Deep Learning Based Mobilenet and Multi-Head Attention Model for Facial Expression Recognition

Aicha Nouisser
Faculty of Sciences of Gafsa,
University of Gafsa,
Tunisia

Ramzi Zouari
National School of Engineering of Sfax,
University of Sfax,
Tunisia

Monji Kherallah
Faculty of Sciences of Sfax,
University of Sfax,
Tunisia

Abstract: *Facial expressions is an intuitive reflection of a person's emotional state, and it is one of the most important forms of interpersonal communication. Due to the complexity and variability of human facial expressions, traditional methods based on handcrafted feature extraction have shown insufficient performances. For this purpose, we proposed a new system of facial expression recognition based on MobileNet model with the addition of skip connections to prevent the degradation in performance in deeper architectures. Moreover, multi-head attention mechanism was applied to concentrate the processing on the most relevant parts of the image. The experiments were conducted on FER2013 database, which is imbalanced and includes ambiguities in some images containing synthetic faces. We applied a pre-processing step of face detection to eliminate wrong images, and we implemented both SMOTE and Near-Miss algorithms to get a balanced dataset and prevent the model to being biased. The experimental results showed the effectiveness of the proposed framework which achieved the recognition rate of 96.02% when applying multi-head attention mechanism*.

Keywords: *Depthwise, pointwise, attention, balanced, skip connection, transfer learning.*

Received March 31, 2023; accepted May 10, 2023
https://doi.org/10.34028/iajit/20/3A/6

1. Introduction

Facial expressions recognition is the most way to identify the psychological and emotional state of persons. It can be widely applied in our real life. In fact, it is an effective way of early detection of health problems. Moreover, it helps services providers to evaluate customer satisfaction, and it is considered as concrete evidence to uncover whether an individual is speaking the truth in the context of criminal investigations [6]. With the recent advancements in computer vision and especially artificial intelligence, facial expression recognition systems have shown great potential due to their powerful automatic feature extraction and computational efficiency. However, some difficulties may be faced related to facial accessories, non-uniform illuminations, pose variations, etc., [1]. This leads to obtain a lot of unnecessary and misleading features that degrade the model effectiveness and make the emotion recognition more challenging.

The key concept in facial emotion recognition consists of detecting specific areas in the face like eyes, eyebrow, mouth, noise, and jaw. The emotion can be identified by calculating distances and angles between these different parts. Detecting landmarks on the face is another way to identify the facial expressions [15]. Each emotion is defined by specific positions of these landmarks. These measures are used with machine learning models like k-nearest neighbors, Naïve Bayes, support vector machines and decision trees to make emotion classification [7]. In the last decade, deep learning models and especially Convolution Neural Networks have been widely applied for facial emotion recognition [29]. This is due to their ability to automatically extract high level features from input images [23]. Therefore, it is almost useless to apply handcrafted feature extraction since deep learning models can made both extraction and classification tasks. However, data preprocessing like face detection, data normalization remains essential to ensure the model performance.

The rest of this paper is organized as follow: the next section presents a literature review of facial expression recognition. Section 3 describes the techniques and methods used in the proposed framework. Section 4 discusses the experimental results, and we finish by a conclusion and some prospects.

2. State of the Art

With the advent of variant of deep neural networks, Convolution Neural Networks (CNN) have become a general trend for Facial Expression Recognition (FER) problem. These approaches have replaced the classical and analytical methods, thank to their effectiveness in extracting relevant features from data [5]. In this context, several FER systems based CNNs have been proposed. Zhu *et al*. [30] introduced a convolutional

relation network based few-shot learning to exploit the feature similarity among training samples. This concept was suitable for solving some emotion categories with limited training samples. The proposed approach was evaluated on FER2013 dtaaset and achieved the recognition accuracy of 67.32%. In other work, Khairuddin *et al*. [13] used a pretrained VGG-Net architecture on Fer2013 and achieved the accuracy of 73.28%. Ab-Wahab and *et al*. [1] proposed a new recognition system based EfficientNet-Lite model. In the classification part, *k*-Nearest Neighbors algorithm was applied instead of dense and SoftMax layers. Experiments were done on Fer2013 database and obtained the accuracy of 75.26%. Tripathi *et al*. [25] experimented three deep learning models based AlexNet, VGG19 and ResNet50 respectively. These models were trained with transfer learning and reached the accuracy of 91.89% on FER2013 dataset when using VGG19 model. Lee *et al*. [16] presented an efficient CNN, called EmotionNet, designed for facial emotion classification. This model is composed and prototyping and exploration stages respectively. The first stage is based on residual networks and provides features to the exploration stage which produces the final decision using analytical approaches. This model was evaluated on CK+ dataset and achieved the accuracy of 92.7%. In [24], a new approach based VGG-Spinal network was proposed. In fact, the VGG-16 classification layers were replaced with SpinalNet layers to improve the classification on FER2013 dataset. Pathak *et al*. [22] proposed a recognition model based on pretrained MobileNet model and obtained the accuracy of 71% on CK+ dataset. Recently, Nouisser *et al.* [20] presented a new FER system in which skip connections were added to MobileNet model. Their system achieved the accuracy of 95.46% on FER2013 dataset.

In several works, the combination of multiple models was applied to improve the model generalization. Yaseen *et al*. [28] proposed a hybrid system based on primary and secondary basic CNNs (P-CNN and S-CNN). The first model is designed to classify the image based on the two primary emotions of showing happy or sad. The second CNN performs a final classification based on the sub-category of emotions from the results of the P-CNN. They obtained the accuracies of 94.12% and 97.07% on JAFFE and FER2013 databases respectively. Bahri *et al*. [3] used hybrid deep learning model containing AlexNet and VGG19 networks. A crop transformation was applied on images to focus on the face. They reached the accuracies of 96.970% and 70.24% on CK+ and FER2013 databases. In addition to model hybridization, data augmentation was applied increase the number of training samples to ensure the model generalization. This technique was applied by Kumar *et al*. [14]. They did slight rotation and inclination on images and obtained the accuracy of 83% on Fer2013 database when using a baseline CNN. In [4], horizontal flip, shift, scaling, and rotation are the data augmentation methods used to increase the size of dataset. This approach is combined with FerConvNet model and achieved the accuracy of 85% on Fer2013 validation dataset.

Based on the literature review, almost of FER systems are based on pretrained deep neural networks. In this study, we opted for MobileNet architecture thanks to its great performance and low memory consumption. Moreover, we applied skip connections and multi-attention model to MobileNet in order to improve the model performances.

3. Proposed Framework

The proposed framework was evaluated on FER2013 dataset, which has made a problem of overfitting and do not give good results for the validation and test subsets. To solve this problem, we inspected the dataset to visualize the images. We found that some images are not fully focused on the face or do not contain faces. So, we did face cropping to maintain the face region and discarding irrelevant parts of the image. On the other hand, we notice that FER2013 is unbalanced dataset where images are uniformly distributed in the different classes. This can lead to a biased model. To overcome this problem, we proposed applying both over-sampling and over-sampling techniques to obtain uniformly distributed database.

3.1. Image Pre-Processing

This step consists of browsing all images in the database, detecting face landmarks, and cropping the images according to these key points. We followed the method used in [12], where Histogram of Oriented Gradient was applied. Thereafter, a linear Support Vector Machines (SVM) was applied to decide whether the pixels inside a sliding window are relevant or not. This method allows defining 68 landmarks relative to eyes, eyebrow, nose, lips, and face contour. This processing allows us to obtain images containing only the face region. However, some images were removed from the database because they do not contain images of faces. Figures 1 and 2 show the result of the application of the detector for both right and wrong face images from FER2013 database.

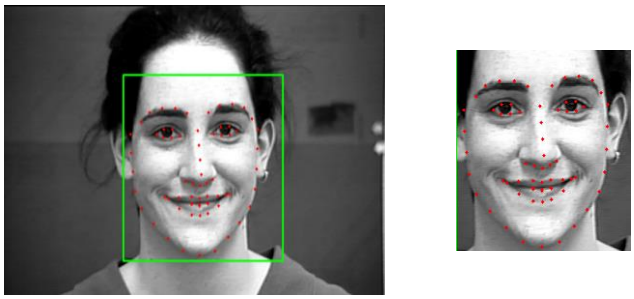

Figure 1. 68 facial landmarks detection.

Figure 2. Wrong facial images from FER2013 dataset.

3.2. Database Balancing

After applying data processing, we obtained an unbalanced dataset where classes are not represented equally. In this case, the classifier will be biased towards the majority class. To handle with this problem, several techniques have been proposed like oversampling and under-sampling [11]. Oversampling is the process of generating new samples from the minority class, while under-sampling reduces the samples of the majority class to match the number of samples in the minority class. This can be done by removing random selecting samples, or those with a minimum average distance to further minority class. Park *et al*. [21] suggested that combining both over and under sampling leads to obtain most lifelike dataset and accurate results. For this reason, we applied these two techniques only on FER2013 training dataset to maintain the same structure of validation and test subsets, and to be able to correctly evaluate the performance of our model.

In the under-sampling step, we used Near-Miss method which involves 2-steps algorithm [18]. First, for each minority class sample, their *m* nearest neighbors will be kept. Then, the selected majority class samples are the ones for which the average distance to the *m* nearest neighbors is the largest (Figure 3). In the oversampling step, we applied Synthetic Minority Over sampling (SMOTE) technique [6]. This approach is mainly adapted for oversampling tabular data. It was inspired by a technique that proved successful in handwritten character recognition. Oversampling has the advantage of generating new samples that do not perturb the training data. For each minority class sample, synthetic samples are generated along the line segments joining it with the k-nearest neighbors (Figure 4).

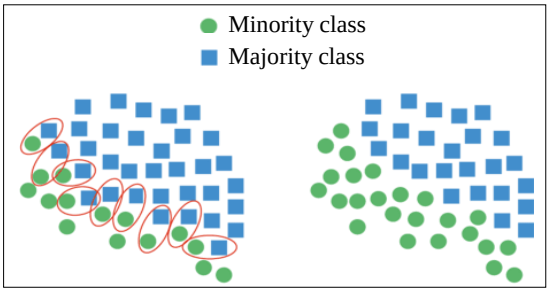

Figure 3. Ner-Miss under-sampling concept.

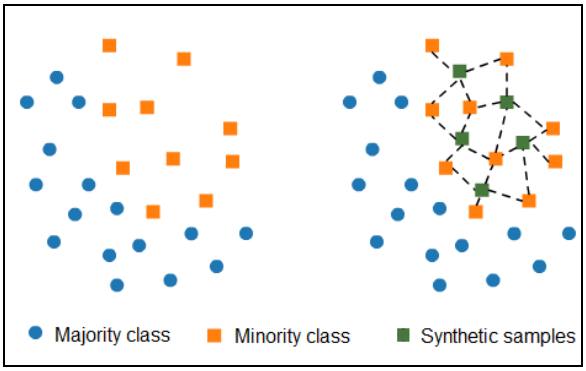

Figure 4. SMOTE oversampling concept.

The generated samples will be added to the training data to form the whole training database. After this stage, we obtained a balanced database distributed as described in Table 1.

Table 1. FER2013 training dataset distribution.

	Samples before balancing	**Samples after balancing**
Happy	7215	2652
Neutral	1965	2694
Sad	4830	2677
Fear	4097	2668
Angry	3995	2701
Surprise	3171	2554
Disgust	436	2657

3.3. Recognition Model

The proposed model is based on deep neural networks and more precisely CNN, which has become dominant in various computer vision tasks. Since facial emotion recognition is intended to be implemented on embedded device, we opted for MobileNet architecture which is a class of efficient models designed for mobile and embedded vision applications [9]. MobileNet is a streamlined architecture that uses depthwise separable convolutions to build light weight deep neural networks. It consists of a depthwise convolution, i.e., a spatial convolution performed independently over every channel of an input, followed by a pointwise convolution, i.e. a regular convolution with 1x1 windows, projecting the channels computed by the depthwise convolution onto a new channel space. The depthwise separable convolution operation should not be confused with spatially separable convolutions,

which are also often called separable convolutions in the image processing. Their mathematical formulation is as follow (Equation 1):

$$G_{k,l,m} = \sum_{i,j} K_{i,j,m} \times F_{(k+i-1,l+j-1,m)} \quad (1)$$

With K is the depthwise kernel of size ($D_k \times D_k \times M$). The m^{th} filter is applied to the m^{th} channel in the feature maps F, to produce the m^{th} channel of the output G. The depth-wise convolution does not combine the input channels to create new features like conventional convolution. Thus, pointwise is used to apply a linear combination of the depthwise outputs.

In standard convolution, with N kernels of size ($D_K \times D_K \times M$), there are ($N \times D_K \times D_K \times M$) multiplications every time the kernel moves. However, in the depth-wise convolution there are ($D_K \times D_K \times M$) multiplications, plus ($N \times 1 \times 1 \times M$) other multiplications in the pointwise convolution [10]. Thus, there is much lower trainable parameters to optimize. With less computations, the network is able to process in a shorter amount of time. Figures 5 and 6 illustrates the principle of depthwise and pointwise convolutions respectively.

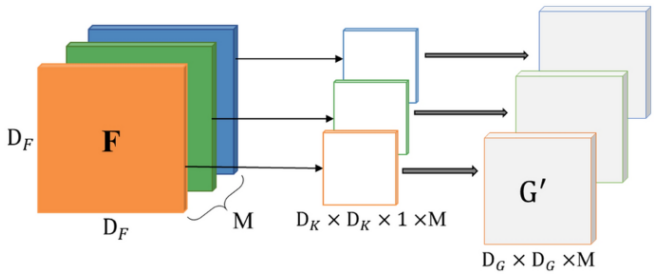

Figure 5. Depth-wise convolution layer.

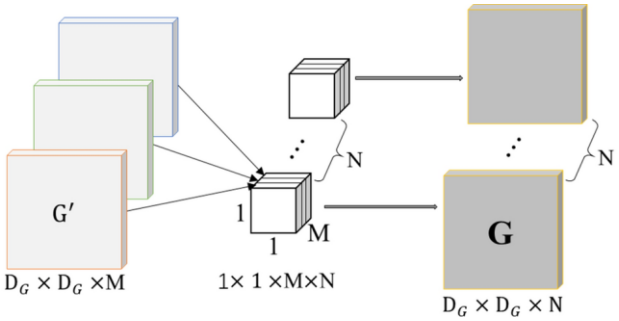

Figure 6. Point-wise convolution layer.

MobileNets architecture begins with a regular convolution followed by 26 successive depth-wise and pointwise layers of different number of kernels. In the classification part, Average Pooling layer is applied before the final output layer. Due to the large number of layers, several problems can be occurred such as vanishing and/or exploding gradient [27]. To overcome this problem, we decided to add skip connections between successive depthwise and pointwise convolution layers (Figure 7).

A part of skip connections, we implemented Multi-Head Attention (MHA) mechanism to improve the model effectiveness (Figure 8). In fact, MHA is a mechanism used to provide an additional focus on a specific component in the data. It enables the network to concentrate on a few aspects at a time and ignoring the rest [26]. MHA consists of several attention layers running in parallel, instead of performing one single attention function. In particular, the input consists of queries and keys of dimension d_k (Q and K respectively), and values of dimension d_v (V). The output of the attention model is done by computing the scaled dot product of the queries with all keys and applying a SoftMax function to obtain the weights on the values V (Equation 2). The attention mechanism is linearly projected h times with different learned weights (W^Q, W^K, W^V). These different representation subspaces are concatenated into one single attention head to form the final output result (Equation 3).

$$Attention(Q,K,V) = softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \quad (2)$$

$$\begin{cases} MultiHead(Q,K,V) = Concat(head_1,...,head_h) \\ head_i = Attention(QW^Q, KW^K, VW^V) \end{cases} \quad (3)$$

In our study case, MHA was applied to the output of each depthwise separable convolution layer. We used a particular version of attention model called self-attention, in which query, key and value inputs are the same. The calculation process follows these steps: First, we made the dot product (MatMul) of query and keys tensors and scale the obtained scores. Next, we apply a SoftMax function on these scores to obtain attention probabilities. Finally, we take a linear combination of these distributions with the value input tensors and concatenate them into one channel.

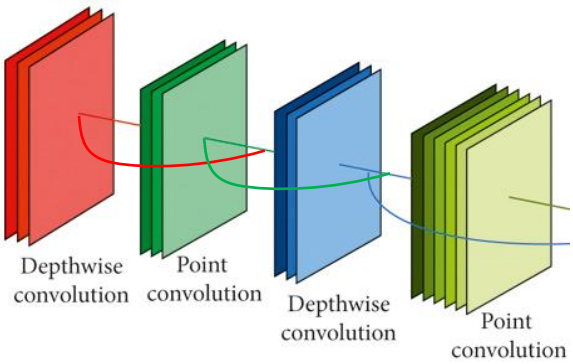

Figure 7. MobileNet with skip connections architecture.

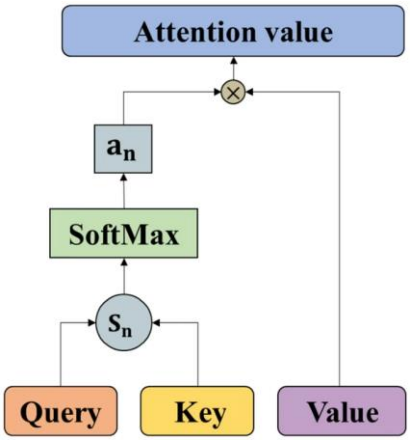

Figure 8. Attention model architecture.

4. Experiments and Results

4.1. Database Description

The experiments were done on FER2013 database [8]. It was firstly used in ICML 2013 Workshop in the contest of Challenges in Representation Learning. The dataset contains 35887 human facial images non uniformly distributed, resized to (48×48) pixels and converted to grayscale. FER2013 is initially divided into training set (28709 images), a public test set (3589 images) used as validation set, and a private test set (3589 images), usually considered as the test set for final evaluations. After applying face cropping, under sampling, and oversampling steps, we obtained a balanced dataset with different images sizes (section 3). So, we resized the images in (48×48) pixels using padding technique instead of standard resize which introduce unnecessary stretching. When applying padding technique, the content of the original image is kept as faithful as possible, and the aspect ratio is also preserved. Figure 9 shows some images generated by SMOTE algorithm.

Figure 9. Images generated by SMOTE algorithm.

4.2. Experimental Results

The recognition model is mainly based on MobileNet model, with the addition of skip connections and multi-head attention mechanism in the extraction part. During the experiments, the training dataset was divided into batches of size 32, with shuffling option to make different min-batch samples in each epoch. Moreover, categorical cross entropy method was used to compute the loss between desired and calculated outputs at each iteration [17]. The model was trained using Adam (Adaptive Moment Estimation) optimizer with an initial learning rate of 0,001. This value can be reduced by a factor of 0.5 once learning stagnate (Table 2). Moreover, early stopping approach is applied as a regularization method. It consists of stopping the training process early before it has overfit the training dataset. In the multi-head attention model, we employ 8 parallel attention layers or heads. For each of these, we use 64 units for both query, key and value linear projections.

In the experiments, three models were evaluated: standard MobileNet, MobileNet with skip connections, and MobileNet with both skip connections and multi-head self-attention models. In all experiments, transfer learning was applied during the training stage. In fact, the weights of the extraction part are considered as non-trainable parameters. However, the parameters of skip connections, multi-head attention, and extraction parts are to optimize. The best performance was obtained by the third model which achieves the validation accuracy of 96.02% after 75 epochs (Table 3). Furthermore, the learning rate was decreased from 0.001 to 0.00025.

Table 2. Hyperparameters configuration.

Hyperparameter	Value
Batch size	32
Optimizer	Adam
Learning rate (Lr)	0.001
Learning rate decrease factor	0.5
Epochs	100

Table 3. Experimentation results.

Model	Accuracy	Recall	Precision
MobileNet	0.9127	0.7821	08214
MobileNet + skip connections	0.9564	0.8169	0.8703
MobileNet + Skip connection + MHA	**0.9602**	**0.8214**	**0.8961**

4.3. Comparative Study and Discussion

The proposed model achieves the state-of-the-art performance on FER2013 dataset, although the other approaches are based on efficient deep learning models (Table 4). The obtained result is justified by the importance of the steps implemented in the recognition framework. In fact, we have eliminated all irrelevant information from images thanks to the face cropping method. In addition, we discovered wrong images that does not contain faces. Moreover, we applied under sampling and oversampling techniques to obtain balanced dataset and prevent the model to being biased. All these preprocessing steps allowed us to obtain tidy data. On the other hand, the addition of skip connections and multi-head attention mechanism have an impact in improving the model effectiveness.

Table 4. Experimentation results.

Work	Model	Accuracy (%)
[13]	VGG-Net	73.28
[1]	CNN-KNN	75.26
[3]	EfficientNet-VGG16	90.6
[25]	VGG-19	91.89
[20]	MobileNet + Skip connections	95.46
Our work	**MobileNet + Skip connections + MHA**	**96.02**

5. Conclusions

In this study, we developed a new system for facial expression recognition based on MobileNet with the addition of skip connections and multi-head attention mechanism. We have applied data preprocessing on FER2013 dataset to eliminate irrelevant parts in the images and crop them on the face area. To overcome the disadvantages of unbalanced dataset, under-sampling and oversampling were applied to obtain balanced dataset. In the training stage, several techniques like early stopping, minibatch shuffling and learning rate scheduling were applied to prevent the model from overfitting problem. As perspective, we intended to develop a new recognition system based on vision transformers which is applied over patches of the image.

References

[1] Ab Wahab M., Nazir A., Ren A., Noor M., Akbar M., and Mohamed A., "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi," *IEEE Access*, vol. 9, pp. 134065-134080, 2021. DOI: 10.1109/ACCESS.2021.3113337

[2] Amato, G., Falchi, F., Gennaro, C., and Vairo, C. "A Comparison of Face Verification with Facial Landmarks and Deep Features," *in Proceedings of 10th International Conference on Advances in Multimedia*, Athens, pp. 1-6, 2018.

[3] Bahri S., Samsinar R., and Denta P., "Pengenalan Ekspresi Wajah untuk Identifikasi Psikologis Pengguna Dengan Neural Network dan Transformasi Ten Crops," *RESISTOR (Elektronika Kendali Telekomunikasi Tenaga Listrik Komputer)*, vol. 5, no. 1, pp. 15-20, 2022. https://doi.org/10.24853/resistor.5.1.15-20

[4] Bodavarapu P. and Srinivas P., "Facial Expression Recognition for Low Resolution Images Using Convolutional Neural Networks and Denoising Techniques," *Indian Journal of Science And Technology*, vol. 14, no. 12, pp. 971-983, 2021. https://doi.org/10.17485/IJST/v14i12.14

[5] Canal F., Müller T., Matias J., Scotton G., De Sa Junior A., Pozzebon E., and Sobieranski A., "A Survey on Facial Emotion Recognition Techniques: a State-of-The-Art Literature Review," *Information Sciences*, vol. 582, pp. 593-617, 2022. https://doi.org/10.1016/j.ins.2021.10.005

[6] Chawla N., Bowyer K., Hall L., and Kegelmeyer W., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. https://doi.org/10.1613/jair.953

[7] Chowdary M., Nguyen T., and Hemanth D., "Deep Learning-Based Facial Emotion Recognition for Human-Computer Interaction Applications," *Neural Computing and Applications*, pp. 1-18, 2021.

[8] Goodfellow I., Erhan D., Carrier P., Courville A., Mirza M., Hamner B., … , and Bengio Y., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *in Proceedings of International Conference on Neural Information Processing*, Daegu, pp. 117-124, 2013.

[9] Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., and Adam H., "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.0486*, 2017. https://doi.org/10.48550/arXiv.1704.04861

[10] Kaiser L., Gomez A., and Chollet F., "Depthwise Separable Convolutions for Neural Machine Translation," *arXiv preprint arXiv:1706.03059*, 2018. https://doi.org/10.48550/arXiv.1706.03059

[11] Kaur P. and Gosain A., "Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," *in Proceedings of ICT Based Innovations*, Singapore, pp. 23-30, 2018.

[12] Kazemi V. and Sullivan J., "One Millisecond Face Alignment with an Ensemble of Regression Trees," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, pp. 1867-1874, 2014.

[13] Khairuddin Y. and Chen Z., "Facial Emotion Recognition: State of the Art Performance on FER2013," *arXiv preprint arXiv:2105.03588*, 2021. https://doi.org/10.48550/arXiv.2105.03588

[14] Kumar K. and Reddy Y., "Facial Emotion Recognition Using Machine Learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 4, no. 4, pp. 1828- 1833, 2022. https://doi.org/10.31979/etd.w5fs-s8wd

[15] Kumar Y., Verma S., and Sharma S., "Multi-Pose Facial Expression Recognition Using Hybrid Deep Learning Model with Improved Variant of Gravitational Search Algorithm," *The*

International Arab Journal on Information Technology, vol. 19, no. 2, pp. 281-287, 2022. https://doi.org/10.34028/iajit/19/2/15

[16] Lee J. R., Wang L., and Wong A., "Emotionnet Nano: An Efficient Deep Convolutional Neural Network Design for Real-Time Facial Expression Recognition," *Frontiers in Artificial Intelligence*, vol. 3, pp. 1-9, 2021. https://doi.org/10.3389/frai.2020.609673

[17] Li B., Yao Y., Tan J., Zhang G., Yu F., Lu J., and Luo Y., "Equalized Focal Loss for Dense Long-Tailed Object Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Louisiana, pp. 6990-6999, 2022.

[18] Mani I. and Zhang I., "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction," *in Proceedings of Workshop on Learning from Imbalanced Datasets*, Washington, pp. 1-7, 2003.

[19] Mehendale N., "Facial Emotion Recognition Using Convolutional Neural Networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, pp. 1-8, 2020.

[20] Nouisser A., Zouari R., and Kherallah M., "Enhanced Mobilenet And Transfer Learning For Facial Emotion Recognition," *in Proceedings of the International Arab Conference on Information Technology*, Abu Dhabi, 2022. 10.1109/ACIT57182.2022.9994192

[21] Park S. and Park H., "Combined Oversampling and Undersampling Method Based On Slow-Start Algorithm for Imbalanced Network Traffic," *Computing*, vol. 103, no. 3, pp. 401-424, 2021.

[22] Pathak A., Bhalsing S., Desai S., Gandhi M., and Patwardhan P., "Deep Learning Model for Facial Emotion Recognition," *in Proceedings of ICETIT: Emerging Trends in Information Technology*, Delhi, pp. 543-558, 2019.

[23] Pecoraro R., Basile V., Bono V., and Gallo S., "Local Multi-Head Channel Self-Attention for Facial Expression Recognition," *arXiv preprint arXiv:2111.07224*, 2021. https://doi.org/10.3390/info13090419

[24] Santoso B. and Kusuma G., "Facial Emotion Recognition on Fer2013 Using Vggspinalnet," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 7, pp. 2088-2102, 2022.

[25] Tripathi M., "Face Emotion Recognition Using A Convoluting Neural Network," *Journal on Image and Video Processing*, vol. 12, no. 1, pp. 2531-2536, 2021.

[26] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., and Polosukhin I., "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1-11, 2017.

[27] Wang W., Li Y., Zou T., Wang X., You J., and Luo Y., "A Novel Image Classification Approach Via Dense-Mobilenet Models," Mobile Information Systems, pp. 1-8, 2020. https://doi.org/10.1155/2020/7602384

[28] Yaseen A., Shaukat A., and Alam M. "Emotion Recognition From Facial Images Using Hybrid Deep Learning Models," *in Proceedings of 2nd International Conference on Digital Futures and Transformative Technologies*, Rawalpindi, pp. 1-7, 2022. 10.1109/ICoDT255437.2022.9787474

[29] Zahara L., Musa P., Wibowo E. P., Karim I., and Musa S. B., "The Facial Emotion Recognition (FER-2013) Dataset For Prediction System of Micro-Expressions Face Using The Convolutional Neural Network (CNN) Algorithm Based Raspberry Pi," *in Proceedings of 5th International Conference on Informatics and Computing*, Gorontalo, pp. 1-9, 2020. 10.1109/ICIC50835.2020.9288560

[30] Zhu Q., Mao Q., Jia H., Noi O., and Tu J., "Convolutional Relation Network for Facial Expression Recognition in the Wild With Few-Shot Learning," *Expert Systems with Applications*, vol. 189, pp. 1-9, 2022. https://doi.org/10.1016/j.eswa.2021.116046

Aicha Nouisser, she is a PhD student at the faculty of sciences of Gafsa, Tunisia. She obtained a master's degree in computer science for the University of Gafsa in 2022. Her main research interest concerns the application of deep learning models for biometrics.

Ramzi Zouari, he is a PhD in Engineering of informatic systems. He obtained the doctorate diploma from the national school of engineering of Sfax, Tunisia in 2020. His research interests include handwriting trajectory modeling and recognition. Moreover, he focuses his research on the application of deep neural networks models on smart applications.

Monji Kherallah, received the Ing. Diploma degree and the PhD in electrical engineering, in 1989 and 2008, respectively from University of Sfax (ENIS), Tunisia. He is a member in Research Group of Intelligent Machines: REGIM. His research interest includes the following projects: "i-Bag", "i-House" and "i-Car". The techniques used are based on methods intelligent, such as neural network, logic fuzzy, genetic algorithm, etc. He is a reviewer of several international journals.