

Enhancing emotion recognition model for a student engagement use case through transfer learning

Ikram Qarbal, Nawal Sael, Sara Ouahabi

Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Casablanca, Morocco

Article Info

Article history:

Received Mar 1, 2024

Revised Nov 6, 2024

Accepted Nov 14, 2024

Keywords:

Computer vision

E-learning

Emotions recognition

Facial expressions

Student concentration

Student engagement

Transfer learning

ABSTRACT

Distance education has been prevalent since the late 1800s, but its rapid expansion began in the late 1990s with the advent of the online technological revolution. Distance learning encompasses all forms of training conducted without the physical presence of learners or teachers. While this mode of education offers great flexibility and numerous advantages for both students and teachers, it also presents challenges such as reduced concentration and commitment from students, and difficulties in course supervision for teachers. This article aims to study student engagement on distance learning platforms by focusing on emotion detection. Leveraging various existing datasets, including the Facial Expression Recognition 2013 (FER2013), the Karolinska Directed Emotional Faces (KDEF), the extended Cohn-Kanade (CK+), and the Kyung Hee University Multimodal Facial Expression Database (KMU-FED), the proposed approach utilizes transfer learning. Specifically, it exploits the large number and diversity of images from datasets like FER2013, and the high-quality images from datasets like KDEF, CK+, and KMU-FED. The model can effectively learn and generalize emotional cues from varied sources by combining these datasets. This comprehensive method achieved a performance accuracy of 96.06%, demonstrating its potential to enhance understanding of student engagement in online learning environments.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ikram Qarbal

Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sik

Hassan II University of Casablanca

Casablanca, Morocco

Email: Ikram.ql.11@gmail.com

1. INTRODUCTION

Today, distance learning represents an integral part of our lives and an irreversible step forward in the way we study. Traditional teaching methods involving physical attendance at classes and training courses are now only part of the methods adopted worldwide. The main aim of these distance-learning methods is to transmit information and disseminate knowledge everywhere, without being limited by geographical boundaries, financial challenges, or other difficulties encountered in face-to-face education. Several factors have contributed to the popularization of these modern methods, the most essential of which is the rise of the internet and information and communication technologies [1]. The advent of the internet and improved broadband access have been the main driving forces behind the expansion of distance learning. Thus, an abundance of e-learning platforms and tools have been launched in recent years, offering flexibility for learners and teachers in terms of scheduling and location. This improvement is due to the technology that has created online communication tools such as webinars, video conferencing, and discussion forums, which facilitate collaboration between students and teachers. Finally, and yet importantly, the COVID-19 pandemic [2] that

contributed directly to the growth in distance learning adoption due to school closures, prompted schools to switch to e-learning modes to ensure educational continuity [3]. All of these factors have contributed to the rapid spread of distance learning, making this approach an essential component of contemporary education. Distance learning has several advantages [4] for students and teachers alike, not least flexibility—the learner decides when and where they can attend training with a schedule and pace to suit. Distance learning also offers ease of access to the course; the learner can study in international institutions without the need to travel, which contributes to the dissemination of science and knowledge.

No one can deny that these modern forms of learning have ceded many advantages in the field of education, but they have also faced many challenges, such as the absence of direct physical contact between students and teachers, which reduces the instantaneous control of the teaching operation [5], and a lack of concentration and engagement from the student. To enhance the distance learning experience and make it more comparable to traditional classroom settings, it is essential to maintain direct communication between students and teachers. This can only be achieved by providing teachers with visibility into the state of their students, enabling them to assess the effectiveness of information reception. This connection and understanding are crucial elements for fostering effective engagement and interaction in the online learning environment.

Emotions represent an adaptive response to environmental or internal stimuli, generating behavioral, physiological, and cognitive reactions. They play a powerful role in human behavior [6]. Although the human brain controls human actions, emotional intervention is not negligible, it influences one's thoughts, actions, and interactions with others. Emotions influence not only actions but also a person's abilities: a happy person is usually motivated, committed, positive, and focused, while a sad or angry person is generally unable to concentrate or perform any task. Ekman [7] develops the study of emotions as they relate to facial expressions. Following Darwin's theory, Ekman *et al.* [8] proved that emotions and facial expressions are universal; that is, they are expressed in the same way in all cultures, countries, and origins. Griffiths [9] also summarized these emotions under six primary emotions: anger, disgust, joy, fear, surprise, and sadness. Emotions are highly relevant factors that can describe the state of a student during an online course [10], and they play a significant role in distance learning, just as they do in face-to-face learning [11]. Emotions can have both positive and negative influences on students' learning experiences. Therefore, monitoring these indicators by an intelligent system can offer teachers visibility of student engagement and course follow-up.

In this study, we propose a transfer learning-based approach to recognize students' emotions, providing teachers with valuable insights into their emotional states during distance learning courses. Our approach leverages datasets of varying quantity and quality, harnessing the strengths of each to address common challenges in emotion recognition. The system comprises two main components: first, recognizing student's emotions, and second, determining the engagement level based on the emotion and its weight. This model is developed through our proposed method, which involves training several models (convolutional neural network (CNN), visual geometry group (VGG) 19-layer network, and MobileNet) on the Facial Expression Recognition 2013 (FER2013) dataset, selecting the best-performing one, and then further fine-tuning it using the Karolinska Directed Emotional Faces (KDEF), the extended Cohn-Kanade (CK+), and Kyung Hee University Multimodal Facial Expression Database (KMU-FED) datasets.

The remainder of this paper is organized as follows. Section 2 provides an overview of related studies on emotion recognition and engagement detection in online learning environments. Section 3 presents the proposed method. Section 4 discusses the experimental results obtained after training the models and discuss the results obtained. Finally, section 5 concludes the study and discusses future works.

2. RELATED WORK

Over the last few years, significant efforts have focused on detecting students' emotions to analyze their engagement in distance learning courses. Hasnine *et al.* [12] proposed a facial recognition model to identify students attending the course remotely; then, the detected face is sent to the detection system, which is divided into two systems: emotion detection and eye movement detection, the results of which are merged into an equation that deduces the state of the student: is he or she engaged or not? The proposed emotion detection model developed on the FER2013 dataset performed 68%. Gupta *et al.* [13] presents two steps to determine student engagement: first, detecting the student's face using the Faster R-CNN object detection model trained on the WIDER FACE dataset. Emotion detection: to perform this task, the author used four different types of datasets and three different deep learning models. The experimental results show that the proposed system achieves accuracies of 89.11%, 90.14%, and 92.32% for Inception-V3, VGG19, and ResNet-50, respectively. Revina and Emmanuel [14] surveyed techniques used for FER, this study compared algorithms based on preprocessing, feature extraction, and classification. To judge the performance of these techniques, they are based on dataset analysis, complexity rates, and accuracy. As a result, for preprocessing, the ROI segmentation method gives the highest accuracy of 99%; for feature extraction, GFs have an accuracy between 82.5% and 99%, and support vector machine (SVM) classification yields an accuracy of 99% for

Japanese female facial expression (JAFEE) and CK+. Li and Lima [15] proposed a feature extraction method using the ResNet-50 deep residual network. The training was performed on a dataset collected by an experienced photographer who used a Canon digital camera to capture each subject's facial expressions ten times for 20 subjects of different ages, careers, and races, including seven types of facial emotion images: joy, sadness, fear, anger, surprise, disgust, and neutral. In the final analysis, the dataset was comprised of 700 images. This approach achieves a performance of 95.39%. Meriem *et al.* [16] found a strong relationship between emotions and student concentration. To study student emotions, four types of datasets were pooled, and four pre-trained models were used to create an emotion detection system. The following results were obtained: 85%, 86%, 87%, and 64.5% for the Xception, VGG16, VGG19, and Alexnet models, respectively. Kusuma *et al.* [17] experimented with varying data distributions, use/non-use of batch normalization, clustering layer, 61 freezing certain layers, and optimizer selection as stochastic gradient descent (SGD), Adam to propose the best combination that yields the highest performance. After 23 models, the researchers found that using an unbalanced dataset, an unfrozen layer, and an SGD optimizer provided the highest accuracy (69.40 %) as a result of FER2013. Alshamsi *et al.* [18] presents achieved 96.3% accuracy on CK+, 91.9% on the JAFEE, and 90.8% on KDEP. These high performances are achieved by feature extraction and analysis of the feature descriptors center of gravity (COG) and face landmarks using a SVM algorithm. Wang *et al.* [19] found that facial components (such as eyes, mouth, and nose) are the most influential features for perceiving the emotion expressed on the face, unlike other areas such as hair and ears. Therefore, they focused on these features in their study and considered the facial area and its components as input information to train their models. Therefore, the method they proposed combined several facial sub-regions to achieve a result of 59.97% on the static facial expressions in the wild (SFEW), 67.7% FER2013, and CK+ (99.07%). Debnath *et al.* [20] proposed a model for detecting the seven basic emotions of anger, disgust, fear, happiness, neutrality, sadness, and surprise from the behavioral aspects of a man or woman. They adopted the "ConvNet" model, which is an approach based on a combination of several techniques. They used a local binary pattern model (LBP), region-based oriented FAST, and rotated BRIEF (ORB) to extract facial characteristics. They opted for a CNN to perform classification. Using this approach, they achieved a performance of 92.05% on the JAFEE and 98.13% on the CK+ dataset. According to Kim *et al.* [21], a data standardization and cleaning technique with different FER datasets was proposed to improve the FER model system. This data normalization and cleaning technique achieved a 5% increase in validation accuracy and a 2% decrease in validation loss. The related works analysis is presented in Table 1.

Table 1. Comparison of the existing models for emotion recognition

Article	Models	Datasets	Results (%)
[12]	CNN	FER2013	68
[13]	Inception-V3	FER-2013+ CK(+) + RAF-DB+ OWN dataset	89.11
	VGG19		90.14
	ResNet-50		92.32
[14]	GF	JAFEE et CK+	82.5
	SVM		99
[15]	ResNet-50	A data set of 700 images collected by an experienced photographer CK & CK+, FER2013, and JAFEE.	95.39
[16]	Xception		85
	VGG16		86
	VGG19		87
	Alexnet		64.5
[17]	Fine-Tuned VGG-16 with optimizer SGD	FER2013	69.40
[18]	SVM	CK+	96.3
		JAFEE	91.9
		KDEF	90.8
[19]	CNN	SFEW	59.97
		FER2013	67.7
		CK+	99.07
[20]	CNN	JAFEE	92.05
		CK+	98.13

The literature on emotion recognition in distance learning highlights both significant advancements and ongoing challenges. Various studies have explored different methodologies, demonstrating the potential of facial recognition and deep learning models to accurately identify student emotions. However, real-world applications face many challenges [22] such as pose variation, occlusions, and ethnic diversity, which impact model robustness. Additionally, while deep learning models generally outperform traditional approaches, there is a notable gap in addressing variations in image quality and ensuring these models generalize well across

diverse student populations and classroom conditions. Some studies have utilized datasets with high-quality images but limited in number, which allows models to better represent emotions but fails to generalize due to the lack of diversity in occlusions, ethnicity, and other real-world variations as these datasets are often captured in controlled photo sessions. On the other hand, the FER2013 dataset offers a large number of diverse images but suffers from lower quality, which can hinder detection accuracy. To address these gaps, we propose a novel approach that combines the strengths of both types of datasets. By integrating high-quality images from smaller datasets with the extensive diversity of FER2013, our approach aims to enhance model performance and adaptability. This strategy will ensure reliable emotion recognition across diverse and dynamic educational environments, ultimately contributing to improved student engagement and learning outcomes. This comprehensive solution addresses the current limitations in the literature and paves the way for more effective emotion recognition in distance learning. In the following section, we will explain our proposed approach and go into detail about the development of our model.

3. METHOD

In this study, we propose a novel transfer learning-based approach to improve emotion recognition accuracy for assessing student engagement in online learning environments. Our method leverages the strengths of multiple datasets and deep learning models, structured in a systematic, multi-stage pipeline (Figure 1). First, we utilize the FER2013 dataset to pre-train a base model, establishing a foundational understanding of diverse facial expressions across a large image set. This initial training enables the model to generalize essential features for emotion recognition. Subsequently, we enhance model performance through fine-tuning, incorporating high-quality images from the KDEF, CK+, and KMU-FED datasets. This multi-dataset approach allows the model to capture nuanced expressions more accurately, improving its robustness and adaptability across various student demographics and environmental conditions.

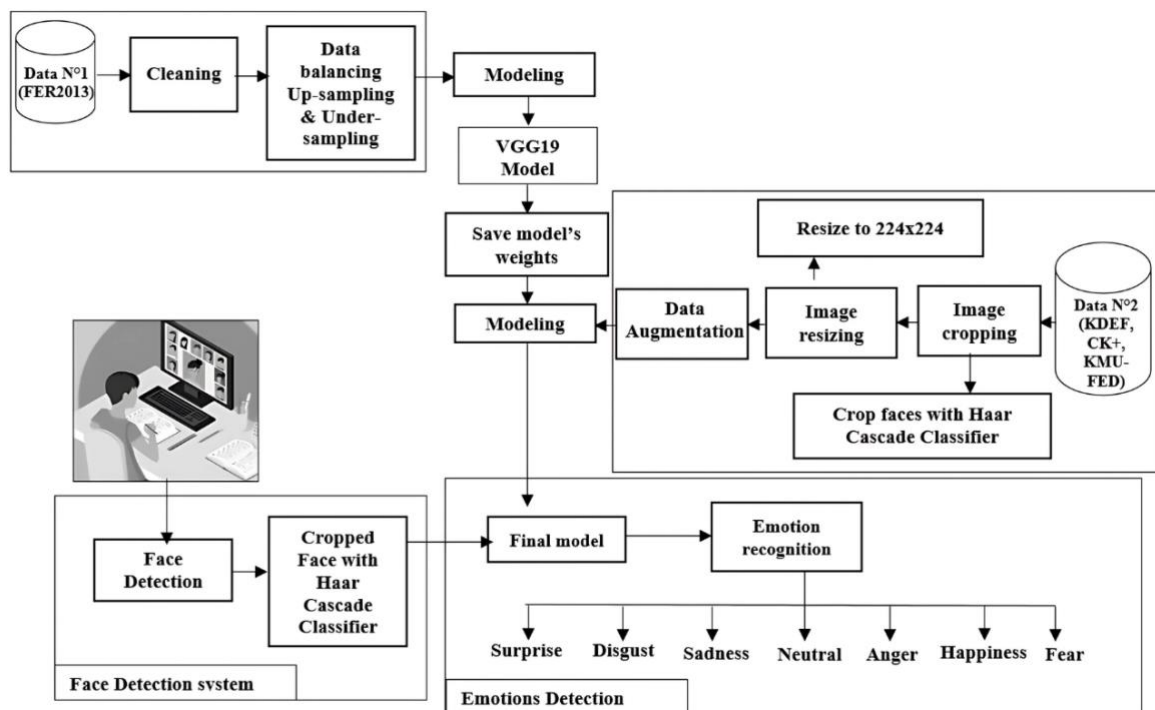


Figure 1. The approach proposed for emotion recognition of students in real-time

3.1. Datasets

To apply our proposed approach, we need different types of datasets to ensure that our model will be able to recognize all possible scenarios and overcome the challenges that our detection system may encounter, such as brightness concerns, diversity, age, and race. In our study we have chosen to work with two types of datasets, the first is FER2013. This dataset was created using Google's Image Search application programming interface (API), and the faces were automatically registered. It contains 35,887 images with a 48×48 resolution.

Faces are labeled as one of six cardinals as well as neutral expressions (0=angry, 1=disgust, 2=fear, 3=happy, 4=sad, 5=surprise, 6=neutral). It has a variation in images: facial occlusion (mainly with the hand), partial faces, low-contrast images, and images with glasses.

The second type of dataset is a combination of KDEF, CK+, and KMU-FED. These datasets are taken primarily for the emotion recognition task and are generally characterized by the clarity of facial expressions but a low quantity of images. KDEF is a set of images divided into 7 classes, with 420 images per class. The images are taken according to selection criteria, i.e., no mustache, beard, and accessories. CK+ is a dataset of seven classes, containing images of 123 adults aged between 18 and 50. KMU-FED is a set of 55 image sequences of 12 subjects captured by a near-infrared (NIR) camera installed on the dashboard or steering wheel.

3.2. Data preprocessing

Data preprocessing is a crucial step that contributes significantly to the success of the learning phase. We performed the following operations to ensure that our data is prepared for the next steps. FER2013 is a large dataset containing over 35,000 images. However, not all the images in this dataset are relevant to the task we want to accomplish. Therefore, we performed a manual cleanup covering all 7 classes of the dataset to eliminate images that would not be useful for learning and that could disrupt our model and reduce its performance. The types of images eliminated include misclassified images, images expressing emotion but classified in a different class, images containing multiple faces, images with no face at all, images with incomplete face parts, and, finally, images with writing on the face. By doing this, we ensure that each class contains images that accurately represent the emotion we want to detect. Furthermore, FER2013 is an unbalanced dataset. For example, the "disgust" class contains around 500 images, while the "happy" class contains over 8,000, which can introduce a bias and cause the model to train primarily on the majority class. To remedy this problem, we added more images to the minority classes to balance the data. The source of these added images generally comes from various other available datasets similar to FER2013, which guarantees reliable training. This pre-processing step, including manual cleaning and balancing of the dataset, was essential to prepare the data for effective training and improve the overall performance of our emotion recognition model.

KDEF, CK+, and KMU-FED contains images with backgrounds. Therefore, before we start training our model, it is important to trim these images to take only the part containing the face. To do so, we perform face detection using the HAAR cascade technique, a machine-learning object detection algorithm widely used to identify objects or features in images or videos [23]. This algorithm is known for its speed and efficiency in detecting objects, including faces. It employs a series of simple rectangular features, known as Haar-like features, and uses a cascading structure with a series of classifiers. This allows it to eliminate non-face regions in an image. Due to its efficiency, the Haar Cascade algorithm is suitable for locating faces in our data in preprocessing. Figure 2 shows the original Figure 2(a) with background and then the transformation done Figure 2(b).



Figure 2. Images of datasets before cropping in (a) original image and after cropping in (b) cropped image

To resize all the samples to a size of (224×224), we explore various resizing methods. Cv2.INTER_LANCZOS4 function is the most suitable in our case and allows obtaining the best results. In addition, we applied data augmentation to generate additional training data. It involves applying transformations such as rotation, zoom, shear, and flip to existing data to obtain new images. The new dataset size after data augmentation is presented in Table 2.

Table 2. The distribution of images by class in KDEF, CK+, KMU-FED, and the three combined

Variable	KDEF	KMU-FED	CK+	KDEF_KMU_CK+
Angry	420	196	43	2,937
Disgust	420	120	59	2,583
Fear	420	200	25	2,859
Happy	420	210	66	3,214
Neutral	420	0	575	3,676
Sad	419	180	28	2,735
Surprise	419	200	80	3,030

3.3. Modeling

The data was divided into three sets with an 80/10/10 split: 80% for training, 10% for validation, and 10% for testing. The training set was used to train the model and learn patterns from the data. The validation set was used during training to fine-tune the model's hyperparameters and monitor for overfitting. The test set provided an unbiased evaluation of the model's performance after training was complete. The data was shuffled before splitting to ensure a uniform distribution and prevent bias. This approach ensures that the model is trained effectively and can generalize well to new, unseen data. To choose the best-performing model, we have first compared the most used and efficient models mentioned in our literature review. Three models were selected: CNN, VGG19, and MobileNetV2.

3.3.1. Convolutional neural network model

Our CNN model has four phases. The first phase of the model contains 4 convolutional layers, which start with an input layer for an image of $224 \times 224 \times 3$, and a convolution is performed on this input. This is followed by batch normalization to obtain the inputs for the next layer. In the next layer, max pooling is performed with a pool size of 2×2 . Dropout is then performed at a rate of 0.25. The first 4 layers start with convolution and end with dropout. The second phase begins with a flatten layer. This flatten layer converts the two-dimensional data into a one-dimensional array. For the third phase, we created fully connected layers. Finally, for the model output, we put in a Dense layer with the Softmax activation function and the number of classes we have (7 emotions) to perform classification.

3.3.2. VGG19 model

VGG is a deep CNN used for image classification. The VGG created it at Oxford in 2014. VGG19 is a 19-layer version of the VGG network (3 fully connected, 16 convolutional, 1 softmax, and 5 max pool layers). The input to VGG-based CNN is a 224×224 RGB image that is preprocessed by a preprocessing layer. After preprocessing, they are passed to through the weight layers of the VGG19 model to 19 weight layers and 3 fully connected layers. It comprises two fully connected layers of 4,096 channels, followed by a completely connected 1,000-channel layer to anticipate 1,000 labels. Softmax feature is used for grouping by the last FC layer. To adapt this model to our problem, we thawed the output layer and replaced it with an output that meets our needs.

3.3.3. MobileNetV2 model

For MobileNet, we chose the same hyperparameters for construction. For compilation and training, we followed the same approach as for the VGG19 model [24]. Mobile Net is also a pre-trained model like VGG19, so we made the same modifications as in the previous architecture to adapt this model to our problem and get an output of seven classes.

To identify the most suitable hyperparameter configurations for our models, we employed the Adam technique with Categorical cross-entropy. We curated a set of hyperparameters aimed at enhancing the models' performance. Table 3 presents the adopted hyperparameters.

Table 3. The hyperparameters adopted for training our models

Types of hyperparameters	Hyperparameters	Proposed values
Layer hyperparameters	Dropout	25%
	Kernel size	3×3 and 5×5
	Final layer activation function	Softmax
	Hidden layer activation function	Rectified linear unit (ReLU)
	Padding	Same
Compiler hyperparameters	Optimization function	Adam
	Error function	Categorical cross-entropy
	Learning rate	0,0001
Execution hyperparameters	Batch size	32
	Number of epochs	100
	Early stopping	Patience = 10

4. RESULTS AND DISCUSSION

4.1. First comparison

In this section, we compare the performance of three model architectures-CNN, VGG19, and MobileNetV2-on two types of datasets: FER2013 (Data N°1) and a grouped set composed of KDEF, KMU-FED, and CK+ (Data N°2). The goal is to determine which model architecture performs best for subsequent use in our model. The results of training the CNN, VGG19, and MobileNetV2 architectures on these datasets are presented in Table 4.

Table 4. The results of models on the two types of datasets

	Data	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)
CNN	Data N°1	61.06	63.75	61.92	61.33
	Data N°2	71.02	71.69	67.25	66.97
MobileNet	Data N°1	62.6	70.85	61.47	62.63
	Data N°2	94.92	94.53	94.52	94.48
VGG19	Data N°1	70.40	72.98	71.09	71.74
	Data N°2	97.31	97.19	96.99	97.09

From Table 4, we conclude that VGG19 is the model that gave us the best results on both datasets. Therefore, to verify that our VGG19 model is not overfitting, we assessed its performance by analyzing both loss and accuracy curves for the two proposed datasets. The results are illustrated in Figures 3 and 4. For the first dataset N1 (Figure 3). The VGG19 architecture gave us a performance of 70.4% but from epoch 8, our model began to overfit, prompting us to save the training state with the best performance. Figure 4 shows that the VGG19 pre-trained model performed well. The training and validation curves increased rapidly from epoch N°2 until epoch 15.

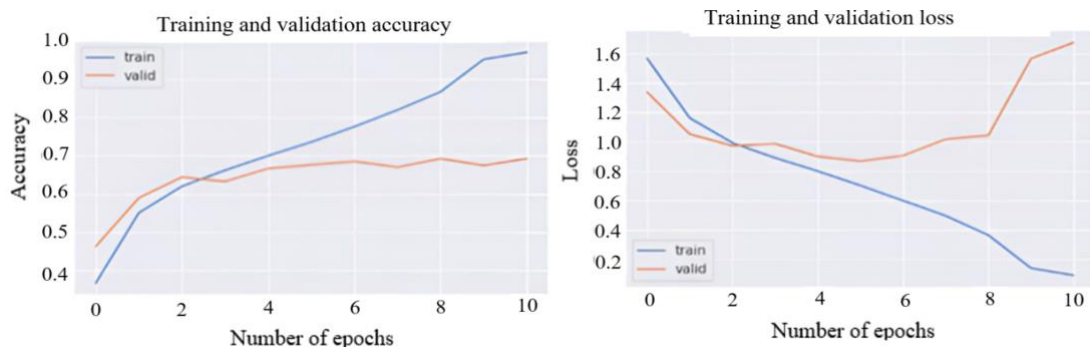


Figure 3. Performance and error evolution of training and validation per epoch for dataset N°1

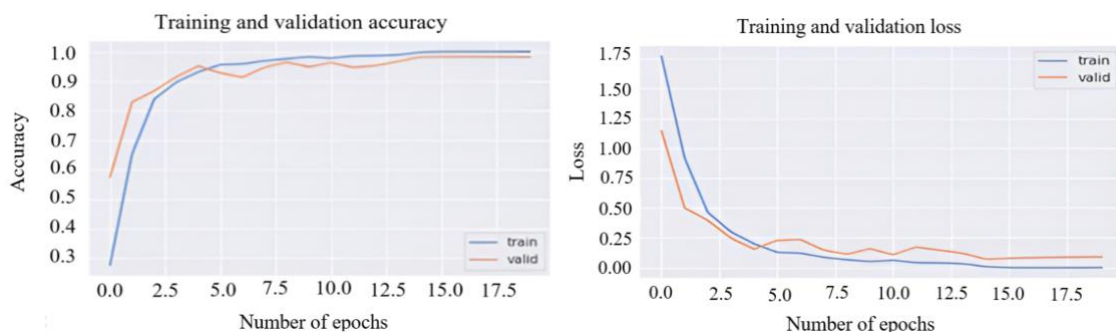


Figure 4. Training and validation performance and error evolution by epoch. For Data N°2

4.2. Our transfer learning approach results

For our proposed approach, we selected the VGG19 model, which gave us the best results on FER2013, and trained it on the second dataset (CK+, KDEF, and KMU-FED). This approach leverages the

benefits of transfer learning, where the knowledge gained from one dataset is used to improve performance on another dataset. The results of our proposed approach were as follows: accuracy of 96.06%, precision of 96.97%, recall of 96.76%, and F1-score of 96.84%. These high-performance metrics demonstrate the effectiveness of our method. The success can be attributed to the initial training on FER2013, which provided a solid learning base. By fine-tuning the model on the more diverse second dataset, we were able to enhance its generalization capabilities and achieve superior results.

To ensure our model is not overfitting, we evaluated its performance using loss and accuracy curves, as shown in Figure 5. Our proposed approach demonstrated significant performance improvement starting from epoch 2. The pre-trained model benefited from the knowledge acquired during its initial training, which is evident from its starting performance of over 80%. The training concluded by epoch 8, with the model achieving over 96% performance, indicating a stable and well-generalized learning process.

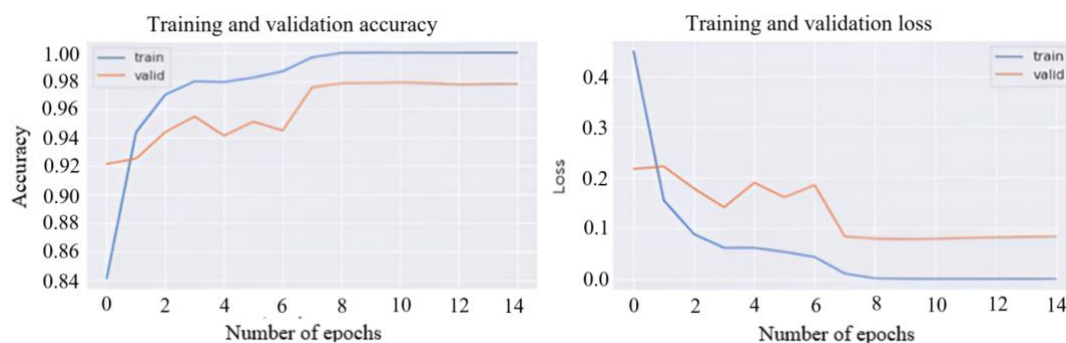


Figure 5. Examples of images from FER2013

The high accuracy of 96.06% achieved by our transfer learning approach can be attributed to several factors. Firstly, VGG19's deep architecture allows it to capture complex patterns and nuances in facial expressions, which are crucial for accurate emotion recognition. The initial training on FER2013 likely provided the model with a strong foundation of generalized features, which were then refined with the more specific and diverse samples from CK+, KDEF, and KMU-FED. VGG19's superior performance compared to CNN and MobileNetV2 can be linked to its deeper layers and more complex structure, making it more adept at recognizing subtle emotional cues. The diverse nature of the second dataset (CK+, KDEF, KMU-FED) contributed to the improved performance by enhancing the model's ability to generalize across different scenarios. The early performance improvements and stable high accuracy indicate that the model effectively retained and utilized the knowledge from its initial training, rather than merely memorizing the new data. These findings have significant implications for the field of emotion recognition, particularly in educational technology, where accurate emotion recognition can enhance the personalization of learning experiences and improve student engagement. By accurately detecting and classifying emotions, our system can predict student engagement levels during distance learning, allowing educators to tailor their teaching strategies to maintain or increase engagement.

4.3. Approach evaluation

To ensure our model's robustness in a variety of settings, we ran an experiment with three distinct models: VGG19 trained only on FER2013, VGG19 trained exclusively on KDEF, CK+, and KMU-FED, and our proposed technique. Eighty photos were chosen at random from the internet for this experiment. We conducted this experiment because we believed that training and validation measures alone are insufficient to fully assess a model's performance in the field of emotion recognition. Variations that are not entirely captured by typical datasets are frequently seen in real-world applications. To accurately evaluate these models' efficacy, it is therefore imperative to test them in a variety of uncontrolled settings. Here is how we collected and annotated these images.

4.3.1. Image collection

To get a wide range of images, we used an organized approach that included the use of search engines with selected keywords. Some of the keywords we used included "happy face," "sad face," "angry face," "surprised face," and a lot of others, to cover the entire scale of emotions we aimed to recognize. This search technique diversified the images and ensured that they were from different sources.

4.3.2. Image annotation

After collecting the images, we manually annotated each one to ensure accuracy. This involved carefully observing each image and labeling it with the appropriate emotion. The manual annotation process was the most crucial part of the dataset maintaining processes, ensuring that every image was rightly categorized according to the emotion it showed. Several annotators performed this process to eliminate the bias and enhance the annotation's reliability.

4.3.3. Experiment results

The results of our experiment showed that the VGG19 model trained on FER2013 made 60 correct predictions out of 80, achieving a performance of 75%. In comparison, the VGG19 model trained on KDEF, CK+, and KMU-FED correctly predicted 53 out of 80 images, giving a performance of 66.25%. Our proposed approach demonstrated superior performance by correctly predicting 75 emotions out of 80 images, which corresponds to a performance of 93.75%. The results indicate that leveraging a combination of datasets through transfer learning can greatly enhance the generalization capability of emotion recognition models, making them more reliable for practical applications. The results of our experiment are presented in Table 5. Through this experiment, we demonstrated that our proposed approach is more effective in real-world scenarios compared to models trained solely on individual datasets. This practical evaluation is crucial for applications in emotion recognition, where real-world variability must be accounted for to ensure robust model performance.

Table 5. The results of our experiment

Models	Accuracy of the model (%)	Number of correct predictions
VGG19 on FER2013	70.40	60 on 80
VGG19 on KDEF, CK+, KMU-FED	97.31	53 on 80
Our approach	96.06	75 on 80

4.4. Student engagement weighting

To enhance the utility of our emotion recognition system for educational settings, we have developed a method to translate the seven detected emotions into three levels of student engagement: Highly engaged, engaged, and disengaged. This translation is based on the concentration index (CI) calculated using the dominant emotion probability (DEP) [25]. The CI is further refined through the application of emotion weights, as detailed in Table 6 and calculated using (1), which collectively enable a nuanced understanding of student engagement states.

$$CI = DEP \times EW \quad (1)$$

Where EW represents the emotion weight.

Table 6. Weight for corresponding emotion

Detected Emotion	EW
Neutral	0.9
Happy	0.6
Surprised	0.6
Sad	0.3
Disgust	0.2
Anger	0.25
Fear	0.3

Emotions such as happiness, surprise, and neutrality, which yield higher concentration indices, are categorized as indicators of engagement, reflecting a student's active involvement and positive response to the learning material. Conversely, emotions such as sadness, anger, fear, and disgust, associated with lower concentration indices, are categorized as indicators of disengagement, reflecting a lack of interest or negative response to the learning environment. By simplifying the emotional data into binary engagement levels using this model, educators can more easily interpret and respond to the emotional states of their students, facilitating timely and effective interventions to maintain or increase student engagement. To further refine the engagement categorization, the CI is used to classify engagement into three levels based on the ranges provided in Table 7. By applying these CI ranges, educators can distinguish between highly engaged, engaged, and disengaged students, providing a more nuanced understanding of student engagement. This classification helps in identifying students who may need additional support or intervention to improve their learning experience.

Table 7. Engagement detection from CI

Engagement type	CI (%)
Highly-engaged	> 65
Engaged	25-65
Disengaged	< 25

5. CONCLUSION

Distance learning is becoming increasingly essential. The adoption of this modern method of learning and teaching comes with its advantages and, above all, its challenges. The subject of student engagement during virtual courses is becoming increasingly interesting, especially its relationship with the emotions that accompany the distance learning experience. In this work, we studied the different types of datasets dedicated to the task of emotion recognition and introduced our approach based on transfer learning to be able to create a system for detecting emotions expressed from the student's face during his learning session to subsequently determine his rate of engagement. Our proposed approach which consists of training VGG19 on FER2013 and then fine-tuning the resulting weights on the combined CK+, KDEF, and KMU-FED datasets gave us the best results. We were able to take advantage of the diversity and quantity of images presented by FER2013 and benefit from the quality of images presented by the CK+, KDEF, and KMU-FED datasets, by adopting this approach, we were able to achieve a performance of 96,06%. While our study has provided valuable insights into the development and performance of our facial emotion recognition system, it has limitations in the datasets used. The sample size remains relatively modest, and certain demographic groups may be underrepresented. In our further research, we first intend to overcome this limitation. In addition, the study carried out in this article is essentially based on the analysis of students' emotions in an e-learning environment; however, emotions are not the only indicator for assessing learner engagement, but their association with concentration rate. Therefore, to work on other aspects to deduce student concentration and engagement by linking the study of emotions with posture, eye movement, and other indicators that can help supervise student engagement.




REFERENCES

- [1] J. Reichert-Schlaß, O. Zlatkin-Troitschanskaia, K. Frank, S. Brückner, M. Schneider, and A. Müller, "Development and evaluation of digital learning tools promoting applicable knowledge in economics and german teacher education," *Education Sciences*, vol. 13, no. 5, 2023, doi: 10.3390/educsci13050481.
- [2] L. Mishra, T. Gupta, and A. Shree, "Online teaching-learning in higher education during lockdown period of COVID-19 pandemic," *International Journal of Educational Research Open*, vol. 1, 2020, doi: 10.1016/j.ijedro.2020.100012.
- [3] S. Naseer and H. Z. Perveen, "Perspective chapter: Advantages and disadvantages of online learning courses," *Massive Open Online Courses - Current Practice and Future Trends*, 2023, doi: 10.5772/intechopen.1001343.
- [4] K. Kubikova, A. Bohacova, J. Slowik, and I. Pavelkova, "Student adaptation to distance learning: An analysis of the effectiveness, benefits and risks of distance education from the perspective of university students," *Social Sciences and Humanities Open*, vol. 9, 2024, doi: 10.1016/j.ssaho.2024.100875.
- [5] I. Rotnitsky, R. Yavich, and N. Davidovich, "The impact of the pandemic on teachers' attitudes toward online teaching," *International Journal of Higher Education*, vol. 11, no. 5, pp. 18–38, 2022, doi: 10.5430/ijhe.v11n5p18.
- [6] E. M. Polo, A. Farabbi, M. Mollura, L. Mainardi, and R. Barbieri, "Understanding the role of emotion in decision making process: using machine learning to analyze physiological responses to visual, auditory, and combined stimulation," *Frontiers in Human Neuroscience*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1286621.
- [7] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993, doi: 10.1037/0003-066X.48.4.384.
- [8] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–717, 1987, doi: 10.1037/0022-3514.53.4.712.
- [9] P. E. Griffiths, "III. Basic emotions, complex emotions, machiavellian emotions," *Royal Institute of Philosophy Supplement*, vol. 52, pp. 39–67, 2003, doi: 10.1017/s1358246100007888.
- [10] K. Altuwairqi, S. K. Jarraya, A. Allinjawi, and M. Hammami, "A new emotion-based affective model to detect student's engagement," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 1, pp. 99–109, 2021, doi: 10.1016/j.jksuci.2018.12.008.
- [11] X. Zheng, S. Hasegawa, M. T. Tran, K. Ota, and T. Unoki, "Estimation of learners' engagement using face and body features by transfer learning," *Artificial Intelligence in HCI*, pp. 541–552, 2021, doi: 10.1007/978-3-030-77772-2_36.
- [12] M. N. Hasnine, H. T. T. Bui, T. T. T. Tran, H. T. Nguyen, G. Akçapınar, and H. Ueda, "Students' emotion extraction and visualization for engagement detection in online learning," *Procedia Computer Science*, vol. 192, pp. 3423–3431, 2021, doi: 10.1016/j.procs.2021.09.115.
- [13] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11365–11394, 2023, doi: 10.1007/s11042-022-13558-9.
- [14] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021, doi: 10.1016/j.jksuci.2018.09.002.
- [15] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, 2021, doi: 10.1016/j.ijcce.2021.02.002.
- [16] B. Meriem, H. Benlahmar, M. A. Naji, E. Sanaa, and K. Wijdane, "Determine the level of concentration of students in real time from their facial expressions," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 159–166, 2022, doi: 10.14569/IJACSA.2022.0130119.




- [17] G. P. Kusuma, A. Jonathan, and P. Lim, "Emotion recognition on FER-2013 face images using fine-tuned VGG-16," *Advances in Science, Technology and Engineering Systems*, vol. 5, no. 6, pp. 315–322, 2020, doi: 10.25046/aj050638.
- [18] H. Alshamsi, V. Kepuska, and H. Meng, "Real time automated facial expression recognition app development on smart phones," in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2017*, 2017, pp. 384–392, doi: 10.1109/IEMCON.2017.8117150.
- [19] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on auxiliary models," *Algorithms*, vol. 12, no. 11, 2019, doi: 10.3390/a12110227.
- [20] T. Debnath, M. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-11173-0.
- [21] J. H. Kim, R. Mutegeki, A. Poullose, and D. S. Han, "A study of a data standardization and cleaning technique for a facial emotion recognition system," in *Proceedings of the Korea Telecommunications Society*, 2020, pp. 1193–1195.
- [22] S. M. González-Lozoya, J. D. L. Calleja, L. Pellegrin, H. J. Escalante, M. A. Medina, and A. Benitez-Ruiz, "Recognition of facial expressions based on CNN features," *Multimedia Tools and Applications*, vol. 79, no. 19–20, pp. 13987–14007, 2020, doi: 10.1007/s11042-020-08681-4.
- [23] A. B. Shetty, Bhoomika, Deeksha, J. Rebeiro, and Ramyashree, "Facial recognition using Haar cascade and LBP classifiers," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 330–335, 2021, doi: 10.1016/j.gltp.2021.08.044.
- [24] P. Sharma, M. Esengönül, S. R. Khanal, T. T. Khanal, V. Filipe, and M. J. C. S. Reis, "Student concentration evaluation index in an E-learning context using facial emotion analysis," *Communications in Computer and Information Science*, vol. 993, pp. 529–538, 2019, doi: 10.1007/978-3-030-20954-4_40.
- [25] P. Sharma *et al.*, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," *arXiv-Computer Science*, pp. 1–18, 2019.

BIOGRAPHIES OF AUTHORS






Ikram Qarbal    completed her master's degree in 2023 from the Faculty of Sciences, Casablanca, Morocco, and her bachelor's degree in 2021 from the Faculty of Sciences and Techniques, Mohammedia, Morocco. She is currently pursuing her doctoral degree in the domain of Data Science at the Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Morocco. Her research interest is in the field of machine learning, deep learning, AI, and computer vision in the field of e-learning. She can be contacted at email: ikram.ql.11@gmail.com.



Nawal Sael    teacher-researcher since 2012, Authorized Professor since 2014, and Professor of Higher Education in the Department of Mathematics and Computer Science at the Ben M'Sick Faculty of Sciences in Casablanca, Morocco since 2020, and her engineer degree in software engineering from ENSIAS in 2002. Her research interests include data mining, machine learning, deep learning, and the internet of things. She can be contacted at email: saelnawal@hotmail.com.



Pr. Sara Ouahabi    is a Habilitated Professor (PH) at the mathematics and Department of Computer Science of Hassan 2 University and a member of the Computer Science and Information Processing Laboratory at the Faculty of Science Ben M'sik. Her research includes artificial intelligence, computer security, the semantic web, e-learning, computer communications (networks), and the internet of things (IoT). She conducts cutting-edge research in these areas while actively engaging in higher education to educate the next generation of computer professionals. She can be contacted at email: sara.ouahabi@gmail.com.