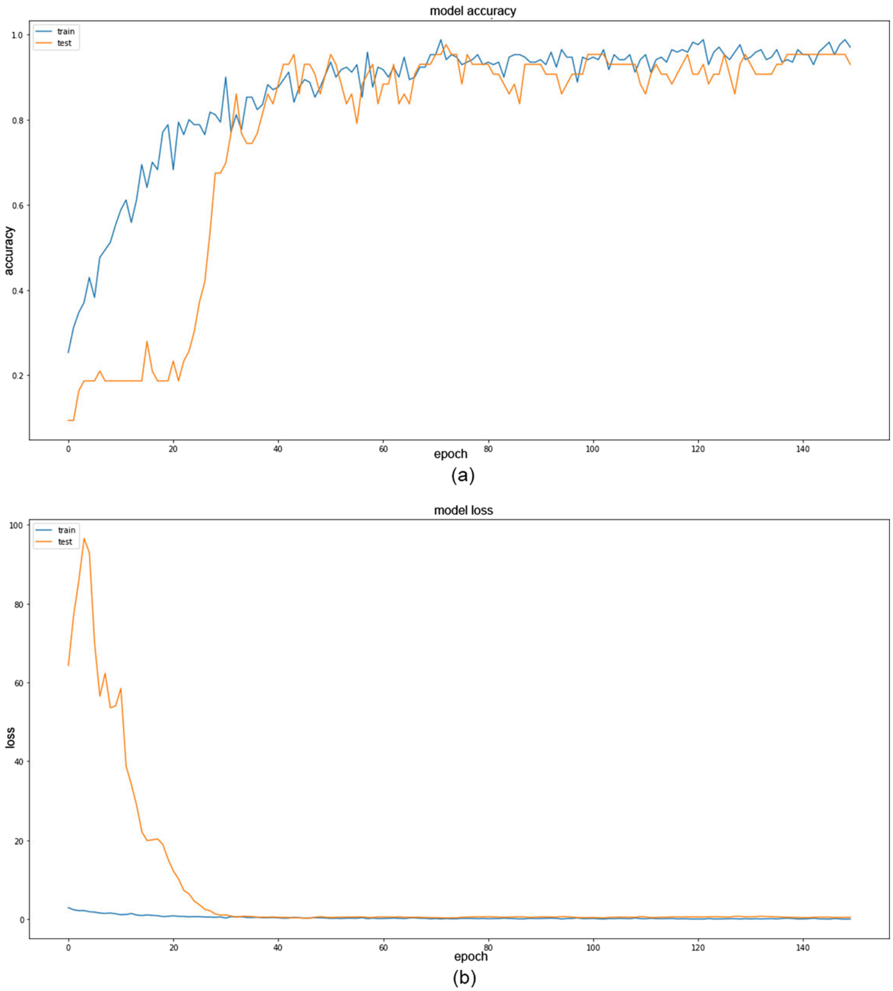


Fig. 9 Training and testing accuracy (a) and loss (b) of the fine-tuned VGG-16 model on JAFFE dataset

schemes of split (80–20%) and 10-fold CV, respectively. It should be noticed that the size of the KDEF dataset is much larger than JAFFE and contains multiple views of the face. Hence, the recognition accuracy of the JAFFE dataset is higher than KDEF.

The proposed FER method is examined for eight differents pre-trained DCNN models: VGG-16, VGG-19, Inception v3, DenseNet-121, DenseNet-169, MobileNet, ResNet-50, and Xception, as shown in Table 6. The experiments were performed using two different splitting modes for JAFFE and KDEF datasets: 20% for testing

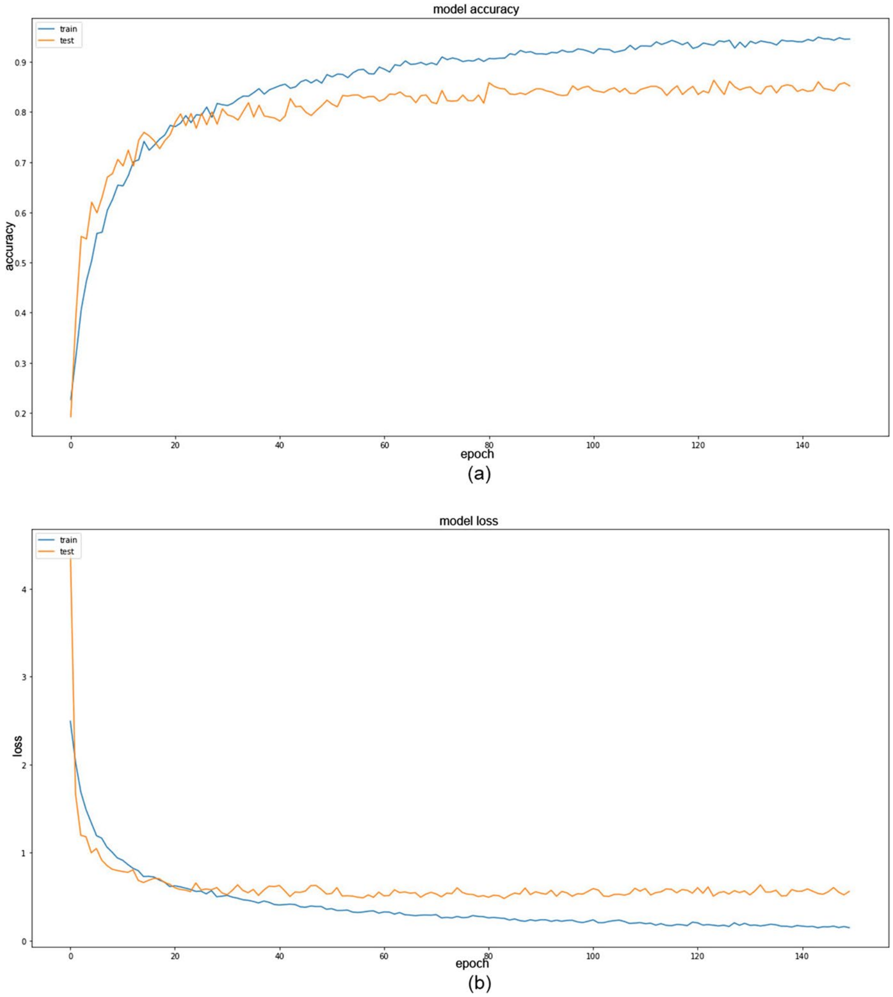


Fig. 10 Training and testing accuracy (a) and loss (b) of the fine-tuned VGG-16 model on KDEF dataset

(i.e., 80% for training) and a 10-Fold Cross-Validation (CV). The results show that the VGG-16 model achieved the best classification accuracies of 97.7% and 86.33% for JAFFE and KDEF datasets on the selected 20% test data case, followed by VGG-19 with the same accuracy of 97.7% for JAFFE and 83.40% for KDEF.

ResNet-50 presented the worst results with an accuracy of 83.72% for JAFFE and 50.4% for KDEF. Moreover, on the 10-Fold Cross-Validation case, accuracy varied from 97% to 99% for the JAFFE dataset and 81% to 99% for the KDEF dataset. The VGG-16 model achieved the best accuracies of 98% and 99% for JAFFE and KDEF datasets, respectively.

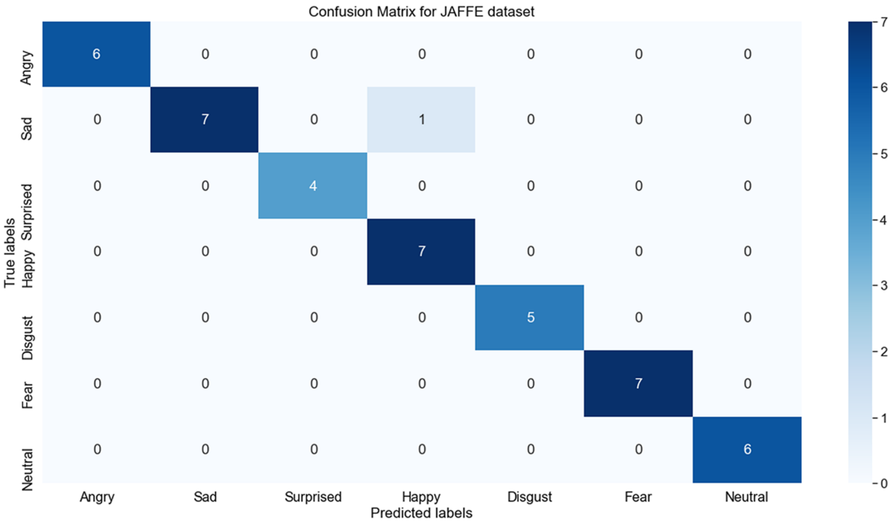


Fig. 11 Confusion matrix of the fine-tuned VGG-16 model on JAFFE dataset

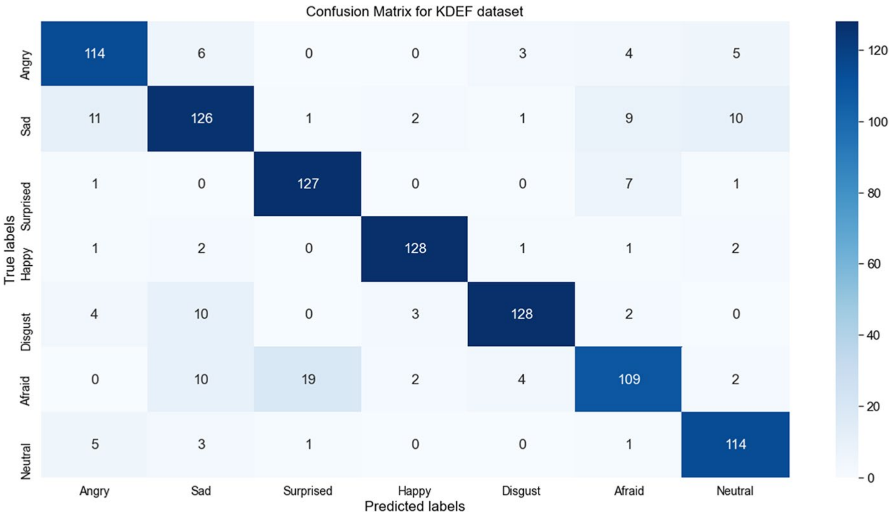


Fig. 12 Confusion matrix of the fine-tuned VGG-16 model on KDEF dataset

4.5 Result visualization of facial emotion recognition and engagement detection of deaf and hard-of-hearing students

Our proposed model has been tested on images of seven deaf and hard-of-hearing students from the faculty of sciences Rabat, Mohammed V University in Rabat, who participated in the experiment (4 males and 3 females) there of them were

Table 5 Precision, Recall, and F1-score of our proposed model on JAFFE and KDEF datasets

	JAFFE			KDEF		
Emotion type	Precision	F1-score	Recall	Precision	F1-score	Recall
Angry	1.00	1.00	1.00	0.84	0.84	0.84
Sad	1.00	0.93	0.88	0.80	0.79	0.77
Surprised	1.00	1.00	1.00	0.87	0.89	0.90
Happy	0.88	0.93	1.00	0.98	0.97	0.96
Disgust	1.00	1.00	1.00	0.95	0.89	0.84
Afraid	1.00	1.00	1.00	0.77	0.78	0.79
Neutral	1.00	1.00	1.00	0.86	0.90	0.96

Table 6 Comparison of the accuracies with different pre-trained deep CNN models on JAFFE and KDEF datasets

	JAFFE		KDEF	
Pre-trained DCNN Model	20% test	10 Kfold CV	20% test	10 Kfold CV
VGG-16	97.70%	98% ± 0,07	86.33%	99% ± 0.03
VGG-19	97.70%	98% ± 0.04	83.40%	98% ± 0.03
Inception v3	86.04%	97% ± 0.08	72.34%	86% ± 0.02
DenseNet-121	93.02%	98% ± 0.07	80.10%	90% ± 0.02
DenseNet-169	93.02%	98% ± 0.07	79.30%	96% ± 0.03
MobileNet	95.35%	97% ± 0.08	82.70%	92% ± 0.01
ResNet-50	83.72%	98% ± 0.06	50.40%	81% ± 0.05
Xception	86.04%	98% ± 0.07	69.40%	90% ± 0.03

three wearing glasses and two in a half-left profile view, as shown in Fig. 13. It can be seen clearly from the figure that all the faces were recognized and marked by the red rectangular outlines. Then, each emotion is represented with black text, and the red bar represents its probability. The dominant emotion label with the maximum value is represented with red text at the top of each rectangle. Subsequently, the engagement type calculated from the dominant emotion of each deaf and hard-of-hearing student is represented with white text at the bottom of each rectangle. Of the total seven faces, five were labeled "happy" and two were labeled "neutral". In addition, six students were nominally engaged and one student was highly engaged. Figure 14 shows the percentage of the engagement level of the deaf and hard-of-hearing students in the classroom. According to the achieved results, the proposed approach has shown remarkable performance in evaluating the facial expressions and the engagement of deaf and hard-of-hearing students in a classroom environment.

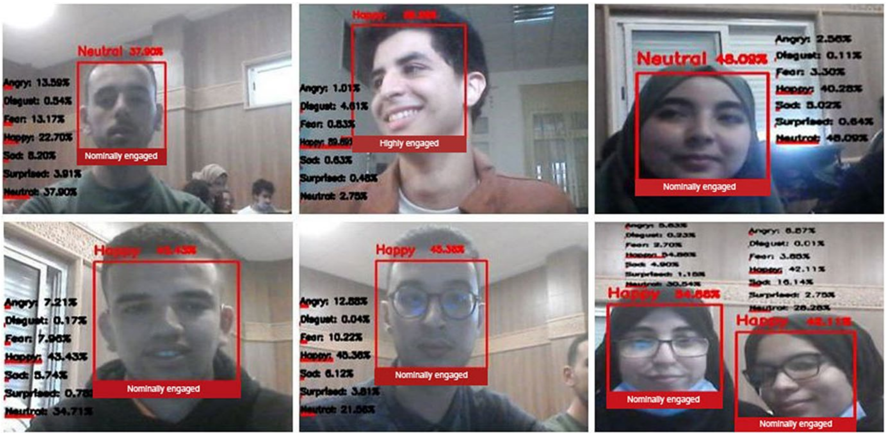
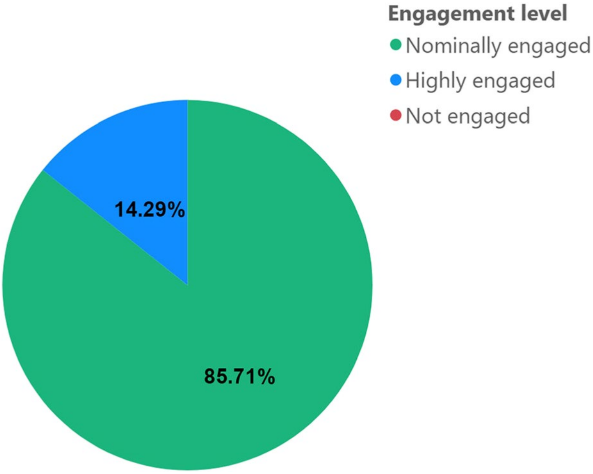


Fig. 13 Facial emotion recognition and engagement detection of deaf and hard-of-hearing students

Fig. 14 Engagement level of the seven deaf and hard-of-hearing students in the class



4.6 Performance comparison with prior studies

In this section, the performance of the proposed FER method is compared with other state-of-the-art methods on JAFFE and KDEF datasets. Table 7 presents the methods used in prior studies to recognize facial emotions, publication year, the total of the sample of the two datasets, the data splitting method, and the test accuracy. In the 10-Fold CV case, the multi-layer perceptron (MLP) and deep belief networks (DBNs) approach (Zhao et al., 2015) gave 90.95% accuracy for JAFFE. However, the proposed approach achieved the best recognition accuracy of 98% for JAFFE and 99% for KDEF.

Table 7 Comparison between the proposed method and existing methods on JAFFE and KDEF datasets

	Total of samples			Test accuracy (%)	
Methods	JAFFE	KDEF	Data splitting method	JAFFE	KDEF
Contourlet transform and RDA-based boosting algorithm (Lee et al., 2012)	210		LOOCV	96.43%	
Boosted deep belief network (BDBN) (Liu et al., 2014)	213		LOSO-CV	91.80%	
Multi-layer perceptron (MLP) and Deep belief networks (DBNs) (Zhao et al., 2015)	213		10-Fold CV	90.95%	
Histogram of oriented gradients (HOG) and SVM (Liew & Yairi, 2015)	213	980	90-10%	89.50%	80.20%
CNN and Convolutional Autoencoder (CAE) (Hamester et al., 2015)	213		LOOCV	95.80%	
Improved gradient local ternary pattern (IGLTP) and SVM (Holder & Tapamo, 2017)	213		LOOCV	84.50%	
Hybrid CNN and RNN network (Jain et al., 2018)	213		70-30%	94.91%	
Histogram of oriented gradients (HOG) and and SVM (Eng et al., 2019)	213	4900	70-30%	76.19%	80.95%
CNN (Sari et al., 2021)	213	490	80-20%	86.24%	82.38%
Our proposed approach with VGG-16	213	4900	80-20%	97.70%	86.33%
based transfer learning model			10-Fold CV	98%	99%

For the 80–20 split, the proposed method shows an accuracy of 97.7% and 86.33% for JAFFE and KDEF outperforming the CNN method (Sari et al., 2021), which gives an accuracy of 86.24% for JAFFE and 82.38% for KDEF, respectively. The proposed approach with VGG-16 based transfer learning model outperformed any other state-of-the-art method for JAFFE and KDEF datasets.

5 Conclusion

One of the biggest challenges in education is having a system that detects the engagement of deaf and hard-of-hearing students. In this work, we proposed a novel approach for evaluating deaf and hard-of-hearing students engagement from their facial emotions captured by the camera in a classroom environment using a VGG-16 based transfer learning model with a fine-tuning strategy. Three different engagement levels are produced by our system: ‘highly engaged’, ‘nominally engaged’, and ‘not engaged’.

Our research compared the influence of different optimization algorithms on model performance and conducted comparison analysis with eight different pretrained DCNN models using two validation schemes split (80–20%) and 10-fold CV on JAFFE and KDEF datasets. It turned out that the VGG-16 model achieved the best classification accuracy of 97.7% and 86.33% for JAFFE and KDEF on 20% of test samples. Similarly, in the 10-Fold CV case, the VGG 16 model achieved the best classification accuracy of 98% and 99% for JAFFE and KDEF datasets. Moreover, the SGD with Nesterov's momentum has the lowest training loss compared with the other optimization algorithms. According to the obtained results, the proposed method outperformed other state-of-the-art methods and has proven to be successful in recognizing deaf and hard-of-hearing students engagement based on facial emotions in a classroom environment.

The proposed system can help teachers to adapt the teaching material based on the engagement level of each deaf or hard-of-hearing student. We have tested our system with seven deaf and hard-of-hearing students in a classroom enviromnent. The results reveal that the proposed system correctly identifies the students engagement from facial expressions. In future research, we will evaluate the engagement from more features such as gaze behavior, and body movements to improve the performance of classroom teaching methods for deaf and hard-of-hearing students.

Acknowledgements We wish to thank the seven deaf and hard-of-hearing students for their participation in the experiment.

Author contributions All authors contributed to the conception and design of the study. Imane Lasri performed the material preparation, data collection, literature survey, analysis and wrote the paper. Anouar Riadsolh and Imane Lasri conducted the experiment. Mourad Elbelkacemi and Anouar Riadsolh participated in manuscript revision and all authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The JAFFE and KDEF datasets were used to support this study and there application details are in https://zenodo.org/record/3451524#.Yc1-OGjMLIU and https://www.kdef.se/download-2/register.html. The datasets are cited at relevant places within the text as references.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283).

Aifanti, N., Papachristou, C., & Delopoulos, A. (2010). The mug facial expression database. In *Proceedings of the 11th international workshop on image analysis for multimedia interactive services (WIAMIS)* (pp. 1–4). Desenzano del Garda, Italy: IEEE.

Aslan, S., Alyuz, N., Tanriover, C., Mete, S., Okur, E., D'Mello, S., & Arslan Esme, A. (2019). Investigating the impact of a real-time, multi- modal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 conference on human factors in computing systems (chi)*. https://doi.org/10.1145/3290605.3300534 (pp. 1–12). Glasgow Scotland, UK: ACM.

Ayouni, S., Hajjej, F., Maddeh, M., & Al-Otaibi, S. (2021). A new ml-based approach to enhance student engagement in online environment. *PLoS ONE*, *16*(11), 0258788. https://doi.org/10.1371/journal.pone.0258788.

Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal of Software Tools*.

Calvo, M., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behavior Research Methods*, *40*(1), 109–115. https://doi.org/10.3758/BRM.40.1.109.

Chollet, F. (2015). *Keras: the python deep learning library*. https://keras.io. Accessed 20 March 2021.

Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions*. arXiv:1610.02357.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2). https://doi.org/10.1037/h0030377.

Ellaban, H., & Elsaeed, E. (2017). A real-time system for facial expression recognition using support vector machines and k-nearest neighbor classifier. *International Journal of Computer Applications*, *159*(8), 23–29. https://doi.org/10.5120/ijca2017913009.

Eng, S., Ali, H., Cheah, A., & Chong, Y. (2019). Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine. *IOP Conference Series: Materials Science and Engineering*, *705*(1), 012031. https://doi.org/10.1088/1757-899x/705/1/012031.

Hamester, D., Barros, P., & Wermter, S. (2015). Face expression recognition with a 2-channel convolutional neural network. In *Proceedings of 2015 international joint conference on neural networks (ijcnn)*. https://doi.org/10.1109/IJCNN.2015.7280539(pp. 1–8). Killarney, Ireland: IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of 2016 ieee conference on computer vision and pattern recognition (cvpr)*. https://doi.org/10.1109/CVPR.2016.90 (pp. 770–778). Las Vegas NV, USA: IEEE.

Holder, R., & Tapamo, J. (2017). Improved gradient local ternary patterns for facial expression recognition. *EURASIP Journal on Image and Video Processing*, *2017*, 42. https://doi.org/10.1186/s13640-017-0190-5.

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. (2017). Densely connected convolutional networks. In *Proceedings of 2017 IEEE conference on computer vision and pattern recognition (cvpr)*. https://doi.org/10.1109/CVPR.2017.243 (pp. 2261–2269). Honolulu, HI, USA: IEEE.

Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, *115*, 101–106. https://doi.org/10.1016/j.patrec.2018.04.010.

Jin, B., Qu, Y., Zhang, L., & Gao, Z. (2020). Diagnosing parkinson disease through facial expression recognition: video analysis. *Journal of Medical Internet Research*, *22*(7), e18697. https://doi.org/10.2196/18697.

Kingma, D., & Ba, J. (2014). *Adam: a method for stochastic optimization*. arXiv:1412.6980v9.

Lasri, I., Riadsolh, A., & El belkacemi, M. (2019). Facial emotion recognition of students using convolutional neural network. In *Proceedings of the third international conference on intelligent computing in data sciences (icds)* (pp. 1–6).

Lee, C., Shih, C., Lai, W., & Lin, P. (2012). An improved boosting algorithm and its application to facial emotion recognition. *Journal of Ambient Intelligence and Humanized Computing*, *3*(1), 11–17. https://doi.org/10.1007/s12652-011-0085-8.

Leo, M., Carcagni, P., Mazzeo, P., Spagnolo, P., Cazzato, D., & Distante, C. (2020). Analysis of facial information for healthcare applications: a survey on computer vision-based approaches. *Information*, *11*(3), 128. https://doi.org/10.3390/info11030128.

Liew, C., & Yairi, T. (2015). Facial expression recognition and analysis: A comparison study of feature descriptors. *IPSJ Transactions on Computer Vision and Applications*, *7*, 104–120. https://doi.org/10.2197/ipsjtcva.7.104.

Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of 2014 IEEE conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2014.233(pp. 1805–1812). Columbus, OH, USA: IEEE.

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In *Proceedings of 2010 IEEE computer society conference on computer vision and pattern recognition - work- shops (cvpr workshops)*. https://doi.org/10.1109/CVPRW.2010.5543262 (pp. 94–101). San Francisco, CA, USA: IEEE.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Proceedings of 3rd IEEE international conference on automatic face and gesture recognition*. https://doi.org/10.1109/AFGR.1998.670949(pp. 200–205). Nara, Japan: IEEE.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate o(1/k2). *Soviet Mathematics Doklady, 27*(2), 372–376.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks : The Official Journal of the International Neural Network Society*, *12*(1), 145–151. https://doi.org/10.1016/S0893-6080(98)00116-6.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*(3), 400–407. https://doi.org/10.1214/aoms/1177729586.

Sari, M., Moussaoui, A., & Hadid, A. (2021). A simple yet effective convolutional neural network model to classify facial expressions. In S. Chikhi, A. Amine, A. Chaoui, D. Saidouni, & M. Kholladi (Eds.) *Lecture notes in networks and systems*. https://doi.org/10.1007/978-3-030-58861-8_14, (Vol. 156 pp. 188–202). Springer.

Shen, J., Yang, H., & Li, J. (2022). Assessing learning engagement based on facial expression recognition in mooc's scenario. *Multimedia Systems*, *28*, 469–478. https://doi.org/10.1007/s00530-021-00854-x.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & et al. (2015). Going deeper with convolutions. In *Proceedings of 2015 IEEE conference on computer vision and pattern recognition (cvpr)*. https://doi.org/10.1109/CVPR.2015.7298594 (pp. 1–9). Boston. MA, USA: IEEE.

Thomas, C., & Jayagopi, D. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM sigchi international workshop on multimodal interaction for education (mie)*. https://doi.org/10.1145/3139513.3139514 (pp. 33–40). Glasgow Scotland, UK: ACM.

Viola, P., & Jones, M. (2001). 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)*. https://doi.org/10.1109/CVPR.2001.990517 (pp. 511–518). Kauai, HI, USA.

Yin, D., Omar, S., Talip, B., Muklas, A., Norain, N., & Othman, A. (2017). Fusion of face recognition and facial expression detection for authentication: a proposed model. In *Proceedings of the 11th international conference on ubiquitous information management and communication (imcom)*. https://doi.org/10.1145/3022227.3022247(pp. 1–8). Beppu, Japan: ACM.

Zeiler, D. (2012). *Adadelta: an adaptive learning rate method*. arXiv:1212.5701.

Zhao, X., Shi, X., & Zhang, S. (2015). Facial expression recognition via deep learning. *IETE Technical Review*, *32*(5), 347–355. https://doi.org/10.1080/02564602.2015.1017542.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Imane Lasri[1] · **Anouar Riadsolh**[1] · **Mourad Elbelkacemi**[1]

Anouar Riadsolh
a.riadsolh@um5r.ac.ma

Mourad Elbelkacemi
mourad_prof@yahoo.fr

[1] Laboratory of Conception and Systems (Electronics, Signals and Informatics), Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco