Indonesian Journal of Electrical Engineering and Computer Science
Vol. 36, No. 2, November 2024, pp. 1199∼1208
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v36.i2.pp1199-1208

An improved student's facial emotions recognition method using transfer learning

Amimi Rajae, Radgui Amina, Ibn El Haj El hassane
CEDOC2IT, National Institute of Posts and Telecommunications, Rabat, Morocco

Article Info

Article history:

Received Mar 19, 2024
Revised Aug 1, 2024
Accepted Aug 5, 2024

Keywords:

Digital classroom
FER system
Students affect states
Transfer learning

ABSTRACT

Instructors endeavour to encourage active participation and interaction among learners. However, in settings with a large number of students, such as universities or online platforms, obtaining real-time feedback and evaluating teaching methodology presents a significant challenge. In this paper, we introduce a student engagement recognition system based on a hybrid method using handcrafted features and transfer learning. The research is conducted on two databases for emotion detection based on facial cues (FER13) benchmarked dataset and our database. We use the local binary patterns (LBP) method combined with pre-trained MobileNet model for feature extraction and classification. The proposed system adeptly discerns students' facial expressions and categorizes their engagement states as either 'engaged' or 'disengaged'. We determine the most effective model by evaluating and comparing several deep learning models, including Inception-V3, VGG16, EfficientNet, ResNet, and DenseNet. Experimental results underscore the efficacy of our approach, revealing a remarkable accuracy, surpassing benchmarks set by state-of-the-art models.

Corresponding Author:

Amimi Rajae
CEDOC2IT, National Institute of Posts and Telecommunications
Rabat, Morocco
Email: amimi.rajae004@gmail.com

1. INTRODUCTION

Recognizing and addressing students' emotions in the learning environment is crucial for enhancing their overall educational experience. Previous research, such as that by Leony *et al.* [1], has confirmed the potential benefits of intelligent systems in providing educators with insights into learners' emotional states. While the subject of evaluating student engagement through facial expression analysis has a rich history, the development of efficient and straightforward methods has been limited. Current student facial expression detection methods face significant challenges: machine learning (ML) algorithms provide limited results, while deep learning (DL) approaches demand extensive datasets [2]. Therefore, there is a need for methodologies that can effectively handle smaller datasets while maintaining high accuracy.

Whitehill *et al.* [3] analyzed engagement levels in students using facial information by employing support vector machines (SVM) along with Gabor filters for feature extraction and classification. The dataset utilized in their research represents a single person captured in a single frame via a webcam. The accuracy of their method is limited, and their study is suited only for online learning settings, thus lacking applicability in various learning environments. Similarly, Tang *et al.* [4] introduced an innovative and efficient prototype system. They employed uniform local gabor binary pattern histogram sequence (ULGBPHS) for feature extraction and K-nearest neighbors (KNN) for classification.

They found similar results to those obtained by Whitehill *et al.* [3]. Consistent with these findings, Monkaresi *et al.* [5], they have collected a spontaneous database of students' facial expressions while interacting with a game-based physics education tool. using the LBP-TOP algorithm, they classified the degree of students' attention based on their facial expressions but their study yielded modest accuracy.

These early studies established the groundwork by employing ML algorithms for feature extraction and classification. Despite their efforts, results remained modest, primarily due to limitations in dataset size and the use of traditional ML approaches. The breakthrough came with Kim *et al.* [6] comprehensive study, which introduced DL into the field of FER for education, signalling a paradigm shift. DL methods consistently exhibited higher accuracy rates, but they necessitated extensive datasets, a challenge often sidestepped by researchers who resorted to generalized facial emotion databases.

To bridge this gap, a few recent studies have focused on creating fitted datasets for student engagement recognition. For instance, in 2019, Ashwin and Guddeti [7] proposed a methodology employing a convolutional neural network (CNN) based on the GoogleNet architecture. The model, trained on both posed and spontaneous databases, achieved competitive results, though it was not verified on small and medium-sized datasets, but it does reveal the efficiency of DL methods. Another study by Summer *et al.* [8] adopted affect-Net as the foundational architecture. Their model was trained on a large spontaneous dataset featuring three engagement levels (low, medium, and high). These findings underscore the efficacy of deep learning approaches, specifically using CNN architectures like GoogleNet and ResNet-50, across diverse scenarios and datasets, enhancing the accuracy of FER systems.

In this paper, we propose a hybrid approach that combines traditional handcrafted features with pretrained DL algorithms, fitted to the task of student engagement recognition. By exploring the strengths of both methodologies, we aim to overcome the challenges posed by dataset scarcity while enhancing the precision and accuracy of FER systems in educational settings.

Our key contributions include the development of an optimized FER model based on transfer learning and the creation of a specialized database carefully annotated for reliable model training and assessment. These affirmative advancements contribute to the comprehensive goal of enhancing real-time student engagement recognition systems within educational settings. We organize this paper as follows: section 2 outlines the research methodology, and section 3 presents the experimental results and discussion, in addition to a comprehensive conclusion of the article and suggestions of potential work for future research.

2. PROPOSED METHOD

Our approach combines MobileNet, a lightweight convolutional neural network, with LBP feature extraction. MobileNet efficiently processes image data, while LBP captures local texture information. This fusion allows for robust and efficient classification, particularly in tasks like facial emotion recognition, where subtle texture variations are crucial.

To train and validate the proposed engagement detection model using LBP-MobileNet, we collected a customised dataset in a real classroom setting. This dataset is specifically designed to capture the nuanced facial expressions and emotions exhibited by students during educational activities, ensuring its relevance and applicability to our research context. Subsequently, we rigorously tested and evaluated our approach on the FER13 dataset, a widely recognized benchmark dataset used by numerous researchers in the field of emotion detection. By conducting experiments on this established dataset, we aim to assess the generalizability and performance of our model across diverse scenarios and datasets.

2.1. FER13 database

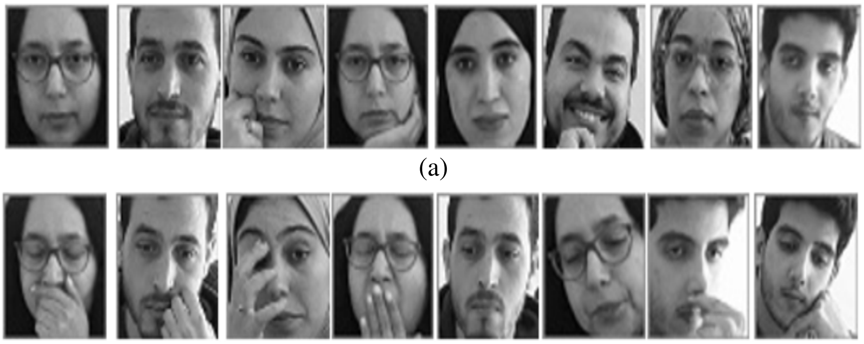
The FER2013 [9] dataset comprises 28,709 images annotated with seven distinct emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. These images were sourced from the internet, providing a diverse collection. Notably, this dataset serves as a prevalent benchmark in the field of facial expression recognition, frequently referenced in research papers for training and evaluating models.

2.2. Own database

Student facial expression databases remain limited, and most authors train their models using general FER databases. Therefore, having an adequate database is crucial to enhance the models' accuracy and provide improved outcomes [10].

The facial expression database we've developed focuses on capturing spontaneous expressions related to engagement, classifying them into two categories: "engaged" and "disengaged." The dataset is substantial, comprising 2 hours of video recordings, 718 facial images, and involving 11 participants, ensuring a balanced demographic with 5 males and 6 females. The facial expressions were captured using a high-resolution 50 MP camera from a frontal view, providing detailed and consistent imagery. To enrich the dataset, a combination of manual and expert labelling methods was employed. This comprehensive database serves as a valuable resource for the development and evaluation of engagement recognition models.

In Figure 1, we present several samples from our dataset. Notably, a unique feature of this dataset is its inclusion of faces with various occlusions, such as beards, scarves, darker skin tones, and instances where individuals have their hands on their faces. These distinctive attributes contribute to improving our suggested model's performance. We will make our database available online to be used in further researches hence we will refer to it as "StuEmo24: Students' Emotions - Student Facial Expression Dataset - 2024". This dataset may significantly advance research in areas like emotion recognition systems and human-computer interaction by providing a useful resource for training and assessing machine learning models in affect recognition tasks.



(a)

(b)

Figure 1. Samples of images from our dataset including two classes (a) engaged and (b) disengaged

2.3. Face detection

Face tracking and detection in computer vision are typically challenging problems that need advanced methods for accurate face recognition. Numerous approaches, including the HAAR cascade classifier [11], histogram of oriented gradient (HOG) [12], LBP [13], and its variants, have demonstrated remarkable face detection performance. However, when faces are tilted, occluded, or multiple faces are present in an image, conventional methods like HAAR Cascade and previously cited approaches, exhibit limitations. In the context of our study, which focuses on analyzing students' engagement in a classroom setting based on facial expressions, a robust method is imperative for recognizing faces under challenging conditions. To address this, we have adopted the single-shot multi-level face localisation in the wild (RetinaFace) method proposed by Deng *et al.* [14]. This method has shown effectiveness in detecting faces under various conditions, offering high accuracy even in complex scenarios.

2.4. Data processing

After the face detection and annotation phase, StuEmo24 comprises 457 images categorized as "engaged" and 261 images as "disengaged." To facilitate training and testing, we divided the images into a training set (90%) and a test set (10%) for each category. Given the modest size of our dataset, which might not be optimal for training a neural network from scratch, we opted for image augmentation using the Keras library. This strategy effectively enlarges the dataset before integrating it into the network to avoid model overfitting and non-convergence.

The augmentation process we used included rotation within a 15-degree range, zooming with a 0.15 range, shearing with a scale of 0.15, horizontal flipping, and width and height shifting with a value of 0.15. These operations collectively contribute to a more diverse and robust training dataset. However, recognizing the limitations of data augmentation in achieving a sufficiently extensive dataset, we recommend considering the transfer learning technique outlined in the following section. Transfer learning, as demonstrated in various classification tasks [15], has proven effective with medium size datasets.

2.5. Feature extraction

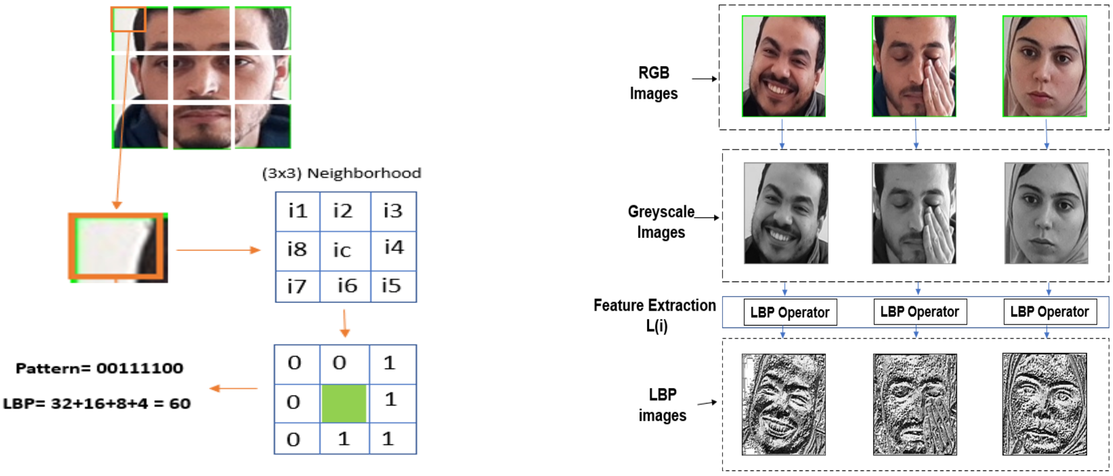
Utilizing facial images for gauging emotions involves the cautious detection of minor changes in facial landmarks. For instance, transitions in the edges, corners, and contours of the eyes, shifting from closed to wide open, can indicate a shift from a dull expression to one of focused concentration. These alterations in key facial regions of interest (ROIs), covering the left and right eyebrows, eyes, nose, and mouth, are filtered through the lens of the LBP.

LBP serve as a method for extracting features in image processing, particularly adept at capturing local texture patterns by assessing pixel intensities within a specified neighbourhood. LBP's resilience to illumination changes enhances model robustness, making it a favourable choice. In this process, we compare each pixel's intensity with its neighbours and assign binary values based on the comparison, resulting in an LBP code for the central pixel, as in (1). This code represents the texture pattern within the neighbourhood. By repeating this process for every pixel in the image, we generate a comprehensive LBP representation, facilitating the examination of the image's texture properties. Then we feed this representation into our neural network.

In our experimental setup, images are captured under varying lighting conditions, posing challenges to model performance. However, the utilization of LBP resolve this issue by making images insensitive to illumination changes, thereby enhancing the robustness of the model. Moreover, LBP's reliance on simple comparison and thresholding operations ensures computational efficiency, rendering it suitable for real-time application. Furthermore, LBP's ability to effectively reduce the dimensionality of the feature space simplifies classifier operations and mitigates the risk of overfitting, thus contributing to the overall effectiveness of the model. Figure 2 shows the application of LBP on StuEmo24.

$$LBP(p) = \sum_{i=0}^{P-1} s(i) \times 2^{i} \tag{1}$$

Where:
- LBP(p) is the local binary pattern for the centre pixel p.
- P is the number of sampling points in the neighbourhood N.
- s(i) represents the result of the intensity comparison between the centre pixel p and its i-th neighbour.



(a) (b)

Figure 2. LBP applied on StuEmo24 (a) procedure of LBP method and (b) samples from StuEmo24 with the application of LBP on face images

2.6. Students' engagement recognition model

Deep neural networks (DNNs) have played a transformative role in the field of image classification; however, they still face several challenges that researchers and practitioners continuously work to address. Some of the key challenges include data quality and quantity, computational resources, and training time. Therefore, transfer learning is a powerful tool that overcomes these challenges, enabling the effective reuse of knowledge from pre-trained models and adapting to new tasks.

In our study, we tested different CNNs architectures based on pre-trained models on ImageNet [16] database that comprises more than 14 million images manually annotated. We have evaluated the following state-of-the-art algorithms: EfficientNet, DenseNet, InceptionV3, ResNet, VGG16, and MobileNet.

Our proposed model initializes a neural network using the MobileNet [17] architecture as shown in Figure 3. MobileNet is a lightweight deep neural network based on depthwise convolution that applies a separate convolutional filter to each input channel, leading to a reduction in computation while still capturing spatial information across channels. Its mathematical formulation is as follows:

Given an input feature map $\mathbf{X}$ with depth D_i, a depthwise convolution filter $\mathbf{W}$ with depth D_f, and spatial dimensions $K \times K$, the depthwise convolution operation at spatial location (i, j) can be represented as (2):

$$DConv(i, j, d_f) = \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} X(i + k, j + l, d_i) \times W(k, l, d_f) \tag{2}$$

where:
- X(i+k, j+l, d$_i$) represents the input feature map value at spatial location $(i + k, j + l)$ and depth d_i.
- W(k, l, d$_f$) represents the filter value at spatial location (k, l) and depth d_f.
- DConv(i, j, d$_f$) is the output feature map value at spatial location (i, j) and depth d_f.

The model is configured with a 48x48 input shape, excluding fully connected layers, and leverages pre-trained weights. To capture intermediate features, we extracted the output of the 14th layer from the end of the base model. Subsequently, we applied global max pooling to this chosen layer, reducing spatial dimensions and retaining essential features. We employed a batch size of 25 images and applied SoftMax as the activation function to the output of the fully connected layer (3), along with the Adam optimizer. Given an input vector $\mathbf{z}$, SoftMax computes the output vector $\mathbf{y}$ as (3):

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{3}$$

where:
- e is the base of the natural logarithm.
- z_i is the i-th element of the input vector $\mathbf{z}$.
- $\sum_{j=1}^{n} e^{z_j}$ is the sum of the exponential values of all elements in the input vector.

Our choice of loss function was categorical cross-entropy. The categorical cross-entropy loss function is commonly used in multi-class classification tasks. It measures the dissimilarity between the true distribution of the data and the predicted distribution outputted by the model. Here we used it for binary classification by treating it as a special case where there are only two classes. In this case, we represent the true labels as one-hot encoded vectors with two elements. Mathematically, the categorical cross-entropy loss function for binary classification can be as (4):

$$\text{CCE } = -\sum_{i=1}^{N} y_i \cdot \log(p_i) \tag{4}$$

where:

- N is the number of samples
- y_i is the true label (either 0 or 1)
- p_i is the predicted probability that the sample belongs to class 1

The model was trained over 100 epochs. However, we noted that the accuracy plateaued starting from the 25th epoch.

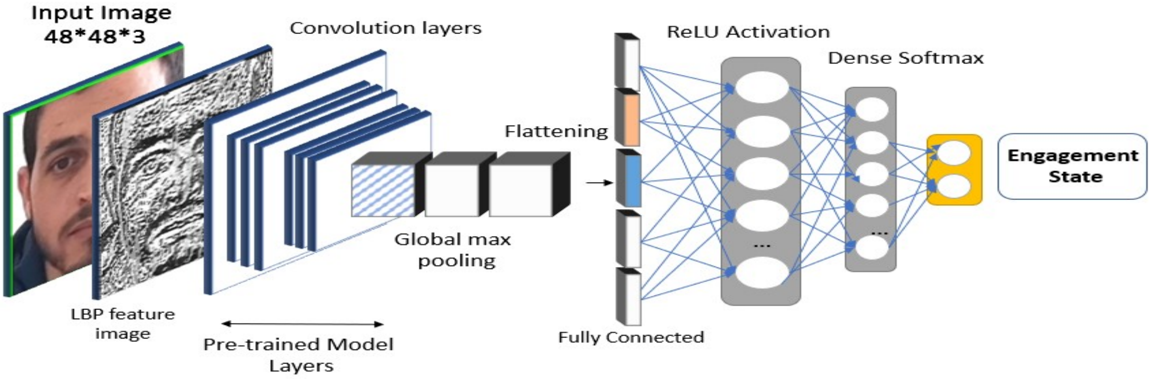


Figure 3. Proposed model architecture

3. RESULTS AND DISCUSSION

3.1. Experimental set-up

We employ a hybrid ML-DL approach utilizing the TensorFlow and Keras libraries for model training. To enhance efficiency, we adopt a transfer learning model based on a pre-trained neural network, ensuring faster and more effective results. The model training is conducted on gradient [18], a platform designed for constructing and scaling real-world machine learning applications. The training environment includes a P5000 machine equipped with TensorFlow version 2.9.11, 30 GB of RAM, an 8 GB CPU, and a 16 GB GPU. To fine-tune the learning process, improve efficiency, and prevent overfitting, we used two callbacks from the Keras library: the EarlyStopping callback monitors validation accuracy, and stops the training process once the model achieves the highest accuracy, saving time and preventing overfitting. The ReduceLROnPlateau callback focuses on adjusting the learning rate during training, monitoring validation accuracy, and reducing the learning rate if accuracy stops improving. Table 1 describes the model features details.

Table 1. Model features details

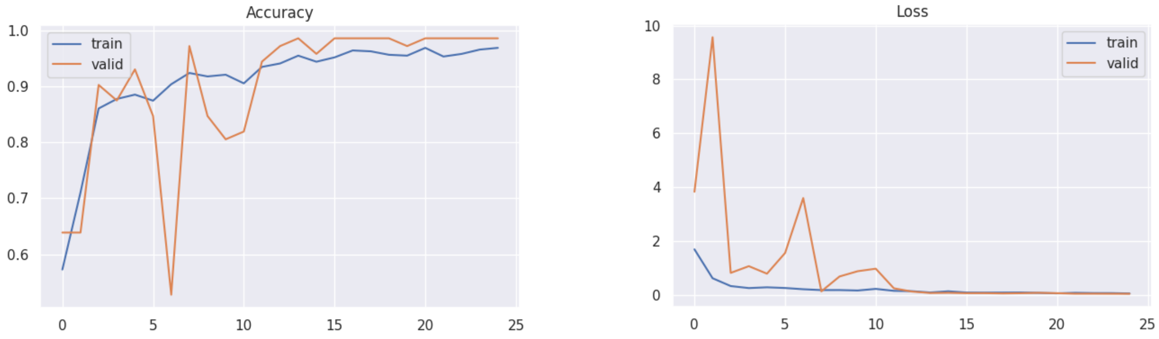
Feature	Batch size	Epochs	Activation function	Optimizer	LR	Loss function
Details	25	20	SoftMax	Adam	0.001	Categorical cross entropy

3.2. Analysis of the results

Our study investigated the performance of a DL model combined with an ML algorithm trained on a medium-sized dataset. Previous studies have typically applied either ML approaches to medium-sized datasets, yielding modest results, or DL approaches to very large but not contextually adapted datasets, resulting in outcomes that are not applicable to real classroom settings. None of the existing literature explicitly addresses the application of performant DL algorithms on small or medium-sized datasets. By exploring the strengths of both methodologies, we aim to overcome the challenges posed by dataset scarcity while enhancing the precision and accuracy of SFER systems.

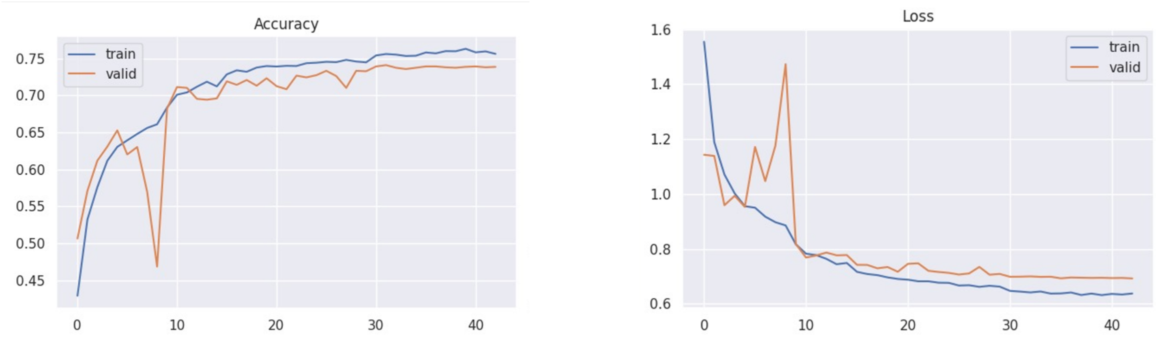
Figure 4, demonstrates that by combining the LBP method along with the pre-trained mobileNet, the model exhibits fluctuations in performance during the first 10 epochs of training. However, over subsequent epochs, the accuracy steadily increases until reaching a remarkable accuracy of 98%. This achievement represents a significant improvement over previous benchmarks, as evidenced by the best achievable accuracy of only 92.3% using ResNet-50 as reported by [19]. This enhanced accuracy holds practical implications, particularly in real-time applications where the ability to accurately detect students' emotional states through facial expressions is of prime importance for providing real-time feedback.

Furthermore, we evaluated the model on the FER13 dataset, where it demonstrated strong performance, achieving an accuracy of 73.6%, as illustrated in Figure 5. This underscores the robustness and generalization capabilities of our approach across different datasets and tasks.



(a) (b)

Figure 4. Optimized model accuracy and loss rate curve on our dataset (a) model accuracy curve and (b) model loss curve



(a) (b)

Figure 5. Optimized model accuracy and loss rate curve on FER13 dataset (a) model accuracy curve and (b) model loss curve

Our experimentation involved testing various models, including EfficientNet and InceptionNet, all of which provided only moderate results, failing to surpass 64% accuracy even with the integration of LBP feature extraction. The lower accuracy of EfficientNet (57%) and InceptionNet (64%) might be due to their trade-off between efficiency and complexity. While they are designed to be computationally efficient, they may struggle to capture nuanced features necessary for accurate facial emotion recognition.

In contrast, employing the VGG16 and DenseNet architectures resulted in a substantial increase in accuracy. Building upon this success, we continued to iterate by incorporating LBP treatment and exploring different neural network configurations. Figure 6 shows the results. Ultimately, our efforts come to an end with the adoption of MobileNet coupled with LBP, achieving an exceptional accuracy of 98.2% on StuEmo24 dataset. This achievement represents a significant advancement, surpassing state-of-the-art results in the field.

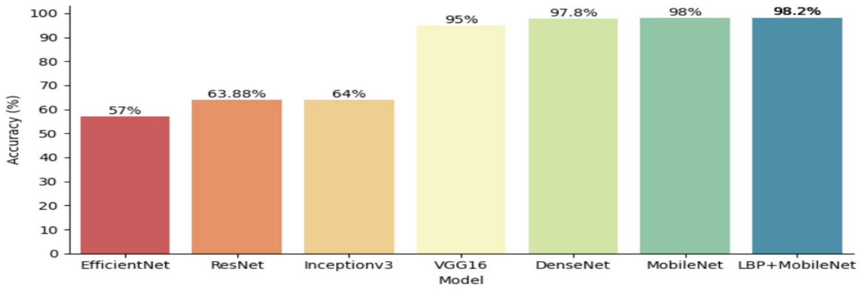


Figure 6. Accuracy of various models applied on our dataset

3.3. Comparison with state-of-the-art results

The literature demonstrates that handcrafted features exhibit limited efficacy in detecting student engagement through facial expressions in real-world settings [3], [5] particularly when compared to deep learning methods [19]-[21]. Our hybrid approach outperforms existing methods, achieving superior results. However, direct comparison with prior methodologies is challenging due to variations in datasets and modalities. Despite these challenges, our results demonstrate comparable and enhanced accuracy and robustness.

Previous research by Ashwin and Guddeti [7] employed a hybrid approach utilizing solely deep learning models as backbones, providing promising results ranging from 70% to 81% accuracy. Additionally, [19] utilized ResNet50 as the base model, achieving an impressive 92.3% accuracy. These findings encouraged us to explore diverse deep-learning models as backbone architectures. Although ML models have shown promising but insufficient results, as demonstrated by Whitehill *et al.* [3], who achieved an AUC of 0.729 using Gabor features in conjunction with SVM. While Gabor features are well-suited for tasks necessitating texture and shape analysis, such as object detection, fingerprint recognition, and biomedical image analysis, we opted for LBP due to their effectiveness in tasks emphasizing texture patterns, including texture classification, face recognition, and emotion recognition. By training MobileNet on the pertinent features extracted from LBP, we effectively improve the nuances of underlying signal fluctuations crucial for engagement recognition. This methodology provides notable improvements in classification accuracy, distinguishing our approach as a promising advancement in the field. Table 1 illustrates the results of our method in comparison with recent methodologies.

Table 2. Comparison of accuracy of state-of-the-art FER models

Author	Method	Accuracy
Ashwin and Guddeti [7]	Two CNN architectures: CNN1 + CNN2	CNN1: 86%, CNN2: 70%
Ma *et al.* [21]	CNN	76.9%
Gupta *et al.* [19]	ResNet-50	92.3%
Pabba and Kumar [20]	Improved version of mini-Xception	72.68%
Shen *et al.* [22]	Lightweight attentional convolutional network (SE-DAN)	56%
Hu *et al.* [23]	Optimized ShuffleNet v2	63.9%
Our method	**LBP + MobileNet**	**98.2%**

However, it is important to acknowledge certain limitations in our approach. While our method demonstrates proficiency in handling occlusions and variations in facial expressions, we have yet to test its performance on a dataset comprising different lighting conditions. Lighting variations can significantly impact the accuracy of facial expression recognition models. Although the integration of LBP mitigates the model's dependency on lighting changes to some extent, further evaluation under diverse lighting conditions is necessary to assess its robustness comprehensively.

3.4. Application and generalisability of the students' engagement recognition system

Our study presents a framework for identifying student engagement. This system provides insightful feedback, which is especially helpful for educators who are just starting out in the field and want to improve their instructional strategies. Furthermore, our method offers instructors and students personalized content feedback, which can further personalize the teaching and learning process. It also allows research into possible relationships between students' emotional states and their academic achievement.

Our system can assist instructors in situations like webinars and large class sizes by taking care of the difficulty of tracking each student's participation in real-time. Our approach is more robust than existing autotutors, and its integration into intelligent tutoring systems could improve personalized learning experiences by replacing traditional methods of predicting emotional states. Due to the spread of COVID-19, there is a growing need to adapt our system to recognize student engagement even when faces are masked. In future work, we aim to customize the system to effectively consider occluded faces, particularly those covered by masks [24], [25].

Beyond educational settings, our emotional state prediction model may have applications in healthcare, where it can assist in patient care by assessing emotional responses to treatments, therapies, and healthcare environments. This could potentially enhance patient satisfaction and overall healthcare outcomes. Additionally, it may offer insights into workplace productivity and well-being, allowing for the assessment of employee engagement, stress levels, and overall well-being in workplace environments. Utilizing this information, organizations can optimize work environments and policies to enhance productivity and employee satisfaction. In each domain, the adaptation of this system necessitates the re-engineering of contextual features customized to the specific requirements and conditions of that domain.

4. CONCLUSION

The primary objective of this study was to develop a framework for identifying student engagement through facial expressions, utilizing transfer learning. The personalized insights gained from monitoring student engagement have the potential to enhance instructional strategies and offer more engaging educational experiences. In addressing the challenges posed by low accuracy in machine learning methods and the need for larger datasets in deep learning approaches, our research demonstrated that the fusion of handcrafted features, particularly LBP, with the pre-trained MobileNet model leads to significant success on medium-sized FER datasets. Through rigorous training with data augmentation and evaluation against the FER13 benchmarked dataset, this hybrid method exhibited robustness in predicting learners' emotional states across diverse learning environments. Moving forward, it is recommended that future research works on expanding the dataset to include images of participants from diverse cultural backgrounds and wider age ranges. Additionally, integrating cognitive engagement assessments, such as cognitive quizzes, could clarify the relationship between students' engagement states and their academic outcomes. By addressing these recommendations and continuing to refine our methodologies, we can further advance the understanding and application of facial expression recognition systems in educational contexts and beyond where analyzing user emotions is critical, such in advertisement domain, online shopping and video consumption.

REFERENCES

[1] D. Leony, P. J. Muñoz-Merino, A. Pardo, and C. D. Kloos, "Provision of awareness of learners' emotions through visualizations in a computer interaction-based environment," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5093–5100, 2013, doi: 10.1016/j.eswa.2013.03.030.

[2] A. Devarapalli and J. M. Gonda, "Investigation into facial expression recognition methods: a review," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 31, no. 3, pp. 1754–1762, 2023, doi: 10.11591/ijeecs.v31.i3.pp1754-1762.

[3] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Jan. 2014, doi: 10.1109/TAFFC.2014.2316163.

[4] C. Tang, P. Xu, Z. Luo, G. Zhao, and T. Zou, "Automatic facial expression analysis of students in teaching environments," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9428, Springer International Publishing, 2015, pp. 439–447.

[5] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, Jan. 2017, doi: 10.1109/TAFFC.2016.2515084.

[6] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally aware AI smart classroom: current issues and directions for engineering and education," IEEE Access, vol. 6, pp. 5308–5331, 2018, doi: 10.1109/ACCESS.2018.2791861.

[7] T. S. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, Oct. 2020, doi: 10.1007/s10639-019-10004-6.

[8] O. Sumer, P. Goldberg, S. Dmello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1012–1027, 2023, doi: 10.1109/TAFFC.2021.3127692.

[9] I. J. Goodfellow et al., "Challenges in representation learning: a report on three machine learning contests," in *Neural Networks*, 2015, vol. 64, pp. 59–63, doi: 10.1016/j.neunet.2014.09.005.

[10] R. Amimi, A. Radgui, and H. Ibn El Haj El, "A survey of smart classroom: concept, technologies and facial emotions recognition application," in *Lecture Notes in Networks and Systems*, 2023, vol. 544 LNNS, pp. 326–338, doi: 10.1007/978-3-031-16075-2_23.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. I-511-I–518, doi: 10.1109/cvpr.2001.990517.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2000, vol. 1842, pp. 404–420, doi: 10.1007/3-540-45054-8_27.

[14] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019, [Online]. Available: http://arxiv.org/abs/1905.00641.

[15] M. Zanotti, "Transfer learning in image classification: how much training data do we really need?," *Towards data science.* 2020, Accessed: Aug. 29, 2022. [Online]. Available: https://towardsdatascience.com/transfer-learning-in-image-classification-how-much-training-data-do-we-really-need-7fb570abe774.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 2, pp. 1097–1105, 2012.

[17] A. G. Howard *et al.*, “MobileNets: efficient convolutional neural networks for mobile vision applications.” *arXiv preprint*, 2017, doi: 10.48550/arXiv.1704.04861.

[18] J. Caron and J. R. Markusen, “Paperspace.” pp. 1–23, 2016, [Online]. Available: https://docs.paperspace.com/.

[19] 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.

[20] C. Pabba and P. Kumar, “An intelligent system for monitoring students’ engagement in large classroom teaching through facial expression recognition,” *Expert Systems*, vol. 39, no. 1, p. e12839, 2022, doi: 10.1111/exsy.12839.

[21] S. Ma et al., “A face sequence recognition method based on deep convolutional neural network,” in *Proceedings - 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science*, DCABES 2019, 2019, pp. 104–107, doi: 10.1109/DCABES48411.2019.00033.

[22] J. Shen, H. Yang, J. Li, and Z. Cheng, “Assessing learning engagement based on facial expression recognition in MOOC’s scenario,” *Multimedia Systems*, vol. 28, no. 2, pp. 469–478, 2022, doi: 10.1007/s00530-021-00854-x.

[23] Y. Hu, Z. Jiang, and K. Zhu, “An optimized CNN model for engagement recognition in an e-learning environment,” *Applied Sciences (Switzerland)*, vol. 12, no. 16, p. 8007, 2022, doi: 10.3390/app12168007.

[24] B. Hdioud and M. E. H. Tirari, “Facial expression recognition of masked faces using deep learning,” *IAES International Journal of Artificial Intelligence IJ-AI*, vol. 12, no. 2, pp. 921–930, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp921-930.

[25] A. Rajae, R. Amina, and I. El Haj El Hassane, “Towards face-to-face smart classroom that adheres to covid’19 restrictions,” in *Lecture Notes in Networks and Systems*, 2023, vol. 625 LNNS, pp. 47–58, doi: 10.1007/978-3-031-28387-1_5.

BIOGRAPHIES OF AUTHORS

Amimi Rajae is a researcher at the National Institute of Posts and Telecommunications in Morocco. She holds an engineering diploma in Networks and Information Technologies with a specialization in Multimedia. Amimi’s research interests span across image processing, machine learning, image analysis, and pattern recognition. With a passion for advancing knowledge and innovation, she is dedicated to exploring the intersection of technology and education to enhance learning experiences. Amimi actively contributes to research in her field and seeks to make meaningful contributions to the academic community. For inquiries or collaborations. She can be contacted at email: amimi.rajae004@gmail.com.

Radgui Amina is an associate professor in computer vision at INPT Morocco. She obtained the M.Sc. and the Ph.D. degrees in computer sciences and telecommunications in 2004 and 2010, respectively, at the Mohammed V University. In 2011, she joined the Institute of Posts and Telecommunications (INPT) as assistant professor. She is also a member of the multimedia, signal, and communication systems (MUSICS) research team, department at INPT since 2011. Her research interests include computer vision, image processing, and the related applications for medical research. She can be contacted at email: radgui@inpt.ac.ma.

Ibn El Haj El Hassane is a professor of multimedia signal processing and communications at the INPT, Rabat, Morocco, from 2000 until now. He is also member of the MUSICS research group at the INPT. He is senior member of IEEE. His research interests include digital signal, speech processing, image and video processing, digital watermarking and data hiding, deep learning, cognitive radio, and multimedia communications. He can be contacted at email: ibnelhaj@inpt.ac.ma.