Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **1**, March **2026**

Facial Expression Recognition via Transfer Learning: A Comparative Study on Deep CNNs with fine tuning strategies

Hanane Hadjira Nedjar[1,*]

[1]SIMPA Laboratory
University of Science and Technology of Oran – Mohamed Boudiaf, Algeria
hananehadjira.nedjar@univ-usto.dz

Abdelkrim Mebarki[1]

[1]University of Science and Technology of Oran – Mohamed Boudiaf, Algeria
abdelkrim.mebarki@univ-usto.dz

Sara Samra Benharrats[2]

[2]University of Oran 1 - Ahmed Ben Bella
Sidi Chami Psychiatric Hospital
benharrats.sarra@univ-oran1.dz

*Corresponding author: Hanane Hadjira Nedjar

Received August 17, 2025, revised December 19, 2025, accepted December 22, 2025.

Abstract. *Facial Expression Recognition (FER) is a crucial component in affective computing and human-computer interaction, enabling systems to interpret human emotions from visual cues. This study presents a comparative analysis of seven state-of-the-art transfer learning models—ResNet18, ResNet50, VGG16, VGG19, DenseNet121, MobileNetV2, and EfficientNetB0—for FER tasks using the CK+ dataset. A uniform preprocessing pipeline and training configuration were applied to ensure fair evaluation across all models. The results reveal that ResNet18 and DenseNet121 achieve the highest classification accuracy (94.93%) with relatively low computational costs. The study further applies a progressive fine-tuning strategy, improving performance stability and training efficiency. These findings offer practical insights for selecting and adapting deep learning architectures for emotion recognition in constrained or real-time environments*
Keywords: Facial Expression Recognition (FER), Transfer Learning, Deep Convolutional Neural Networks, CK+ Dataset, Fine-tuning, Emotion Recognition.

1. **Introduction.** Facial expression recognition (FER) has become a critical task in computer vision, particularly in the fields of human-computer interaction, social robotics, affective computing, and behavioral analysis. The ability to automatically recognize human emotions from facial images can significantly enhance the development of intelligent systems that respond to user emotions in real-time [1]. Traditional FER approaches typically rely on handcrafted features, like LBP (Local Binary Patterns) [2], HOG (Histogram of Oriented Gradients) [3], or GF (Gabor Filters) [4], followed by classification algorithms like SVM (Support Vector Machines) or KNN (K-Nearest Neighbors) [5]. While these methods have shown promising results in controlled settings, they often struggle to generalize across the different FER databases and in real-world scenarios due to their sensitivity to features extraction methods. With the advent of deep learning, convolutional neural networks (CNNs) have demonstrated outstanding performance in image classification tasks, including FER [6]. However, training deep models from scratch requires large annotated datasets [6], which are not always available for facial expression analysis. To address this challenge, transfer learning has emerged as an effective solution by leveraging the feature representations

learned from large-scale datasets, such as ImageNet, and fine-tuning them for specific tasks with smaller datasets [7]. In this work, we conduct a comprehensive comparison of seven popular transfer learning models: ResNet18, ResNet50, VGG16, VGG19, DenseNet121, MobileNetV2, and EfficientNetB0 on the CK+ dataset, a well-known benchmark for facial expression recognition. The models are evaluated under the same training and testing conditions to ensure a fair comparison and allows for objective selection of best-performing models for the next progressive fine-tuning step.

2. **Related work on deep learning for facial expression recognition.** Facial mimicry convey a wide spectrum of emotions that are universally recognizable, making facial expression recognition (FER) a vital component of affective computing. FER has found extensive applications in various domains, including driver fatigue detection, human computer interaction, assistive robotics, digital entertainment, and healthcare. In medical contexts, FER contributes to understanding patients emotional states, which supports early intervention and improves the quality of care [8]. Additionally, facial recognition systems are increasingly integrated into security frameworks to enhance privacy and authentication. Deep learning, as a specialized branch of machine learning, employs multilayered artificial neural networks to automatically extract and learn hierarchical representations from data [9]. Within the domain of computer vision, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged as leading architectures, particularly noted for their effectiveness in feature extraction from raw inputs. In the context of image analysis, CNNs are especially well-suited, as their convolutional layers systematically capture spatial patterns and feature hierarchies an essential capability for accurate facial expression recognition. Numerous studies in the literature have explored CNN-based architectures for facial expression recognition (FER). For instance, the authors in [10] implemented a conventional CNN composed of two convolutional-pooling layers to classify facial expressions from a self-collected image dataset. In a more complex approach, Mollahosseini et al. [11] extended the classical CNN architecture by incorporating four inception modules, thereby enhancing the network's depth and capacity for feature abstraction based on spatio-temporal manifold (STM) modeling and expressionlet learning (STM-ExpLet). Ruiz-Garcia et al. [12] proposed initializing the CNN using encoder weights derived from a stacked convolutional autoencoder. This weight initialization strategy demonstrated superior performance compared to conventional random initialization methods. In another study, authors [13] investigated a hybrid deep learning architecture that combined CNNs with Recurrent Neural Networks (RNNs) to capture both spatial and temporal dependencies in facial expression data. Additionally, Liliana [14] employed a deep CNN model consisting of 18 convolutional layers and four subsampling layers, aimed at achieving robust FER performance through a deeper hierarchical representation. This study is fundamentally grounded in the use of pre-trained deep learning models and transfer learning (TL) techniques. A range of established pre-trained architectures is examined to determine the most effective model for facial expression recognition (FER). The following subsections offer a concise overview of Convolutional Neural Network (CNN) architectures and the rationale for employing transfer learning in this context.

2.1. **Convolutional Neural Network (CNN).** Owing to its inherent structural design, the Convolutional Neural Network (CNN) is particularly well-suited for image-related tasks [15]. A typical CNN is composed of an input layer, multiple convolutional and pooling hidden layers, followed by an output layer. The core operation of convolution, a mathematical process involving two functions, produces a third function that represents a modified version of the original input. In the context of CNNs, small-sized kernels (e.g., 3*3 or 5*5) slide across the input image to extract useful local features through convolutional operations. Pooling, on the other hand, serves as a non-linear downsampling mechanism. It reduces spatial dimensions by aggregating information from non-overlapping regions, thereby producing a more compact feature representation. Figure 1 illustrates a basic CNN architecture comprising two convolutional-pooling layers. The first convolutional layer applies convolution to the input image, generating convolved feature maps (CFMs), which are then passed to a pooling layer to produce the first set of subsampled feature maps (SFMs). This process is repeated in the second convolutional-pooling layer. Subsequently, the output of the second pooling layer is flattened and passed to a fully connected (dense) layer, where each neuron is linked to all activations from the previous layer. The final layer, often referred to as the loss layer, defines the objective function and guides the training process by penalizing deviations between predicted and actual outputs. Such CNN architectures are widely adopted for pattern recognition tasks involving small-sized input images (e.g., 48*48), such as handwritten digit classification. Further details regarding CNN design and functionality can be found in the literature [16].

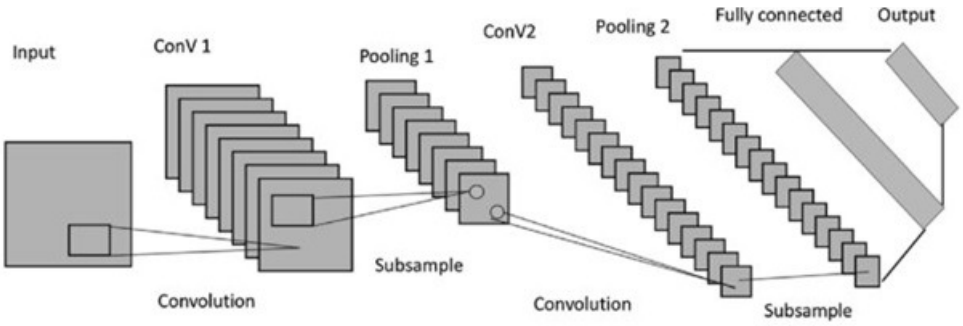

FIGURE 1. General scheme of a Transfer Learning (TL) CNN model

2.2. **Transfert Learning.** Facial Expression Recognition (FER) using a pre-trained Convolutional Neural Network (CNN) model through an appropriate transfer learning (TL) strategy constitutes the primary contribution of this work. Authors in [17] provided valuable insights into the internal representations learned by CNNs. Their findings indicate that the initial layers of a CNN capture low-level features such as edges and corners, while intermediate layers detect more complex patterns like textures and shapes. Higher layers progressively learn abstract and task-specific representations. Since low-level visual features are generally consistent across images, the representations learned by the lower layers of CNNs for FER are often transferable from models trained on other image-based tasks, such as object classification. Given the computational demands and data requirements of training CNNs from scratch with randomly initialized weights, leveraging pre-trained models and fine-tuning them via TL presents a more efficient and effective solution for FER.

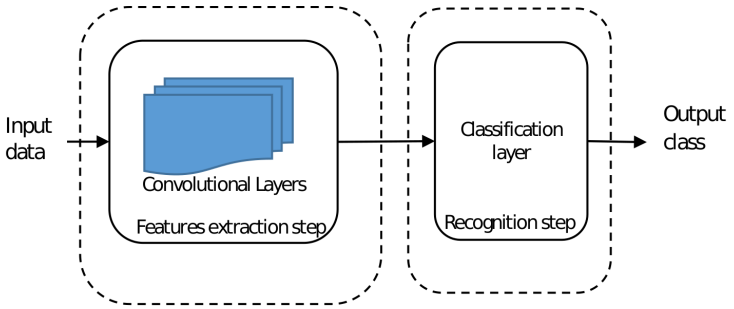

FIGURE 2. Basic topology of a Convolutional Neural Network (CNN) model

Figure 2 illustrates the general architecture of a transfer learning (TL)-based CNN model. In this configuration, the convolutional layers (originating from a pre-trained model) are used for low and mid-level feature extraction, while the original classification head is removed. In its place, a new task-specific classifier is added and fine-tuned on the target FER dataset to adapt the model to the specific classification and recognition task. The new added classifier component typically consists of one or more dense layers, which are fully connected and responsible for performing the final classification specific to the target task. Three commonly adopted strategies are used for fine-tuning in TL:

- training the entire model
- training a subset of layers while freezing the others
- training only the newly added classifier while keeping the convolutional base frozen.

For similar recognition tasks, fine-tuning only the classifier and/or a few upper layers is typically sufficient. Conversely, for tasks that are significantly different from the original recognition task, it becomes necessary to fine-tune the entire network to achieve optimal performance. Accordingly, the fine-tuning process is applied to the added classification layers and, depending on task similarity, to a selected subset or the entirety of the convolutional layers. CNNs are designed to handle complex image data using multiple hidden layers, which can make training difficult, especially with high-dimensional inputs. Each CNN architecture differs in how these layers are organized and connected, influencing performance and

efficiency [6]. One of the earliest model in large-scale image recognition came with AlexNet [18], which introduced a five-layer convolutional structure and achieved impressive results on the ImageNet dataset. Building on this, ZfNet [19] improved efficiency by replacing large convolutional filters with smaller ones, achieving similar accuracy with fewer parameters. While ZfNet required around 1 million images for training, AlexNet needed approximately ten times more data to achieve similar performance. Following these two models, the VGG family brought deeper networks. VGG-16 [20] introduced 13 convolutional layers and consistently used small 3×3 filters to increase depth while controlling complexity. VGG-19 extended this architecture further with 16 convolutional layers, offering even more representational power. Skip connections are an important idea used in many recent CNN models. This concept was first introduced in Residual Networks (ResNet) [21]. A skip connection takes the input of a layer and adds it to the output of a later layer. This helps the network keep useful information and makes training easier. It also reduces the problem of vanishing gradients. There are several versions of ResNet with different depths. Some common examples are ResNet-18, ResNet-50, and ResNet-101. The number in the name shows the total number of layers. However, the number of convolutional layers is usually one less than the depth. DenseNet [22] improved the idea of skip connections by introducing dense connections between layers. In this model, each layer receives inputs not only from the previous layer but from all earlier layers. Also, the output of each layer is passed to all following layers. This is done by concatenating the feature maps from the previous layers as input to the current layer. In a traditional CNN with N layers, there are N direct connections. In contrast, DenseNet has N(N+1)/2 direct connections. Because every layer has access to all earlier features, there is less information loss, and the network becomes more compact and efficient. However, this dense connectivity increases the number of input features in deeper layers, which can raise the computational cost. To reduce this, DenseNet uses 1×1 convolutions to shrink the number of channels, improving both speed and memory usage. Each layer's output is computed by applying a non-linear function to the concatenated outputs of all previous layers. DenseNet comes in different versions. For example, DenseNet-161 contains 157 convolutional layers organized into four modules. Recently, researchers have introduced new CNN architectures that are computationally efficient yet powerful, offering strong performance with fewer parameters. Among these, MobileNet and EfficientNet stand out as two popular lightweight models. MobileNet is designed for mobile and embedded applications [23]. It uses depthwise separable convolutions to reduce the number of computations and parameters while maintaining good accuracy [23]. This makes it ideal for real-time image processing tasks on resource-limited devices. On the other hand, EfficientNet introduces a compound scaling method that uniformly scales the model's depth, width, and input resolution [24]. It achieves state-of-the-art performance on many benchmarks using fewer parameters and lower computational cost compared to traditional models. Both architectures reflect a shift toward efficient deep learning models that are suitable for deployment in practical, real-world environments. Recent studies have highlighted the effectiveness of transfer learning in FER, where models pre-trained on large image datasets are fine-tuned on facial expression datasets. For instance, transfer learning with VGG-19, ResNet50V2, and DenseNet-121 has yielded significant performance improvements by selectively freezing layers and applying regularization strategies to mitigate overfitting. These methods have reported accuracy gains compared to earlier approaches.

3. **Methodology.** In this study, we evaluate and compare the performance of several deep learning models for facial expression recognition (FER) using the widely adopted benchmark datasets: CK+. These datasets were selected for their diversity in image quality, annotation schemes, and representativeness of human emotions, which support a robust comparative analysis.

3.1. **Extended Cohn-Kanade (CK+) Dataset.** The CK+ dataset includes 593 sequences from 123 subjects, with the final frame of each sequence annotated with one of seven emotion labels: anger, contempt, disgust, fear, happiness, sadness, and surprise. For this study, only the peak expression frames were used, resulting in approximately 327 labeled images. The dataset features both posed and spontaneous expressions under controlled conditions. Similar preprocessing steps were applied: conversion to RGB, resizing to 224×224 pixels, and normalization. Images were normalized using the mean and standard deviation values of the ImageNet dataset to ensure compatibility with pre-trained transfer learning models. Some samples images from the CK+ dataset are presented in Figure 3.

3.2. **Preprocessing.** To ensure consistency and compatibility with pretrained CNN models, all images from the, CK+ dataset underwent a standardized preprocessing pipeline. The preprocessing steps were as follows:

FIGURE 3. Samples images from CK+ dataset

- Cropping : Cropped face portion from the image is considered as input in the FER task to enhance facial properties.
- Resizing: All facial images were resized to 224×224 pixels, which is the standard input dimension required by most CNN-based architectures such as ResNet, VGG, DenseNet, and EfficientNet. This resizing step ensures uniformity across datasets and aligns with the expectations of models pretrained on ImageNet.
- Normalization: Image pixel values were normalized using the mean and standard deviation values of the ImageNet dataset, mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]. This step aligns the image intensity distribution with that of the training set used for pretrained models, improving convergence during fine-tuning.
- Dataset Splitting: Dataset was randomly split into training (70%), validation (15%), and test (15%) subsets. A fixed random seed was applied during the split to guarantee reproducibility of results across different runs. This strategy ensures a balanced and consistent evaluation of model performance during training, hyperparameter tuning,and final testing.

These preprocessing steps were crucial in preparing the datasets for the transfer learning experiments conducted in this study, ensuring fairness, consistency, and replicability across all comparative evaluations.

3.3. **Models.** In this study, we adopted transfer learning to leverage the feature extraction capabilities of deep convolutional neural networks pretrained on the ImageNet dataset. The following state-of-the-art architectures were selected due to their proven performance and varying depth, parameter complexity, and design paradigms:

- ResNet18
- ResNet50
- VGG16
- VGG19
- DenseNet121
- MobileNetV2
- EfficientNet-B0

These models span a wide range of computational complexities and depths, making them suitable for a comparative study of performance versus efficiency in facial expression recognition (FER) tasks. Figure 4 presents the detailed schematic architecture of the proposed method. It includes the selection of pre-trained models trained on the ImageNet dataset, modification of their final layers, preprocessing of the CK+ dataset, and an initial training phase to identify the best-performing models. These selected models are then fine-tuned to produce the final versions, which are subsequently evaluated during the validation step.

3.3.1. *Architectural Modifications.* To adapt deep CNN models for the FER task, we replaced the final classification layer of each architecture with a new fully connected layer outputting 7 classes, corresponding to the seven universal facial expressions: anger, disgust, fear, happiness, neutral, sadness, and surprise. All pretrained layers are retained to exploit the general visual features learned on the ImageNet dataset, while the new classification layer was trained on the CK+ datasets.

3.3.2. *Parameter Count and Model Size.* To further contextualize the trade-off between performance and efficiency, we computed the number of trainable parameters for each model.

Table 1 summarizes the parameter counts, which range from lightweight architectures such as MobileNetV2 (2.2M parameters) to heavier models like VGG19 (exceeding 143M parameters). This comparison helps to evaluate how model complexity impacts training time, memory consumption, and classification accuracy in FER tasks.

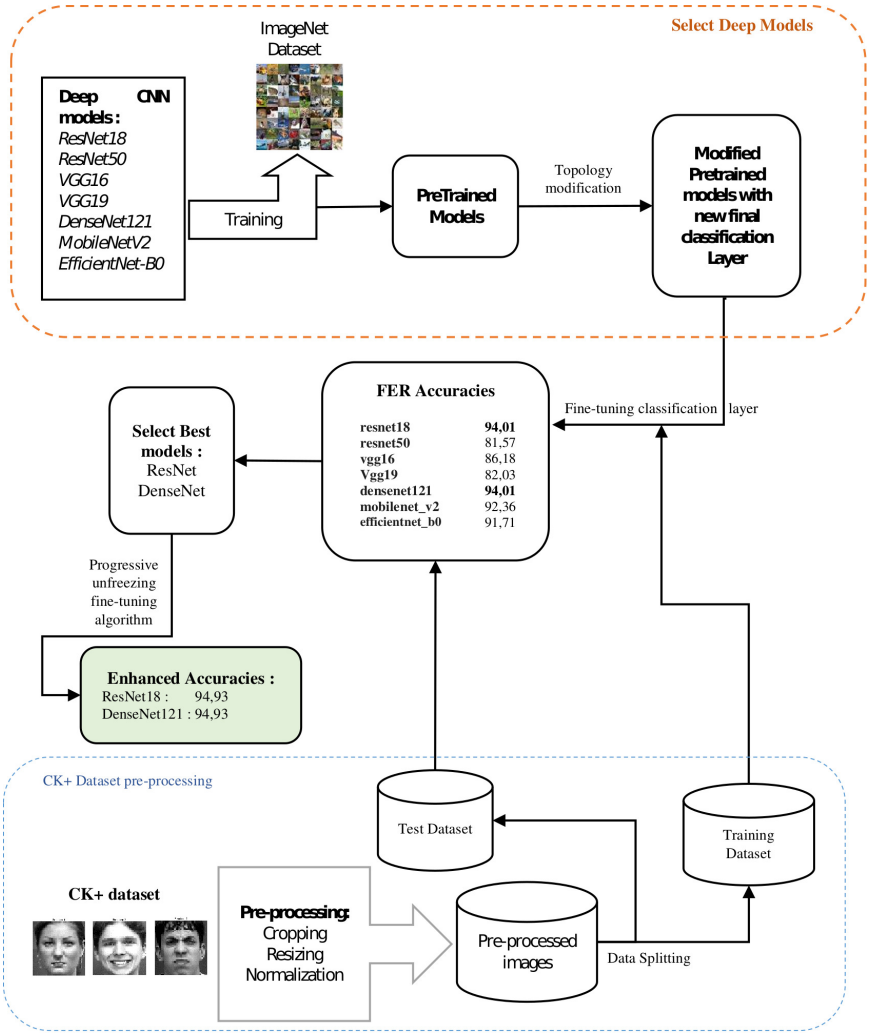

FIGURE 4. General scheme of the proposed comparative facial expression recognition (FER) system based on transfer learning using several deep CNN models.

TABLE 1. Model Parameters Comparison

Model	Parameters
resnet18	11,180,103
resnet50	23,522,375
vgg16	134,289,223
Vgg19	143,667,719
densenet121	6,961,031
mobilenet_v2	2,232,839
efficientnet_b0	4,016,515

3.4. **Training Configuration.** Fine-tuning is a crucial step in transfer learning, and a carefully designed strategy is employed to optimize the performance of the selected models. In the first stage, the newly added layers are fine-tuned, as their weights are randomly initialized and require learning from the target dataset. After this initial training phase, the best-performing models are identified based on their test accuracies. In the second stage, a progressive fine-tuning strategy is applied to these selected models. Specifically, portions of the pre-trained convolutional base are gradually unfrozen and fine-tuned in a step-by-step manner. Figure 5 illustrate the general scheme of progressive fine-tuning

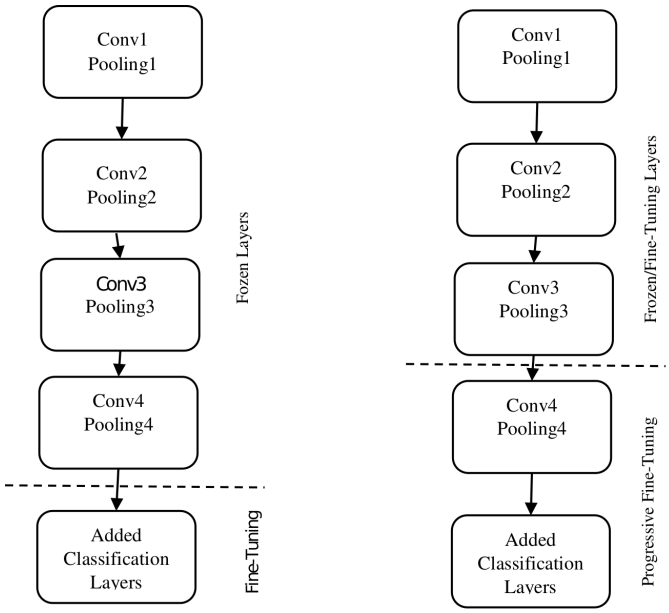

FIGURE 5. Progressive fine-tuning

If the added layers and the pre-trained layers are trained simultaneously from the beginning, the randomly initialized weights of the added layers can produce unstable gradients. These poor gradients may propagate into the already well-trained parts of the network, potentially degrading overall performance. To avoid this issue, a pipeline fine-tuning strategy is employed. In this approach, the training begins with the added layers only. Once these layers are stabilized, selected blocks of the pre-trained CNN are progressively unfrozen and fine-tuned in a step-by-step manner. This gradual adaptation allows for more stable learning and leads to better accuracy.

We have employed the Adam optimization algorithm [25] during the progressive fine-tuning phase. Adam is a widely used optimizer that combines the strengths of two earlier methods: AdaGrad [26], which adapts the learning rate for each parameter based on accumulated past gradients, and RMSProp [27], which adjusts learning rates using an exponentially weighted moving average of squared gradients. Adam maintains two momentum terms, governed by the hyperparameters β_1 and β_2, which estimate the first and second moments of the gradients, respectively. These parameters, together with the learning rate, play a critical role in controlling the convergence behavior and overall performance of the training process.

The learning rate was set to 0.0005, based on preliminary experiments that balanced convergence speed and stability. The Cross-Entropy Loss function was employed as the objective function, as it is suitable for multi-class classification problems like facial expression recognition (FER), where the task involves predicting one of seven mutually exclusive emotion categories.

4. **Results and discussion.** Table 2 presents the test set accuracies for the TL deep models and time consumed in 20 epochs of training:

TABLE 2. Obtained results of TL Deep Models

Model	**Accuracy (%)**	**Time (Minutes)**	**Error**
resnet18	94.01	1.2	0.001
resnet50	81.57	2.4	0.023
vgg16	86.18	3.5	0.074
Vgg19	82.03	4.0	0.281
densenet121	94.01	2.5	0.002
mobilenet_v2	92.36	1.3	0.008
efficientnet_b0	91.71	1.5	0.024

It is noteworthy from Table 2 that ResNet18 and DenseNet121 achieve the highest test accuracy (more than 94%) with the minimum error reached in training phase. This performance indicates its superior

ability to generalize across different facial expression datasets. Vgg16 and Vgg19 are too heavy for this task (3.5 and 4 mins for less than 86% accuracy). Furthermore, Figure 6 and Figure 7 illustrates the evolution of the Loss functions and the accuracies reached during the training phase for all models in the first step. As observed, Vgg16 and Vgg19 are the worst models in termes of accuracies, DenseNet121 consistently reaches a lower training error compared to the other models, further confirming its robustness and effectiveness for facial expression recognition

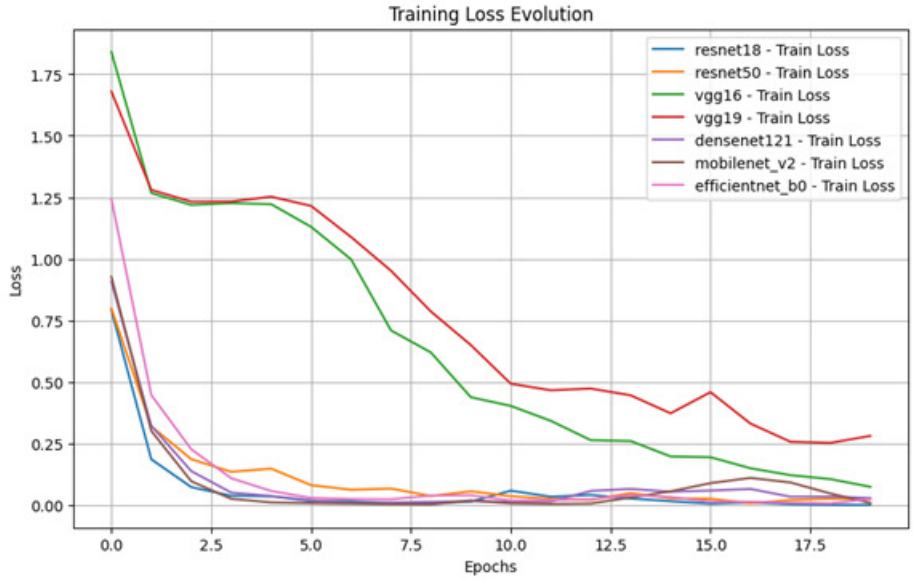

FIGURE 6. Loss functions evolution in the training step

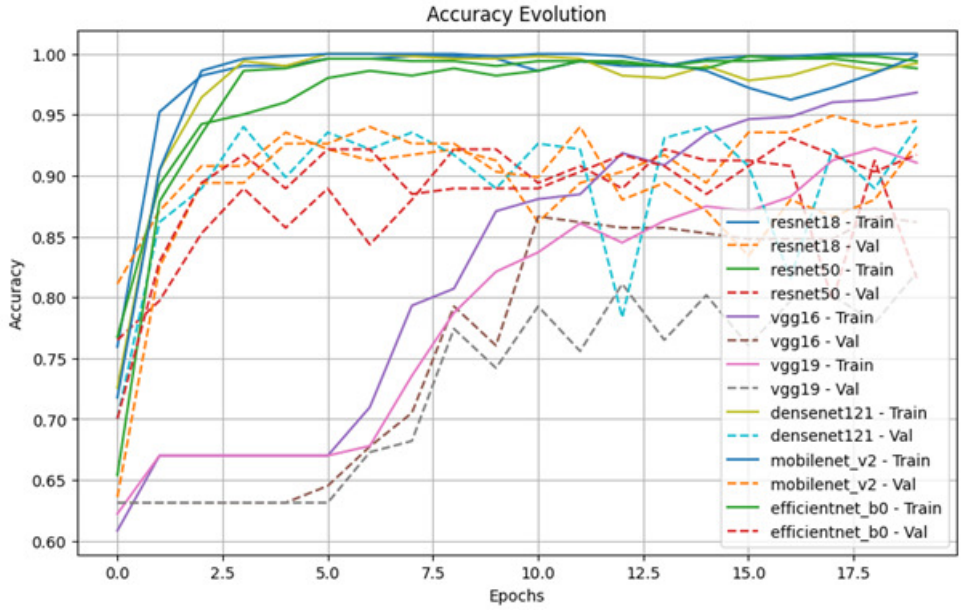

FIGURE 7. Accuracies evolution in the training phase

Based on these results, the best-performing models selected for the second stage of progressive fine-tuning were ResNet18 and DenseNet121. The test accuracy of both models improved equally, reaching 94.93Table 3 presents a comparative summary of several studies on facial expression recognition (FER) using the CK+ dataset.

The study by Wu et al. [28] employs a method termed the Generic Model (GM), which incorporates a modified CNN model with spatial normalization for classification. For feature extraction, the authors

TABLE 3. Comparison method applied on CK+ dataset

Author	**Method**	**Accuracy (%)**
Wu & Lin [28]	GooGleNet	85.71
	AlexNet	85.87
	AlexNet + SVM	86.83
	GM	86.83
	GM + AFM	87.78
	GM + W-AFM	88.25
	GM + W-CR-AFM	89.25
Mollahosseini et al. [11]	STM + SVM	91.13
	STM-ExpLet + SVM	94.19
Sawardekar & Naik [29]	LBP with CNN	90.00
Sanin et al. [30]	Cov3D + LogitBoost	92.30
Ouellet et al. [31]	DCNN + SVM	94.40
Makhmudkhujaev et al. [32]	LPDP	94.50
Bashar et al. [33]	MTP + SVM	94.93
Turan et al. [34]	LGBPHS + SLPM-NN	92.23
	LPQ + SLPM-NN	94.61
	LGBPHS + SVM	91.91
	LPQ + SVM	94.93
Kim et al. [35]	LBP + deep neural network	96.46
Zeng et al. [36]	Deep learning + handcrafted feature	97.35
Debnath et al. [37]	LBP + ORB + CNN model	98.13
Our study	**TL-DCNN**	**94,93**

utilize Adaptive Feature Mapping (AFM), along with its weighted variant (W-AFM), and Weighted Center Regression AFM (W-CR-AFM). They evaluate their approach by comparing its performance against two well-established CNN architectures, AlexNet and GoogleNet. Their experimental results demonstrate that the GM combined with W-CR-AFM achieves superior performance, attaining an accuracy of 89.25%.

Mollahosseini et al. [11] proposed a dynamic facial expression recognition framework based on spatio-temporal manifold (STM) modeling and expressionlet learning (STM-ExpLet). Their approach captures the temporal evolution of expressions by decomposing data into localized spatio-temporal modes, achieving robust alignment and discriminative representation. The method demonstrated strong performance, reporting accuracies of 91.13% for STM and 94.19% for STM-ExpLet.

Sanin et al. [30] introduced spatio-temporal covariance descriptors (Cov3D) for feature extraction, coupled with LogitBoost classifiers, achieving 92.30% accuracy.

Ouellet et al. [31] employed a Deep Convolutional Neural Network (DCNN) model for feature extraction, followed by classification using a Support Vector Machine (SVM). Their DCNN architecture comprises seven layers (including a logistic regression layer), with five convolutional layers and two fully-connected layers. Features of interest were extracted from the fifth layer (prior to any fully-connected transformation) and the sixth layer. These extracted features were then fed into an SVM for classification.

Makhmudkhujaev et al. [32] proposed the Local Prominent Directional Pattern (LPDP), a novel descriptor that extracts statistical neighborhood information to generate robust facial features. Their method achieved state-of-the-art performance with 94.5% recognition accuracy.

In [33], the authors propose an effective appearance-based facial feature descriptor called the Median Ternary Pattern (MTP), a novel local texture operator for facial expression recognition (FER). By thresholding pixel intensities against the local median grayscale value and quantizing the neighborhood into three levels, MTP robustly encodes texture information. When combined with an SVM classifier, this approach achieves 94.93% recognition accuracy.

Turan et al. [34] employ two feature extraction methods—Local Gabor Binary Pattern Histogram Sequence (LGBPHS) and Local Phase Quantization (LPQ)—for facial representation. They evaluate classification performance using both Support Vector Machine (SVM) and Neural Network (NN) classifiers, with Subspace Learning Projection Matrix (SLPM) applied for dimensionality reduction in the NN-based approach. The optimal configuration achieves 94.93% accuracy using LPQ with SVM.

Kim et al. [35] proposed an efficient facial expression recognition algorithm that achieved 96.46% accuracy by combining appearance and geometric features using deep neural networks. Their dual-network approach integrates static appearance features, extracted from LBP-based Action Unit (AU) information, with dynamic geometric features derived from facial landmark movements between neutral and peak expressions. This feature fusion results in a more robust representation for accurate facial expression analysis.

In [36], the authors proposed a general framework for embedding hand-crafted features into a deep network for improved feature learning. This approach uses deep metric learning to guide the deep network with hand-crafted features, enabling the learning of more discriminative representations. The learned deep features are then fused with the hand-crafted features through a fusion network for final recognition. This method achieved an accuracy of 97.35%.

The study in [37] proposed a model called ConvNet, which combines features extracted from Local Binary Pattern (LBP), region-based Oriented FAST and Rotated BRIEF (ORB), and a Convolutional Neural Network (CNN) for facial expression recognition. These features were fused and used as input to train the ConvNet model. By leveraging this fusion-based approach, the method achieved a recognition accuracy of 98.13%.

In summary, recent advancements in facial expression recognition (FER) have demonstrated the effectiveness of combining handcrafted features with deep learning approaches, as seen in table 3 by using methods like GM with W-CR-AFM [28], STM-ExpLet [11], Cov3D [30], and MTP [33], achieving accuracies ranging from 89.25% to 94.93%. More recent works have pushed performance further by leveraging deep networks and feature fusion strategies, with Kim et al. [35], [36], and [37] reporting accuracies of 96.46%, 97.35%, and 98.13%, respectively. Within this landscape, our work explores a transfer learning (TL)-based approach using deep convolutional neural network models trained by a progressive fine tune algorithm (the best accuracies was for ResNet18 and DenseNet121). Our study achieves a competitive accuracy of 94.93%, demonstrating the potential of TL-based architectures in delivering high performance with less reliance on complex feature engineering.

5. **Conclusion.** While numerous studies have explored deep learning models for FER, direct and comprehensive comparisons of transfer learning approaches under standardized conditions remain limited. Most existing works either focus on a single model architecture, use different training protocols, or evaluate on multiple datasets without isolating the impact of model choice. To address this gap, our study systematically benchmarks seven diverse and widely-used CNN architectures under identical training, augmentation, and evaluation settings on the CK+ dataset. This controlled comparison not only identifies the most effective base models for FER but also establishes a reproducible baseline for future research. Our investigation demonstrates that among these models, ResNet18 and DenseNet121 achieve superior performance, reaching 94.93Beyond methodological contributions, this work has practical implications for real-world applications. In human-computer interaction, the identified efficient models enable more responsive affective systems. Furthermore, in healthcare domains particularly mental health assessment, reliable FER systems can assist clinicians in monitoring emotional states and detecting early signs of psychological conditions, demonstrating the translational value of robust model benchmarking.

Acknowledgments. This work is supported by the Directorate General of Scientific Research and Technological Development (DGSRTD, URL: www.dgrsdt.dz, Algeria).

REFERENCES

[1] G. Fenfei and A. Mideth, Research on Multi-Task Learning Facial Attribute Recognition Model Based on Adversarial Training and Differential Privacy, *Journal of Information Hiding and Multimedia Signal Processing* , vol. 16, no. 2, pp. 2073-4212, 2025.

[2] D. G. R. Kola and S. K. Samayamantula, A novel approach for facial expression recognition using local binary pattern with adaptive window, *Multimedia Tools and Applications*, vol.80, no.2, pp.2243–2262, 2021.

[3] H. Jo and B. Kwon, Facial emotion recognition using Canny edge detection operator and histogram of oriented gradients, *Journal of Multimedia Information System*, vol.12, no.1, pp.1–12, 2025.

[4] A. Boughida, M. N. Kouahla, and Y. Lafifi, A novel approach for facial expression recognition based on Gabor filters and genetic algorithm, *Evolving Systems*, vol.13, pp.331–345, 2022.

[5] H. I. Dino and M. B. Abdulrazzaq, Facial expression classification based on SVM, KNN and MLP classifiers, *Proc. of the 2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp.70–75, 2019.

[6] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, *Neural computation*, vol.29, no.9, pp.2352–2449, 2017.
[7] J. Pordoy, H. Farman, N. K. Dicheva, A. Anwar, M. M. Nasralla, N. Khilji, and I. U. Rehman, Multi-frame transfer learning framework for facial emotion recognition in e-learning contexts, *IEEE Access*, vol.12, pp.151360–151381, 2024.
[8] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, Development and application of emotion recognition technology—a systematic literature review, *BMC psychology*, vol.12, no.1, p.95, 2024.
[9] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *nature*, vol.521, no.7553, pp.436–444, 2015.
[10] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, Facial emotion recognition using deep convolutional neural network, *Proc. of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp.317–320, 2020.
[11] A. Mollahosseini, D. Chan, and M. H. Mahoor, Going deeper in facial expression recognition using deep neural networks, *Proc. of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1–10, 2016.
[12] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, Stacked deep convolutional auto-encoders for emotion recognition from facial expressions, *Proc. of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp.1586–1593, 2017.
[13] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, Hybrid deep neural networks for face emotion recognition, *Pattern Recognit. Lett.*, vol.115, pp.101–106, 2018.
[14] D. Y. Liliana, Emotion recognition from facial expression using deep convolutional neural network, *J. Phys. Conf. Ser.*, vol.1193, p.012004, 2019.
[15] M. Sahu and R. Dash, A Survey on Deep Learning: Convolution Neural Network (CNN), in *Smart Innovation, Systems and Technologies*, vol.153, pp.317–325, Springer, Singapore, 2021.
[16] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. S. Awwal, and V. K. Asari, A State-of-the-Art Survey on Deep Learning Theory and Architectures, *Electronics*, vol.8, p.292, 2019.
[17] A. Mahendran and A. Vedaldi, Visualizing Deep Convolutional Neural Networks Using Natural Pre-images, *Int. J. Comput. Vis.*, vol.120, pp.233–255, 2016.
[18] G. Antonellis, A. G. Gavras, M. Panagiotou, B. L. Kutter, G. Guerrini, A. C. Sander, and P. J. Fox, Shake Table Test of Large-Scale Bridge Columns Supported on Rocking Shallow Foundations, *J. Geotech. Geoenviron. Eng.*, vol.141, p.04015009, 2015.
[19] M. D. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, *Proc. of the European Conference on Computer Vision (ECCV 2014)*, pp.818–833, 2014.
[20] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, 2014.
[21] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778, 2016.
[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely Connected Convolutional Networks, *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2261–2269, 2017.
[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, *arXiv preprint arXiv:1704.04861*, 2017.
[24] P. Utami, R. Hartanto, and I. Soesanti, The EfficientNet performance for facial expressions recognition, *Proc. of the 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp.756–762, 2022.
[25] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv:1412.6980*, 2014.
[26] P. L. Bartlett, E. Hazan, and A. Rakhlin, Adaptive Online Gradient Descent, *Adv. Neural Inf. Process. Syst.*, vol.20, pp.1–8, 2007.
[27] T. Tieleman, G. E. Hinton, N. Srivastava, and K. Swersky, RMSProp: Divide the gradient by a running average of its recent magnitude, *COURSERA Neural Netw. Mach. Learn.*, vol.4, pp.26–31, 2012.
[28] B. F. Wu and C. H. Lin, Adaptive feature mapping for customizing deep learning based facial expression recognition model, *IEEE Access*, vol.6, pp.12451–12461, 2018.
[29] S. Sawardekar and S. R. Naik, Facial expression recognition using efficient LBP and CNN, *Int Res J Eng Technol (IRJET)*, vol.5, no.6, pp.2273–2277, 2018.

[30] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, Spatio-temporal covariance descriptors for action and gesture recognition, *Proc. of the 2013 IEEE Workshop on applications of Computer Vision (WACV)*, pp.103–110, 2013.

[31] S. Ouellet, Real-time emotion recognition for gaming using deep convolutional network features, *arXiv preprint arXiv:1408.3750*, 2014.

[32] F. Makhmudkhujaev, M. Abdullah-Al-Wadud, M. T. B. Iqbal, B. Ryu, and O. Chae, Facial expression recognition with local prominent directional pattern, *Signal Processing: Image Communication*, vol.74, pp.1–12, 2019.

[33] F. Bashar, A. Khan, F. Ahmed, and M. H. Kabir, Robust facial expression recognition based on median ternary pattern (MTP), *Proc. of the 2013 International Conference on Electrical Information and Communication Technology (EICT)*, pp.1–5, 2014.

[34] C. Turan and K.-M. Lam, Histogram-based local descriptors for facial expression recognition (fer): A comprehensive study, *Journal of visual communication and image representation*, vol.55, pp.331–341, 2018.

[35] J. H. Kim, B. G. Kim, P. P. Roy, and D. M. Jeong, Efficient facial expression recognition algorithm based on hierarchical deep neural network structure, *IEEE Access*, vol.7, pp.41273–41285, 2019.

[36] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, Hand-crafted feature guided deep learning for facial expression recognition, *Proc. of the 2018 13th IEEE international conference on automatic face & gesture recognition (fg 2018)*, pp.423–430, 2018.

[37] 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, p.6991, 2022.