IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Facial Emotion Recognition (FER) through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets

MUSTAFA CAN GURSESLI[1], SARA LOMBARDI[1], MIRKO DURADONI[2], LEONARDO BOCCHI[1], ANDREA GUAZZINI*[2,3], and ANTONIO LANATA*[1]

[1]Department of Information Engineering, University of Florence, Via di Santa Marta 3, 50139 Florence, Italy
[2]Department of Education, Literatures, Intercultural Studies, Languages and Psychology, University of Florence, 50135 Florence, Italy
[3]Centre for the Study of Complex Dynamics, University of Florence, 50019 Sesto Fiorentino, Italy

Corresponding author: Antonio Lanatà (e-mail: antonio.lanata@unifi.it).

*:These authors contributed equally.

ABSTRACT Facial emotion recognition is a crucial process for many applications and is still unresolved. Historically, emotional recognition has usually been achieved through artificial intelligence techniques such as Convolutional Neural Networks. However, this approach is quite expensive in terms of computational power and complexity. To alleviate this problem, we propose a lightweight CNN for facial emotion recognition, called Custom Lightweight CNN-based Model (CLCM), based on the well-known MobileNetV2 architecture. Performance was evaluated on four public datasets, FER-2013, RAF-DB, AffectNet, and CK+, where seven facial emotions were detected. The CLCM model was compared to the well-known MobileNetV2 and ShuffleNetV2 architectures. The CLCM performance results were close to or better than the more complex models. Specifically, in the FER-2013 dataset, CLCM achieved an accuracy of 63%, while MobileNetV2 was 58%, and ShuffleNetV2 65%. In the RAF-DB dataset, CLCM showed 84%, MobileNetV2 73%, and ShuffleNetV2 80%. Lastly, in the AffectNet dataset, MobileNetV2 and ShuffleNetV2 showed an accuracy of 57%, while CLCM showed 54%. The results obtained from this study establish CLCM as one of the efficient models in FER. Although CLCM is a smaller model in terms of parameters (2.3 Million) compared to MobileNetV2 (3.5 Million) and ShuffleNetV2 (3.9 Million), it showed good results in almost all analyses. The reduced CLCM's computational power allows its application in enhanced human-computer interaction, affective computing, and personalized user experience, especially for real-world scenarios such as psychological and medical assessment, automotive (real-time driver emotional state), and vulnerable individual care where limited resource systems are commonly employed and real-time and reliable response are strongly recommended.

INDEX TERMS Facial emotion recognition, Convolutional neural networks, Computer vision, Emotion recognition, Deep Learning.

I. INTRODUCTION

EMOTIONS trigger complex psycho-physiological changes in one's state of mind, deriving from biochemical and environmental interactions [1]. Emotions are the primary determining factors of a healthy sense of oneself and play a pivotal role in a person's daily life [2]. Moreover, emotions directly connected to one's psychology are relayed through the choices of words, thoughts, phrases, mimics, posture, and especially facial expressions [3].

The human body's internal manifestation of emotional response can be seen in an alteration of the Nervous System response, both Central (CNS) and Autonomic (ANS) [9]. Generally, ANS response can be detected by measuring several physiological signs such as Heart Rate Variability (HRV), Electrodermal Activity (EDA), and Respiration. In contrast, CNS response is detected through the analysis of ElectroEncephaloGram (EEG) [10]. Many instrumentation and processing methodologies have been developed to detect ANS variables in this context. However, most tools must be placed on the subject's body to have physical contact, while others are still very intrusive, altering the ANS response itself. However, many studies have been conducted on them over the

TABLE 1. Public Facial Expression Image Databases Used in the Article

Database	Environment	Images	Position	Colour	Emotion Classes	Annotation Methods
AffectNet-7 [4]	Web	1.000.000	Posed & Spontaneous	RGB & BGR	6 basic emotions + Neutral	One annotator per image
RAF-DB [5], [6]	Web	29.673	Posed & Spontaneous	RGB & BGR	6 basic emotions + Neutral and 12 compound emotions	40 Independent annotators
FER-2013 [7]	Web	32.298	Spontaneous	BGR	6 basic emotions + Neutral	Image search API
CK+ [8]	Lab	593	Posed & Spontaneous	RGB & BGR	6 basic expressions + Neutral + Contempt	FACS coded by two annotators

years [11]. On the contrary, this study aims to identify emotional responses using information acquired with systems that avoid physical contact with the subject's body. Specifically, it is focused on cameras for acquiring and monitoring facial expressions.

Humankind's capability of performing and interpreting facial expressions evolved in human phylogenesis. Facial expressions are fundamental in conveying emotions and constructing communication between individuals. According to the famous Ekman and Freis theory, facial expressions are classified into six core categories. These are anger, disgust, fear, happiness, sadness, and surprise [12]. However, beyond these definitions, facial expressions are also affected by variables such as culture, race, language, and religion [13].

In the last decades, the increase in computational power and technological integration into human life enabled a rapid expansion of computer vision jointly with machine learning models in several research fields [14], [15]. Moreover, computer vision algorithms are not limited to designing dimensional information of separate image frames but also interpreting transitory contextual correlations of consecutive frames [16]. In this context, facial expression recognition (FER) by means of computer vision is still challenging due to the computational load and the high number of variables involved in the emotional facial expression process (culture, race, etc.).

Various studies put much effort into defining and interpreting facial emotion indicators [17], [18]. One of the earliest is the Facial Action Coding System (FACS), which explains facial muscle movements by creating Action Units whose groups, made by muscle synergies, have been scientifically proven to interpret emotions efficiently [12], [19]. Ekman and Friesen's psychology-based study continues to be used today with various computer vision techniques. The foundations of these techniques are mainly based on machine learning approaches (such as Gabor Filters, Support Vector Machine, Random Forest, Local Binary Pattern, Nearest Neighbor Algorithm, Histogram of Oriented Gradients, Neural Networks, etc.) [20].

In this context, the evolution made by intelligent methods such as the Convolutional Neural Network (CNN) technique, which is based on a machine learning approach, put it as the most frequently chosen method by scientists for image processing purposes [21]. Currently, this method is regarded as the most effective for facial expression recognition, outperforming other techniques in accuracy. [21].

Furthermore, there are several successful state-of-the-art CNN and Deep Convolutional Neural Networks (DCNN) models for Emotion recognition in the literature, such as VGG16 [22], ResNet50 [23], and Inception v4 [24]. Although these models provide a lot of innovation to the field, they need high computational power due to their complexity and the high number of required parameters [25]. Moreover, due to the wide proliferation of mobile devices in daily life (e.g., home monitoring healthcare [26], natural psychological evaluations [27], education [28]), a crucial challenge arises, for low power consumption and computing, in implementing lightweight and performance-oriented CNNs [29]. In this context, these models gained an essential role in using mobile devices in real-life scenarios. Therefore, the research aims to decrease the number of parameters for more reliable and effective models.

In this direction, in 2017, the creation of lightweight CNN architectures such as MobileNetV1 [30] and ShuffleNet [31], as pioneers in this field, demonstrated that the performance of large models can also be achieved with lightweight architectures [32], [33]. Lightweight architectures, which reached a larger community in the following years with the publication of MobileNetV2 [32], are gaining great attention due to their wide range of applications and compatibility with mobile devices. However, there is still a significant gap in real-life scenarios literature about lightweight CNN architecture for emotion recognition.

This study presents the Custom Lightweight CNN-based Model (CLCM), an architecture designed to address the crucial need for a lightweight facial emotion recognition model. CLCM, based on the MobileNetV2 architecture and using a transfer learning approach [34] might provide a suitable model for limited computational capability devices. The model implements a lightweight structure that provides better performance than existing models. The compact form of the model emphasizes its potential adaptability for implementation on a web-based platform, highlighting its versatility in adapting to different computing environments [35]. All these features make CLCM a suitable model for many real-life scenarios.

In this study, we aimed to propose an improved lightweight CNN for the facial emotion recognition task. Specifically, the objectives of this study are the following:

1. Designing and implementing a CLCM architecture whose performance will be tested on four validated and public datasets.
2. Compare the CLCM model's performance on each dataset with the popular architectures: MobileNetV2

[32] and ShuffleNetV2 [36].

In line with our aims, we brought significant contributions to the existing literature. The model, developed through a transfer learning approach, is compact and holds potential for use in various domains. Its lightweight characteristics make our model highly suitable for use in reduced computational systems such as mobile devices and online computing platforms. Furthermore, this model can face real-time requirement offering a practical application of FER in real-life emotion studies and clinical settings. In addition, it provides the compactness and flexibility required in applied research, such as the biofeedback approach. Lastly, the CLCM model is suitable for scenarios where several models are used simultaneously, which are becoming very popular nowadays. For example, in scenarios where FER and other systems, such as eye-tracking models, can be used in parallel, compactness and low power consumption are strongly required.

The manuscript is organized as follows: Section II shows various related studies on facial expression image datasets, deep learning, and convolutional neural networks. Section III introduces the technical background, model architectures, and how datasets are used for the proposed model. Section IV describes training and testing procedures, preprocessing, and data augmentation. Section V shows how the three lightweight CNNs perform on four different datasets. Section VI focuses on the results performance comparison of the proposed model and existing models and discusses the performance on emotional categorization. Section VII highlights conclusive remarks. Finally, section VII shows the study's conclusions.

II. RELATED WORKS

This section overviews the literature about three crucial dimensions for facial expression recognition tasks: facial expressions image datasets, deep learning approach, and CNN architecture.

A. FACIAL EXPRESSION IMAGE DATASETS

Public datasets are scarce, and they are one of the main variables that largely determine the performance of AI models. We identified four different datasets in the literature with different features. These differences are the annotation method, the number of images, resolution, gender, race, age, etc. Table 1 shows the four databases used in our experiments and their characteristics.

The selected public datasets are among the FER datasets with the most images available in the literature. Moreover, these datasets used for training were specifically selected from datasets not prepared in laboratory conditions. Previous studies revealed that most datasets in the literature had been produced under strict directives in laboratory conditions. This results in the samples being uniform and repetitive, lacking a variety of conditions that each sample repeats. [4], [6], [7], [37], [38]. Furthermore, the most common limitation is the number of images and the imbalance of sample numbers between classes [37]. The selected datasets are Affectnet, RAF-DB, FER-2013, and Cohn-Kanade Dataset (CK+). The next subsections describe the datasets in detail. All datasets used in this study are available on creators' websites. However, most of them are subject to specific permissions from their original authors [4]–[8].

1) AffectNet

The AffectNet [4] is a large-scale database of facial expressions from internet images. AffectNet-7 has seven classes: anger, disgust, fear, happy, neutral, sad, and surprise (6 basic + 1 neutral), and includes manually annotated 287401 images. AffectNet was originally split into training and validation samples. The authors suggested researchers use the validation dataset as a test set. Besides that, the remaining images (around 550000) are automatically annotated [4]. The images are not cropped (every image has different dimensions) and have Grayscale or RGB color pixels. Examples of AffectNet are shown in Figure 1-a.

2) RAF-DB

Real-world Affective Faces Database (RAF-DB) is a large-scale facial expression database [5], [6]. The dataset is based on crowdsourcing annotation (i.e., the image annotation made by groups of people different from the authors), performed independently by about 40 annotators. Moreover, the dataset has 29673 images from the Internet, divided into seven main emotion expressions(6 basic and neutral, 15500 images) and 12 compound emotion expressions. The images are 100x100 pixels with Grayscale and RGB colors. Examples of RAF-DB are shown in Figure 1-b.

3) FER-2013

Facial Expression Recognition 2013 (FER-2013) is one of the first datasets in the literature, created by Kaggle in 2013. The dataset was divided into seven emotion expressions (6 basic and neutral) with 28709 train images and 3589 test images. These images were collected from the Internet, and an algorithm automatically annotated emotion labels [7]. The images are 48 x 48 pixels with Grayscale colors. Examples of FER-2013 are shown in Figure 1-c.

4) The Extended Cohn-Kanade Dataset (CK+)

Extended Cohn-Kanade (CK+) is one of the most popular and first laboratory-controlled facial expression classification databases. The dataset, created with the participation of 123 people, consisted of 593 video recordings (30 Frames per second) of people expressing seven basic emotions. Each video recorded the participant's emotions changing from neutral to the intended emotion. When the participant performs a peaked emotional expression, this frame is taken as a dataset image. As a result of these procedures, 327 labeled images are categorized. These images are 640x490 or 640x480 pixels, with Grayscale and RGB colors [8], [39]. Examples of CK+ are shown in Figure 1-d.

FIGURE 1. Facial expression images of different datasets (a) AffectNet [4], (b) RAF-DB [5], [6], (c) FER-2013 [7], and (d) CK+ [8]. From left to right are the images labeled with anger, disgust, fear, happy, neutral, sad, and surprise.

B. CONVOLUTIONAL NEURAL NETWORKS AND FACIAL EMOTION RECOGNITION

The techniques of machine learning and deep learning have become an integral part of many of the tools that are commonly used in our lives to solve various problems [38], [40]. One of their main applications is solving well-known pattern recognition problems such as object recognition [41], face recognition [42], and image classification [43]. The Convolutional Neural Network, popular among neural networks, is used for many recognition tasks, including emotion recognition [44], [45]. The most important reason is the effective capacity of CNN to extract discriminating features automatically [46]. CNN's journey started with LeCun's handwriting recognition [47] study in 1989 and accelerated its development with competitions held in the following years. Competitions such as ILSVRC-2012 [48], ICML 2013 [49], and EmotiW 2013 [50] showed how the limits of the CNN technique could be stretched even further.

DCNN architectures, designed to solve different sophisticated problems, contain deeper and much more trainable parameters than CNN, but they need more computing power. CNNs and DCNNs architectures such as AlexNet [48] DenseNet [51], VGG16 [22], and ResNet [23] have a high number of parameters, and they are not suitable for training with small datasets, thus causing under-fitting with the model while generalizing, and tapering the performance. Nevertheless, the latest studies aimed at adapting the existing state-of-the-art models to compact algorithms. MobileNetV1 [30] was developed by the Google brain team to obtain more efficient performance with a more compact structure. It is nearly as accurate (in ImageNet) as VGG16 and is 32 times smaller. Then, an improved model, i.e., MobileNetV2 [32], outperformed MobileNetV1 in ImageNet with fewer parameters. Besides, ShuffleNet [31] and EfficientNet b0 [52] also achieved similar performance with low parameter counts.

In addition to the significant improvements mentioned before, many researchers have explored the application of various computer vision techniques to improve emotion recognition accuracy [53], [54]. One notable example is the study of Zeng et al. [55], which achieved enhanced accuracy in the FER task by combining hand-crafted features with deep network learning. Moreover, Amin et al. [56] aimed to improve accuracy by applying different hyper-parameter techniques on the FER-2013 dataset.

In the literature, numerous CNN models have been developed with different structures and approaches, and these models are actively used today for different purposes. Although many of these models mentioned in this section have complex structures, few studies in the literature have focused on lightweight models in the context of FER. However, the increasing use of mobile devices and the internet has led to a rise in the utilization of lightweight models, creating a need for their development. CLCM focused on filling in the highlighted gap on this topic.

III. METHODOLOGY

In this study, we analyzed accuracy per class to test overall accuracy, characterized as the proportion of correct classifications concerning the total sample set. This included determining the amount of correct classifications attributed to each of the seven different emotional categories. This strategy is one of the most widely used methods in the literature to measure the performance of models [57], [58]. Moreover, machine learning performance evaluation was conducted in two phases. First, we trained, validated, and tested CLCM and the popular MobileNetV2 and ShuffleNetV2 architectures using AffectNet, RAF-DB, and FER-2013 datasets separately. Second, we tested all previously trained models on the fourth dataset, CK+.

A. MODELS ARCHITECTURE

This section presents a detailed overview of the architectures of MobileNetV2, ShuffleNetV2, and the CLCM model. Following a general description of the key aspects of these models, the subsequent discussion focuses on the specific characteristics of the CLCM implementation. The architectures included in this study share common features such as a low number of parameters and limited power requirements. Although each model has structural similarities, they have different designs and features, and their main common attributes are compact models. These features are highly critical for the models to work on mobile devices.

1) MobileNetV2

MobileNetV2 is a convolutional neural network architecture based on an inverted residual structure, in which the building block is a separable bottleneck deep convolution with residual connections [32]. Each block contains three convolutional layers followed by Batch Normalization layers. The first layer (Expansion Layer, EL) is a 1x1 Convolution with Relu6 activation is responsible for expanding the number of channels in the data. The second layer is a Depthwise Convolution with Relu6 activation and the final layer, the Projection Layer (PL), is a 1x1 point-by-point Convolution. PL is responsible for reducing the output dimensionality. A residual connection performs the sum of the input to the bottleneck block with

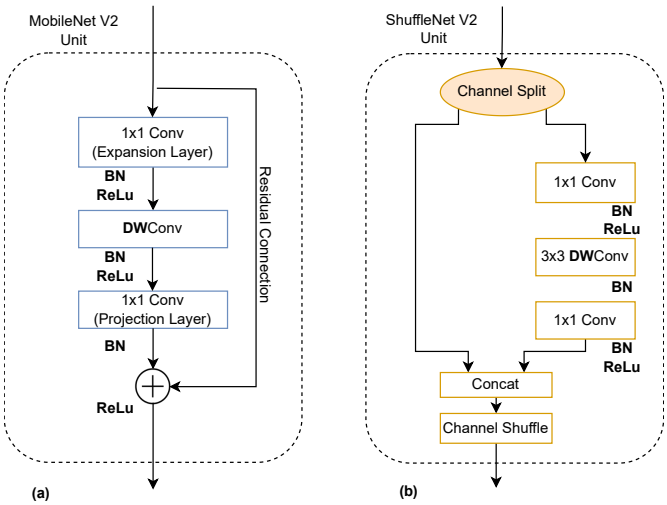

FIGURE 2. **Building blocks of used architectures. (a) shows the building block of MobileNetV2 architecture. (b) represents the building block of ShuffleNet V2 architecture.DWConv: Depth-wise Convolution**

TABLE 2. **MobileNetV2 Architecure**

Input Size	Layer	Stride	# Repetitions
224x224	Conv2D	2	1
112x112	Bottleneck Block	1	1
112x112	Bottleneck Block	2	2
56x56	Bottleneck Block	2	3
28x28	Bottleneck Block	2	4
14x14	Bottleneck Block	1	3
14x14	Bottleneck Block	2	3
7x7	Bottleneck Block	1	1
7x7	Conv2D 1x1	1	1
7x7	Average Pooling		1
1x1	Dense		

the output of the last Batch Normalization layer. Figure 2 (a) shows a graphical representation of the building block.

Globally, the MobileNetV2 architecture contains an initial complete convolutional layer, then 17 building blocks in a row. This is followed by a regular 1×1 convolution, a global average pooling layer, and a classification layer. Table 2 shows the complete architecture of MobileNetV2 [7]. The CLCM architecture's final layer has been adjusted to work with seven classes for emotion recognition paradigm testing.

2) ShuffleNetV2

ShuffleNetv2 is a convolutional model designed for mobile devices. This model builds on the architecture of the popular ShuffleNet and introduces structural changes to increase its speed. Specifically, the ShuffleNet building block is a bottleneck-like structure that uses pointwise group convolutions and a channel shuffle operation to enable information communication between different channel groups. ShuffleNetV2 introduces a "channel split" operator at the beginning of each block, which divides the input into two branches. One of the two branches is used as identity, while the other consists of 3 convolution operations with the same number of input and output channels. After convolutions, the

TABLE 3. **ShuffleNetV2 Architecture**

Input Size	Layer	Stride	# Repetitions
224x224	Conv2D	2	1
112x112	Max Pooling	2	1
56x56	Building Block	2	1
28x28	Building Block	1	3
28x28	Building Block	2	1
14x14	Building Block	1	7
14x14	Building Block	2	1
7x7	Building Block	1	3
7x7	Global Pooling		
1x1	Dense		

two branches are concatenated, and the shuffle operation is applied. The building blocks are stacked to build the overall network structure described in Table 3. The last layer of the architecture was adapted to recognize seven classes. Figure 2 (b) provides a clear picture of the building block.

3) Custom MobileNetV2 (CLCM)

In this study, we adopted a transfer learning approach [34] using the MobileNetV2 architecture and following the same mathematical formulation as the original architecture [32]. Transfer learning is a supervised learning technique that reuses parts of a previously trained model on a new neural network architecture designed for a different task. The assumption is to use a trained model on a sufficiently large and heterogeneous dataset, representing a generic vision model. The feature maps learned from the model can then be exploited without training a complex model architecture from scratch on a large dataset. Two common transfer learning approaches are feature extraction and fine-tuning. The CLCM model applies a fine-tuning approach. The fine-tuning process involves training the deeper layers of the pre-trained network to capture specific dataset features [59]. In this method, the low-level layers of the pre-trained model are typically set as untrainable or "frozen", while the final part of the network, responsible for the original classification task, is usually replaced by a new classifier layer. The final layers of the base model and the new classifier are then trained together to learn the desired task.

In the CLCM model, the first Bottleneck Residual Block has been frozen. In this way, the original weights of the frozen layers are preserved and remain unchanged during the backpropagation phase. The initial layers of a pre-trained network often learn to identify low-level features such as edges and textures [60]. Freezing this block allows the model to retain these low-level features and focus training toward deeper layers, facilitating the acquisition of more complex, task-specific representations. In addition, by avoiding redundant training of the initial layers, there is a decrease in the number of trainable parameters and, consequently, a reduction in the computational load during training.

With regard to the classification layer, we removed the last prediction layer of MobileNetV2, and we added a fully connected neural network to learn specific characteristics of

the dataset. The fully connected network consists of 2 dense layers of 128 and 64 units, respectively, and an output layer (see Figure 3), with several units equal to the categories of emotions to be classified. A ReLu activation layer follows each dense layer and a Dropout layer, with a rate of 0.5, is used as a regularization strategy to prevent model overfitting. The output layer uses the softmax activation function (instead of ReLU) to obtain the probabilities that the input belongs to a particular emotional class. Figure 3 shows the MobileNetV2 core and the custom fully connected layers we use in the model.

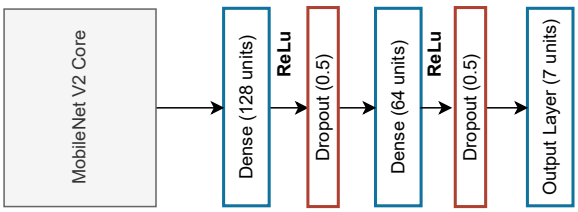

FIGURE 3. Architecture of our CLCM model. The model's core has the same structure as MobileNetV2, while the final classification layer has been replaced by a new fully connected network.

B. DATASETS DESCRIPTION

This study was carried out on different datasets. Images other than the seven expressions commonly adopted in Ekman's theory were excluded from the datasets, and only these seven facial expressions were considered for further analysis. In these datasets, there are images with different sizes, i.e., from 48x48 to 224 x 224 with RGB and BGR pixels.

The Affectnet, RAF-DB, and FER-2013 have been selected for the training phase, and although they have different characteristics, are considered standardized in the literature. More specifically, these datasets were all created in natural environment images. The CK+ dataset was considered only for the testing phase. Previous studies have highlighted that most datasets documented in the literature are prepared from controlled laboratory environments where strict guidelines are enforced. This approach leads to a lack of variety and repetition within samples, thereby reducing the representation of different conditions [4], [6], [7], [37], [38]. Therefore, for the training phase, datasets originating from laboratory environments are deliberately excluded.

Each of these datasets has its own challenges and strengths in this study. AffectNet is the dataset with the highest number of images in the literature, and it is available to all researchers. However, even though the number of images is high, having only one annotator for each image may cause some of the images in the dataset to be mislabeled. On the other hand, RAF-DB is one of the datasets with the highest accuracy rate both in the literature and this study, despite having the lowest number of images among the datasets used for the train. Moreover, although FER-2003 is the first facial emotional recognition dataset created in the literature, it does not perform as well as other datasets in terms of image clarity due to the 48x48 size of the contained images. Finally, the CK+ dataset, which is only used as a test dataset, is the first laboratory-based dataset created in the literature and has a much smaller number of images than the datasets used for the training.

IV. MODEL TRAINING AND TESTING

This section reports on the description of the proposed training and evaluation pipeline (as shown in Figure 4), which includes the data preprocessing, partitioning, data augmentation, model training, and evaluation process.

A. PREPROCESSING AND DATA AUGMENTATION

The data preparation procedure is crucial to ensure the model's good performance. Each image was converted to grayscale and resized to 224x224 pixels. In addition, each sample was normalized to vary between -1 and 1.

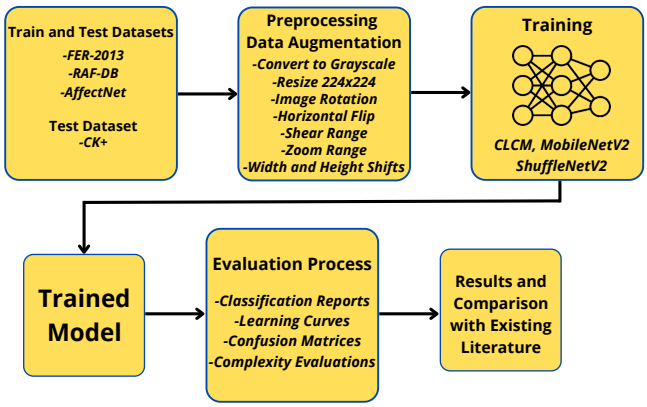

FIGURE 4. Pipeline for the Study

Data augmentation was performed by adding slightly modified copies of existing data to reduce overfitting when training [61]. The preprocessing was carried out without altering the nature of facial expressions. Specifically, it consisted of slight image rotation, horizontal flip, shear and zoom range, and width & height shifts. These techniques are standardized methods used in the literature for the FER task [62], [63]. The use of the horizontal flip technique contributes to increased model robustness and generalization across different spatial orientations and viewpoints [64]. In addition, the rotational transformation of images is used to make models invariant to rotational fluctuations [65]. Furthermore, the use of zooming and shearing techniques is recognized for its effectiveness in increasing the accuracy of facial features and emotional expressions [66].

B. TRAIN, TEST, AND VALIDATION PROCEDURES

The used datasets differ in the number of images. As suggested by the creators of the datasets, for train, testing, and validation processes of AffectNet-7, RAF-DB, and FER-2013 databases, the training sets were kept constant in their original number. The validation set was created by taking 20% of the original train set (80% train, 20% validation), and the test set was kept in its original number. In this way, performance

comparison with previous studies in the literature has been made possible. The validation data were used to select the optimal hyperparameter configuration for the models, as well as to determine the optimal training stop point to obtain the best-performing model. All models were trained separately on each of the 3 databases and tested on the corresponding test set. In addition, each model's performance was evaluated on the popular CK+ dataset, which was used in this work only as a test set to evaluate the model's generalization ability.

Regarding the selection of training hyperparameters, since we had limited computational power available, we adopted a trial-and-error approach [67] using a few of the most commonly used hyperparameters for image classification tasks. We explored values for batch sizes 32, 64, and 128 (limited by our storage capacity), learning rates of 10^{-3} and 10^{-4}, and Adam and Stochastic gradient descent (SGD) optimization functions. The final hyperparameter configuration was chosen based on the overall accuracy of the validation set and the learning curves observed on the training set. Overall accuracy was chosen as the metric for selecting hyperparameters since comparable work in the literature used this metric for model performance evaluation [68]. The selected hyperparameters were a batch size of 64, a learning rate of 0.0001, and the Adam optimizer, as based on our experiments, they produced the best performance for all models. The categorical cross-entropy function was chosen as the loss function since the task is a multi-class problem. The same set of hyperparameters was applied consistently across all training experiments so that the performance of the models and the utilization of different datasets could be compared. All models underwent 100 epochs of training. Best model weights were saved using Keras' ModelCheckpoint callback. The callback was set to store the optimal model based on validation accuracy.

C. MODEL EVALUATION

At the end of the training, each model was evaluated on the data selected for the test set and additionally on the CK+ database. In addition to the overall accuracy, i.e., the percentage of correct classifications out of the total number of samples, we evaluated the accuracy per class, that is, the number of correct classifications for each of the seven emotions.

Regarding the complexity of the investigated models, we analyzed some of the most commonly used metrics to assess the complexity of artificial neural networks: the number of trainable parameters, the number of floating-point operations (FLOPs), and the inference time. FLOPs represent the total number of calculations the model must perform to process an input sample. The total number of floating-point operations was estimated using the TensorFlow Python API. Inference time represents the time it takes for the trained model to process an input and provide the output. To estimate this parameter, we repeated the inference process for 50 input samples and estimated the mean and standard deviation of the prediction times. A low number of FLOPS and inference time is necessary in a real-time emotion recognition scenario.

V. RESULTS

A. ANALYSIS OF THE FER-2013 DATABASE

In the conducted experiments, 32,298 images from the FER-2013 database were utilized for training, testing, and validation purposes. The dataset consists of 7 different emotion categories, each with different image numbers in the train and test sets. Specifically, the train set contains 3,995 images, and the test set includes 958 images.

TABLE 4. Overall Accuracy rate on FER-2013, RAF-DB, AffectNet with original test sets (%)

FER-2013	
Method	Overall Accuracy
MobileNetV2	58
CLCM	63
ShuffleNetV2	65
RAF-DB	
Method	Overall Accuracy
MobileNetV2	73
CLCM	84
ShuffleNetV2	80
AffectNet	
Method	Overall Accuracy
MobileNetV2	57
CLCM	54
ShuffleNetV2	57

The performance of the models in each class can be inferred from the confusion matrix, which summarizes the number of correct and incorrect predictions. Table 8 presents the confusion matrices obtained by the CLCM, MobileNetV2, and ShuffleNetV2 models for the FER-2013 dataset. The diagonal entries represent the accuracy of correctly identified individual expressions. The results indicate that the CLCM model outperformed MobileNetV2 in classifying the categories 'disgust' and 'fear.' In addition, CLCM outperformed ShuffleNetV2 for the emotion categories 'angry,' 'fear,' and 'surprise.' Lastly, ShuffleNetV2 performed better than other models in the 'sad' category. While CLCM performed better in these categories, it achieved comparable results to MobileNetV2 and ShuffleNetV2 in the remaining emotion classes.

In terms of overall accuracy on the FER-2013 test set (shown in Table 4), MobileNetV2 achieved an accuracy of 58%. In contrast, CLCM achieved an accuracy of 63%, and ShuffleNetV2 achieved an accuracy of 65% (see Table 4). Figure 5 shows the training procedure and demonstrates that CLCM experienced fluctuations in the initial 18 epochs while the validation loss accuracies of MobileNetV2 remained consistently close.

CLCM showed a lower training loss than MobileNetV2 when trained on the FER-2013 dataset. It is also observed that the validation loss of ShuffleNetV2 decreases with fluctua-

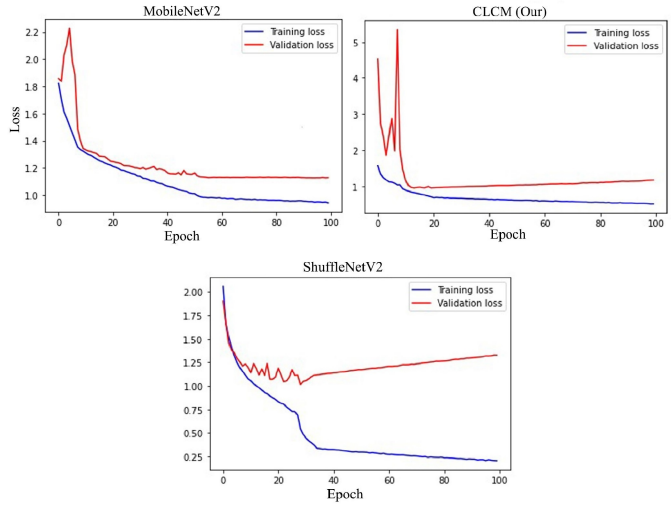

FIGURE 5. Learning Curves (Loss) over the FER-2013

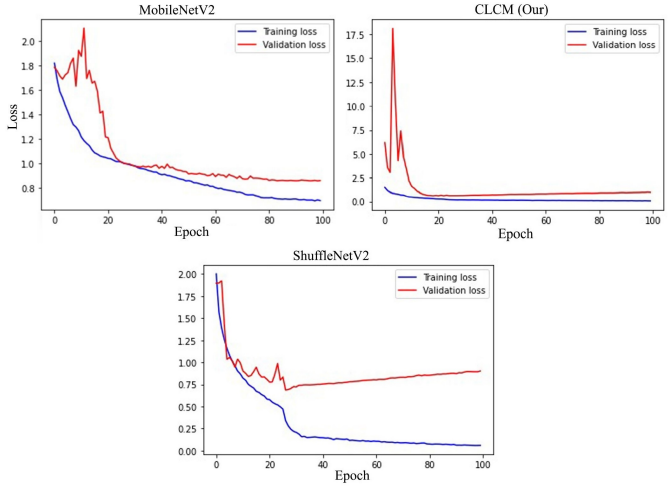

FIGURE 6. Learning Curves (Loss) over the RAF-DB

tions up to 25 epochs, but the validation loss increases in the following epochs.

B. ANALYSIS OF THE RAF-DB DATABASE

All the procedures (testing, training, and validation) were performed with 15339 (Train 12,271 - Test 3,068) images in the RAF-DB database. Twelve compound emotion expressions from RAF-DB have been excluded from the seven main facial emotion recognition.

Results presented in Table 8 for RAF-DB database revealed that the CLCM model scored higher for classifying the categories of 'angry,' 'disgust,' 'fear,' 'neutral,' and 'surprise' compared to MobileNetV2 and ShuffleNetV2. MobileNetV2 performed similar results on the 'sad,' 'neutral,' 'angry,' and 'happy' categories as CLCM for this dataset. Moreover, CLCM performed better than ShuffleNetV2 in discriminating most emotion categories, including 'angry,' 'disgust,' 'fear,' 'happy,' 'neutral,' and 'surprise.'

In terms of overall accuracy, CLCM achieved a performance of 84%, while ShuffleNetV2 achieved 80%, and MobileNetV2 reached 73% of accuracy rates (shown in Table 4). As shown in Figure 6, MobileNetV2 exhibited more fluctuations in the training and validation loss parameters compared to the CLCM model. Additionally, CLCM demonstrated a lower training loss than MobileNetV2 in the RAF-DB dataset. It is worth mentioning that with the increase in the number of epochs during the training period of ShuffleNetV2, the training loss decreased, but the validation loss increased, indicating overfitting of the model.

C. ANALYSIS OF THE AFFECTNET DATABASE

All testing, training, and validation procedures were performed with 287401 (Train 283,901 - Test 3,500) images in the AffectNet database. For the testing for seven major facial emotion recognition, AffectNet-8 has been excluded, and we employed the manual annotated version of AffectNet-7. The original validation set provided was used as a test set recommended by the manuscript [4].

Results presented in Table 8 for AffectNet highlighted that MobileNetV2 and ShuffleNetV2 achieved higher scores in classifying the categories of 'angry,' 'disgust,' 'fear,' and 'surprise.' CLCM performed similarly to MobileNetV2 and ShuffleNetV2 in the 'angry,' 'disgust,' 'fear,' 'surprise,' and 'sad' categories. While ShuffleNetV2 and MobileNetV2 exhibited closely aligned results, CLCM slightly trailed behind them in most categories on the AffectNet dataset.

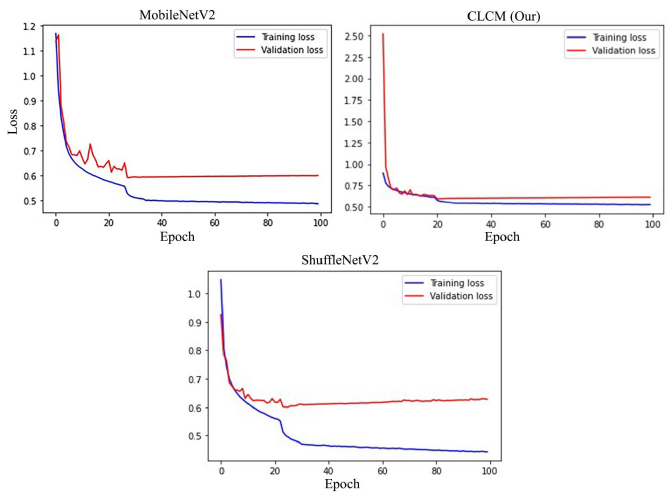

FIGURE 7. Learning Curves (Loss) over the AffectNet

Results of the overall accuracy of test set performance between MobileNetV2 and CLCM are shown in Table 4. The results show that MobileNetV2 and ShuffleNetV2 showed 57% accuracy, and CLCM showed 54% accuracy. MobileNetV2 and ShuffleNetV2 slightly outperformed the CLCM model on the AffectNet dataset. According to the learning curves in Figure 7, MobileNetV2 showed more training and validation

loss fluctuations than the CLCM and ShuffleNetV2 models. Furthermore, MobileNetV2 has a slightly lower training loss than CLCM and ShuffleNetV2.

TABLE 5. Overall accuracy rate on CK+ dataset (%) using models trained on FER-2013, RAF-DB and AffectNet

FER-2013	
Method	Overall Accuracy
MobileNetV2	65
CLCM	71
ShuffleNetV2	71
RAF-DB	
Method	Overall Accuracy
MobileNetV2	47
CLCM	78
ShuffleNetV2	60
AffectNet	
Method	Overall Accuracy
MobileNetV2	91
CLCM	91
ShuffleNetV2	92

D. TEST ANALYSES OF THE TRAINED MODELS WITH CK+

The results obtained by testing the models MobileNetV2, CLCM, and ShuffleNetV2, trained with the AffectNet, FER-2013, and RAF-DB datasets on the CK+ dataset, show considerable similarity with the other findings (as shown in Table 5). In terms of overall accuracy, the CLCM and ShuffleNetV2 models trained on the FER-2013 dataset achieved 71% accuracy, while MobileNetV2 achieved a lower accuracy of 65%. CLCM, trained on the RAF-DB dataset, achieved an accuracy of 78%, followed by ShuffleNetV2 at 60% and MobileNetV2 at 47%. Notably, trained on the AffectNet dataset, ShuffleNetV2 achieved an accuracy of 92%, whereas CLCM and MobileNetV2 showed a slight difference, achieving an accuracy of 91%. It is important to note that MobileNetV2, trained on the RAF-DB dataset, showed limitations in identifying the emotions labeled 'disgust' and 'fear.' Similarly, ShuffleNetV2, trained with the RAF-DB dataset, struggled to identify the emotion category 'fear.' In contrast, CLCM showed mixed results for the emotions categorized as 'happy,' 'sad,' and 'surprise' within the 'fear' emotion category. While the MobileNetV2 and ShuffleNetV2 models could not detect the 'fear' category, CLCM had difficulty distinguishing it from other emotion categories. Furthermore, MobileNetV2, trained on the FER-2013 dataset, showed an inability to identify the 'disgust' class, while CLCM and ShuffleNetV2 had difficulty distinguishing 'disgust' from other categories (Shown in Table 9).

E. ANALYSES OF MODELS FOR GFLOPS AND INFERENCE TIME

The computational effectiveness of models plays a vital role in real-time tasks such as emotion recognition. Therefore, a series of analyses were performed to understand the differences in computational effectiveness between the models. The models' complexity was evaluated based on their trainable parameters, GFlops, and mean inference time. Table 6 shows the evaluation of selected metrics for the MobileNetV2, CLCM, and ShuffleNetV2 architectures. Among the models, CLCM (2,393,191 trainable parameters) performed with the lowest mean inference time of 0.0502 seconds, closely followed by MobileNetV2 (3,511,879 trainable parameters) with a mean inference time of 0.0584 seconds. ShuffleNetV2 (3,997,795 trainable parameters) had a slightly longer mean inference time of 0.0633 seconds.

Further analysis of the differences in inference times was conducted using ANOVA, and the post hoc test revealed statistically significant mean differences between the models (shown in Table 7). Specifically, MobileNetV2 and CLCM showed a significant mean difference of 0.0082 seconds, indicating variability in their inference times. Similarly, a significant mean difference of 0.0048 seconds was observed between ShuffleNetV2 and MobileNetV2. Furthermore, ShuffleNetV2 and CLCM showed a significant mean difference of 0.0130 seconds.

TABLE 6. Complexity Evaluation Results for Models

Model	T.Param	Gflop	Mean Inf. Time [s] (N=50)
MobileNetV2	3,511,879	0.6016	0.0584 ± 0.0035
CLCM	2,393,191	0.5994	0.0502 ± 0.0020
ShuffleNetV2	3,997,795	0.9743	0.0633 ± 0.0062

TABLE 7. ANOVA Post Hoc Test Results for Model Inference Time Comparisons

Models	Comparing Models	Mean Difference
MobileNetV2	CLCM	0,0082*
	ShuffleNetV2	-0,0048*
CLCM	MobileNetV2	-0,0082*
	ShuffleNetV2	-0,0130*
ShuffleNetV2	MobileNetV2	0,0048*
	CLCM	0,0130*

* The mean difference is significant at the 0.05 level.

VI. DISCUSSION

This study implemented model development and training in Python programming using the Keras framework with the Tensorflow library [69] as a backend, within the Microsoft Windows operating system. All machine learning experiments were conducted on a computer with an Intel i7-6700k processor, 32 GB RAM, and a Geforce GTX 1080TI GPU.

Facial emotion recognition is one of the most challenging computer vision tasks. In our study, we propose a lightweight CNN model for facial emotion recognition, and we compare its performance with two popular lightweight architectures (MobileNetV2 and ShuffleNetV2) by reporting the results obtained on four different datasets. All the datasets used for training were created with images collected from the internet (see Table 1). The main reason for choosing these datasets is that data created in the laboratory environment, obtained by executing precise instructions, do not exhibit the range of features that characterize emotions in real-life conditions [6]. In the evaluation phase, to assess the generalization ability of

TABLE 8. Per-Class Performance Analysis: Confusion Matrices on FER-2013, RAF-DB, AffectNet using MobileNetV2, CLCM, and ShuffleNetV2

FER-2013 Dataset (%)

	MobileNetV2							CLCM							ShuffleNetV2						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	40	0	10	2	7	38	3	64	2	13	3	7	10	1	60	1	12	4	9	13	1
Disgust	44	0	6	3	2	42	3	35	44	10	1	1	7	2	23	52	9	6	4	4	2
Fear	21	0	17	3	9	33	17	15	1	50	2	8	14	10	10	0	47	3	9	23	8
Happy	3	0	2	84	5	3	3	4	0	2	82	6	3	3	2	0	1	86	5	4	2
Neutral	2	0	5	6	57	27	3	15	0	8	6	57	12	2	7	1	7	7	59	18	1
Sad	10	0	6	3	18	60	3	16	1	17	2	16	46	2	10	0	14	4	15	56	1
Surprise	3	0	7	3	1	2	84	4	0	11	3	3	1	78	3	1	10	4	3	3	76

RAF-DB Database (%)

	MobileNetV2							CLCM							ShuffleNetV2						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	67	0	0	9	7	13	4	74	12	3	4	4	3	0	68	10	1	7	5	5	4
Disgust	18	0	0	10	28	41	3	9	49	1	9	23	6	3	6	44	1	14	18	13	4
Fear	43	0	0	11	19	20	7	5	1	56	10	4	8	16	5	3	54	11	7	11	9
Happy	0	0	0	94	4	1	1	0	1	0	94	4	1	0	1	1	0	91	4	2	1
Neutral	0	0	1	5	86	7	1	0	2	0	5	87	4	2	1	2	1	6	76	10	3
Sad	1	0	0	7	14	78	0	1	3	2	4	14	75	1	1	3	1	7	12	75	1
Surprise	4	0	0	7	68	7	14	2	2	2	3	6	0	85	2	1	4	4	10	2	77

AffectNet Database (%)

	MobileNetV2							CLCM							ShuffleNetV2						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	57	2	1	4	30	4	2	55	2	1	4	32	4	2	57	2	2	3	30	5	1
Disgust	27	30	2	12	19	9	2	29	24	2	11	20	12	2	27	31	2	10	21	7	2
Fear	7	2	42	5	15	11	18	9	1	35	5	15	14	21	6	1	42	5	16	12	18
Happy	0	0	0	94	6	0	0	0	0	0	95	4	1	0	0	0	0	93	6	0	1
Neutral	4	0	1	10	78	4	3	4	0	0	12	79	3	2	4	0	0	9	80	3	4
Sad	7	1	0	4	30	57	1	8	0	1	5	32	53	1	7	2	0	3	28	57	3
Surprise	3	0	5	20	28	4	40	3	0	4	19	33	5	36	3	1	5	20	27	3	41

TABLE 9. Per-Class Performance Analysis: Confusion Matrices on CK+ dataset

	CLCM Trained with RAF-DB test with CK+							CLCM Trained with AffectNet test with CK+							CLCM Trained with FER2013 test with CK+						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	11	40	0	7	20	22	0	91	4	0	0	2	2	0	42	0	11	0	24	22	0
Disgust	12	78	0	5	3	2	0	8	88	0	2	2	0	0	78	8	3	5	3	2	0
Fear	4	4	20	16	4	40	12	0	4	56	0	0	36	4	0	0	48	4	0	32	16
Happy	0	0	0	100	0	0	0	0	0	0	100	0	0	0	0	0	0	100	0	0	0
Neutral	0	0	1	5	89	5	0	2	0	0	5	91	2	0	0	0	2	0	88	10	0
Sad	0	7	0	0	18	75	0	4	0	0	0	7	89	0	0	0	18	0	11	71	0
Surprise	0	0	2	0	1	0	96	0	0	0	0	1	0	99	1	0	4	1	4	0	90
	MobileNetV2 Trained with RAF-DB test with CK+							**MobileNetV2 Trained with AffectNet test with CK+**							**MobileNetV2 Trained with FER2013 test with CK+**						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	4	0	0	7	22	64	2	87	0	2	0	2	9	0	9	0	0	0	42	49	0
Disgust	71	0	0	5	0	19	5	8	92	0	2	2	0	0	19	0	0	2	5	75	0
Fear	8	0	0	8	40	44	0	0	8	48	0	0	32	12	32	0	8	8	16	16	20
Happy	0	0	0	100	0	0	0	0	0	0	100	0	0	0	0	0	0	100	0	1	0
Neutral	0	0	0	6	89	6	0	2	0	0	4	91	3	0	0	0	0	2	93	4	0
Sad	0	0	0	0	36	64	0	0	0	0	0	7	93	0	4	0	0	0	39	57	0
Surprise	0	0	0	1	95	0	4	0	0	1	0	1	0	98	0	0	6	0	5	1	88
	ShuffleNetV2 Trained with RAF-DB test with CK+							**ShuffleNetV2 Trained with AffectNet test with CK+**							**ShuffleNetV2 Trained with FER2013 test with CK+**						
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	11	4	0	2	31	51	0	89	0	0	0	11	0	0	27	2	13	0	24	33	0
Disgust	42	47	0	7	2	2	0	7	90	0	3	0	0	0	54	37	0	2	2	5	0
Fear	0	12	0	48	8	32	0	0	12	60	0	0	24	4	0	4	20	4	8	48	16
Happy	0	0	0	100	0	0	0	0	0	0	100	0	0	0	0	0	0	99	0	1	0
Neutral	3	6	0	7	58	24	2	2	0	1	5	91	1	0	0	0	1	2	89	7	2
Sad	4	18	0	4	7	64	4	0	0	0	0	4	96	0	0	0	18	0	14	68	0
Surprise	2	1	0	2	5	5	84	0	0	1	0	1	0	98	0	0	4	1	1	0	94

the models, the CK+ database was additionally used as a test set since it is one of the most widely used databases in the literature in emotion recognition applications.

After the end of the training, validation, and testing phases, the analysis indicated that MobileNetV2, trained and tested on the FER-2013 dataset, achieved an overall accuracy of 58%. In comparison, CLCM achieved an accuracy of 63%, while ShuffleNetV2 achieved the highest accuracy of 65% (see Table 4). Notably, ShuffleNetV2 and CLCM markedly outperformed MobileNetV2, while CLCM and ShuffleNetV2 showed comparable performance. Furthermore, evaluated with the CK+ dataset as a test set, these FER-2013 trained models revealed consistent performance trends with previous results. Specifically, MobileNetV2 achieved a performance level of 65%, while CLCM and ShuffleNetV2 reached an overall accuracy of 71% (see Table 5). Moreover, the performance of these models per class shows differences. Specifically, MobileNetV2 performed less than CLCM and ShuffleNetV2 in accurately discriminating between the 'angry' and 'fear' categories. However, CLCM and ShuffleNetV2 achieved close performance in these categories. In addition, MobileNetV2 struggled to discriminate the 'disgust' category, whereas CLCM and ShuffleNetV2 performed similarly and achieved comparable results in this particular category.

TABLE 10. Accuracy per-class comparison on FER-2013, RAF-DB, AffectNet using MobileNetV2, CLCM and ShuffleNetV2

	FER-2013			RAF-DB			AffectNet		
	MobileNetV2	CLCM	ShuffleNetV2	MobileNetV2	CLCM	ShuffleNetV2	MobileNetV2	CLCM	ShuffleNetV2
Angry	0.40	0.64	0.60	0.67	0.74	0.68	0.57	0.55	0.57
Disgust	0	0.44	0.52	0	0.49	0.44	0.30	0.24	0.31
Fear	0.17	0.50	0.47	0	0.56	0.54	0.42	0.35	0.42
Happy	0.84	0.82	0.86	0.94	0.94	0.91	0.94	0.95	0.93
Neutral	0.57	0.57	0.59	0.86	0.87	0.76	0.78	0.79	0.80
Sad	0.60	0.46	0.56	0.78	0.75	0.75	0.57	0.53	0.57
Surprise	0.84	0.78	0.77	0.14	0.85	0.77	0.40	0.36	0.41

TABLE 11. Overall Accuracy Performance Comparison on FER-2013, RAF-DB, AffectNet with different models from literature

FER-2013		RAF-DB		AffectNet	
Method	**Accuracy**	**Method**	**Accuracy**	**Method**	**Accuracy**
FERC [70]	0.54	DNN-CNN [71]	0.65	2Att-CNN [72]	0.49
MobileNetV2	0.58	AlexNet [6]	0.68	VGG-16 [57]	0.51
DenseNet121 [73]	0.59	VGG [6]	0.70	Deep CNN [74]	0.52
ResNet-50 [73]	0.60	MobileNetV2	0.73	DLP-CNN [57]	0.54
TouchyFeely [56]	0.61	Inception-V3 [75]	0.79	P-CNN [57]	0.54
Custom CNN [76]	0.62	ShuffleNetV2	0.80	**CLCM (Ours)**	**0.54**
EXP_DAL_MSE [77]	0.62	ResNet-50 [78]	0.80	EfficientNet-B0 [73]	0.55
MI+MII+MIII [55]	0.62	DenseNet121 [79]	0.81	ResNet-50 [73]	0.55
Tiny XCEPTION [58]	0.63	BaseDCNN [6]	0.82	ShuffleNetV2	0.57
EfficientNet-B0 [73]	0.63	Center loss [6]	0.83	MobileNetV2	0.57
CLCM (Ours)	**0.63**	EfficientNet-B0 [80]	0.84	DenseNet121 [73]	0.58
Inception-V3 [73]	0.65	Custom CNN [81]	0.84	Inception-V3 [73]	0.58
ShuffleNetV2	0.65	**CLCM (Ours)**	**0.84**	Custom CNN [81]	0.61

Notably, all three models showed similar performance levels in the 'happy,' 'neutral,' 'sad,' and 'surprise' classes (see Table 8).

Further analysis of the results showed that MobileNetV2, based on training and testing using the RAF-DB dataset, achieved an overall accuracy of 73%. In comparison, CLCM achieved an accuracy of 84%, while ShuffleNetV2 achieved an accuracy of 80% (see Table 4). These findings show that CLCM performs better on the RAF-DB dataset than ShuffleNetV2 and MobileNetV2. In addition, using the CK+ dataset as a test set, CLCM trained with RAF-DB showed a better performance than other models. In detail, CLCM achieved 78% accuracy while ShuffleNetV2 achieved 60% and MobileNetV2 achieved 47% accuracy (see Table 5). Evaluating the performance of the models trained and tested with RAF-DB for each emotion category, CLCM consistently outperforms MobileNetV2 and ShuffleNetV2 for the majority of emotions, namely 'anger,' 'disgust,' 'fear,' 'neutral' and 'surprise .'Furthermore, ShuffleNetV2 performs better than MobileNetV2 in most emotion categories, with performance levels close to CLCM's. In particular, MobileNetV2 shows failures in correctly recognizing the 'disgust' and 'fear' categories (see Table 8).

In the final phase of the analysis, MobileNetV2, CLCM, and ShuffleNetV2 models, trained on the AffectNet database, were tested on their respective test sets, following the same procedure used on the previous datasets. The results of these tests showed that MobileNetV2 and ShuffleNetV2 achieved an overall accuracy of 57%, while CLCM achieved a slightly lower accuracy of 54% (see Table 4). To extend these results, additional tests were performed using the CK+ dataset as the test set. In this scenario, MobileNetV2 and CLCM achieved an accuracy of 91%, while ShuffleNetV2 achieved a slightly higher accuracy of 92% (see Table 6). These results show that all three models achieved close results in two test scenarios. Evaluating the performance of the models trained and tested with AffectNet for each emotion category, MobileNetV2, and ShuffleNetV2 showed similar or close results in almost all categories. Although all models could recognize each emotion category, all the models had difficulty distinguishing the categories 'disgust' and 'surprise' from other emotions (see Table 8). Lastly, CLCM, ShuffleNetV2, and MobileNetV2 general emotion performance comparison are shown in Table 11.

A critical discussion can be drawn from Table 11, where past literature results are reported. Specifically, the CLCM model achieved an accuracy of 63% in FER-2013 and outperformed the reported methods of TouchyFeely [56], MI+MII+MIII [55], EXP-DAL-MSE [77], and had the same accuracy percentage of Tiny XCEPTION [58]. However, ShuffleNetV2 showed slightly better performance than CLCM with an accuracy of 65%. Furthermore, CLCM achieved 83% on the RAF-DB dataset, outperforming AlexNet [6], VGG [6], BaseDCNN [6], and MobileNetV2 and ShuffleNetV2. Lastly, CLCM achieved a performance of 0.54 on the AffectNet dataset, obtaining the same or better performance as 2att-CNN [72], VGG-16 [57], P-CNN [57], and DLP-CNN [57]. However, it slightly fell behind the 57% accuracy of MobileNetV2 and ShuffleNetV2. This finding may indicate that MobileNetV2 and ShuffleNetV2 have a slight performance advantage over CLCM with datasets containing larger images. It is worth noting that, unfortunately,

many authors did not report confusion matrices in the literature but only overall accuracy. That is the reason we discussed only the overall accuracy.

It is worth noting the results of recent studies using these datasets.(see Table 11). Mozaffari et al. [76] achieved 62% accuracy in their analyses using the FER-2013 dataset. Their model implemented batch normalization and did not balance image classes. Meanwhile, Sarvakar et al. [70] proposed a CNN model that achieved 54% accuracy, and these two models performed lower than the CLCM. Moreover, in studies conducted on the RAF-DB, Huang et al. [71] reported an accuracy of 65% and performed less accuracy than the CLCM model. Nevertheless, Gómez-Sirvent et al. [81] achieved an 84% accuracy, the same performance as CLCM. Lastly, in the largest dataset among recent studies, AffectNet, Tan Quan, et al. [74] obtained an accuracy of 52%, less than the performance of CLCM. Furthermore, Gómez-Sirvent et al. [81] outperformed CLCM by achieving a 61% accuracy (See Table 11).

Furthermore, several popular models cited in the literature, developed for various tasks, have also been tested for FER. For instance, Inception V3, which contains 24 million parameters, achieved an accuracy of 65% on the FER2013 dataset, 79% on RAF-DB, and 58% on AffectNet [73], [75]. Another notable compact architecture is EfficientNet-B0, with 4 million parameters. Studies indicate that EfficientNet-B0 attained an accuracy of 63% on the FER2013 dataset, 84% on RAF-DB, and 55% on AffectNet [73], [80]. Among these models, ResNet-50 stands out as one of the most complex and popular, containing 25 million parameters. It achieved accuracy 60% on FER2013, 80% on RAF-DB, and 55% on AffectNet [73], [78]. Additionally, DenseNet121, known for its comparatively smaller structure with 7 million parameters, achieved accuracies of 59% on FER2013, 81% on RAF-DB, and 58% on AffectNet [73], [79] (See Table 11).

Overall, recent studies and popular architectures demonstrate varying levels of accuracy, with some models outperforming CLCM and others underperforming. It's important to highlight that most of these studies involve larger and more complex models than the CLCM relying on their number of parameters. This is a factor that can directly affect performance. Additionally, it's important to emphasize that the primary objective of CLCM is to achieve optimal performance on devices with limited computational capacity.

In general, CLCM shows notable advances over many existing models. CLCM is characterized by its compact architecture, a feature shared by only a few models in the literature. This compactness and low inference time make CLCM a practical choice for mobile devices. Furthermore, considering that models with low power consumption are known to consume less power, the suitability of their use in mobile devices increases even more.

The results obtained on the CK+ database showed consistent performance of our CLCM architecture using all training datasets. In particular, our architecture showed good generalization capabilities even when trained with datasets with a limited number of images, such as FER-2013 and RAF-DB, cases for which MobileNetV2 and ShuffleNetV2 showed lower performance (Table 5). Although there are better-performing models in the literature [82], [83], these models mostly have numerous parameters and bigger structures than CLCM. We believe that the results obtained are attractive, as our study focuses on creating a lightweight model and reduced computational load. Compared to other models CLCM offers better solutions for the limited computational power devices in terms of inference time and flops. Moreover, our solution can be considered a step forward in developing compact and real-time models for daily life applications.

Reduced model complexity is a fundamental requirement for using artificial intelligence models in a real-time application, which is our goal for future developments. In this study, we assessed the complexity of the models by evaluating the number of trainable parameters, the number of FLOPs, and the inference time. The results showed that the CLCM model had fewer trainable parameters (2.4 million) than MobileNetV2 (3.5 million) and ShuffleNet V2 (3.9 million). A large number of trainable parameters, i.e., the number of parameters whose value is updated during the back-propagation process, impacts the training process by requiring more time and energy consumption. The parameters that identify the complexity of the model during the inference process, that is, the application of the pre-trained model to new data, are of greater interest. The CLCM model showed fewer FLOPs required to process input data compared to MobileNetV2 and ShuffleNetV2, as described in Table 6. This indicates a lower computational load and a higher sample processing speed, as confirmed by inference time analysis. Repeated model application on 50 samples produced a lower mean inference time (0.0502s) for the CLCM model, compared with MobileNetV2 (0.0584s) and ShuffleNetV2 (0.0633s). Anova post-hoc statistical analysis, Table 7, showed that the differences in inference time among the three models were statistically significant, thus confirming a shorter data processing time of the CLCM architecture and making it more suitable for real-time applications.

Nevertheless, the results from the CLCM show better or comparable performance to existing models in the literature. Specifically, the study by Lee and Wong, using the (2+1)D ConvNet ResNet20 model, achieved an inference time of 0.0527 seconds [84]. In contrast, the study by Xu et al. reported an inference time of 0.122 seconds for their model, which was designed to address real-world scenarios [85]. Furthermore, the study shows that the EfficientNet lite0 model has an average inference time of 0.280 seconds [85].

The comparison of flops (i.e., an indirect measure of computational complexity) between several studies highlights differences in performance and model structure. For instance, Barros et al.'s FaceChannel model has 3.6 million parameters and 0.633 Gflops, while Gera et al.'s CERN model has 1.45 million parameters and 1.781 Gflops. Lee et al.'s CAER-Net-S features 2.12 million parameters and 1.717 Gflops [86]–[89]. These studies show different levels of complexity due

to their implementation strategies. However, in comparison to CLCM, they perform less well in terms of Gflops, indicating differences in computational requirements.

Several limitations of this study must be addressed. In our study, we adopted a transfer learning approach by freezing some model layers and employing a new output classification layer. It is important to note that while we observed trends in performance, we did not quantitatively measure the direct impact of freezing specific layers or the final classifier on the overall model performance. Therefore, in future studies, it will be essential to systematically investigate and quantify the effects of different freezing approaches and final layer classifiers on model performance. In addition, we tested only a few hyperparameters in a trial-and-error approach. The results showed overfitting for the ShuffleNetV2 model, suggesting how the chosen hyperparameters were not optimal for that architecture. Therefore, future developments of this study should use different hyperparameter tuning techniques to optimize the models' performance.

Obtained results showed that all models had difficulties in recognizing the emotions "disgust" and "fear." This result is due to the complexity of these two particular facial emotions. Although both emotions have different substructures, they have more complex facial features than other emotions [90], [91]. In addition, these two emotions are the facial features of emotions that even human annotators have difficulty recognizing in daily life [92]. Moreover, most of the studies in the literature, as in this study, indicate that the facial features of "disgust" and "fear" emotions are challenging to recognize [90], [91]. Secondly, in most of the datasets used in the study (AffectNet, RAF-DB, FER2023), there is an imbalance between the number of images of emotion groups. Public datasets show a notable quantity of samples related to emotions characterized by easy identification and annotation, such as happiness. In contrast, the set of facial images corresponding to emotions that are difficult to identify and annotate, including disgust and fear, is limited [37]. This phenomenon affects the performance of the trained models in this study as in many other studies in the literature [37].

In future developments of this study, we will evaluate our model using a balanced training set obtained, for example, by mixing images from multiple data sources. In this regard, the test performed on the CK+ dataset showed that models trained on the AffectNet database, the most extensive database and the one exhibiting the greatest diversity among examples, performed better than models trained on smaller datasets. It suggests how an image set obtained from multiple data sources may improve model generalization capabilities. In addition, various techniques such as oversampling [93] or undersampling [93] may solve the sample imbalance issue in the datasets available in the literature and improve the performance. We also plan to work on model interpretability and experiment with other training strategies, overcoming limitations due to a lack of computational resources. In this regard, we plan to evaluate the integration of advanced architectures such as the Non-Flat Surface Level (NFSL) pyramid interconnection network [94], which represents a promising avenue for optimising image processing applications. Such a method may significantly improve the computational efficiency and performance of our approach. Similarly, the work of Mollajafari et al. [95], applied to cloud environments, offers interesting insights into the balance between computational cost and performance, which may apply to the optimisation process of neural network weights. Future research could explore how these architectural innovations and optimisation principles can be applied to improve our model's time and computational cost without compromising performance in emotion recognition.

Lastly, our future experiments might include evaluating the spatial complexities of our model and comparing it with the studies in the literature.

Future studies should focus on evaluating CLCM and similar models with lightweight architectures in real-world scenarios with different datasets [96]. In addition, particular emphasis should be placed on exploring the feasibility of integrating these models into web-based systems due to their lightweight characteristics. Scenarios in which FER is employed in mobile apps, e.g., real-time facial emotion recognition for special needs children, can shed light on the practical benefits and be helpful for the community [97].

Considering the rapid development of FER technology, its ethical and social implications should not be ignored. The increase in such models and their integration into more mobile devices raises ethical questions that may reduce privacy in both social and online environments. In addition, considering that social and interpersonal dynamics may also be affected, lawmakers and organizations must inform individuals about emerging technologies in the context of ethical issues.

VII. CONCLUSIONS

This paper introduces a lightweight model, called CLCM, to solve an important problem in facial expression recognition. The evaluations show that CLCM performs better than many models in the literature despite its smaller size. These evaluations highlight the potential of CLCM to provide better human-computer interaction, emotional recognition, and personalized user experience in real-world scenarios with limited computational power. Specifically, CLCM has great potential due to the compact structure for real-time emotion-based psychological and biofeedback studies. Moreover, the possible usage of CLCM with mobile devices is useful for daily life applications for the research horizons.

REFERENCES

[1] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The Influences of Emotion on Learning and Memory," *Frontiers in Psychology*, vol. 8, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01454

[2] R. Pandey and A. K. Choubey, "Emotion and health: An overview," *Journal of Projective Psychology & Mental Health*, vol. 17, pp. 135–152, 2010, place: India Publisher: Somatic Inkblot Society.

[3] M. N. Ab Wahab, A. Nazir, A. T. Zhen Ren, M. H. Mohd Noor, M. F. Akbar, and A. S. A. Mohamed, "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi,"

IEEE Access, vol. 9, pp. 134 065–134 080, 2021, conference Name: IEEE Access.

[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan. 2019, conference Name: IEEE Transactions on Affective Computing.

[5] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Li_Reliable_Crowdsourcing_and_CVPR_2017_paper.html

[6] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, Jan. 2019, conference Name: IEEE Transactions on Image Processing.

[7] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Heidelberg: Springer, 2013, pp. 117–124.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[9] L. K. McCorry, "Physiology of the Autonomic Nervous System," *American Journal of Pharmaceutical Education*, vol. 71, no. 4, p. 78, Aug. 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1959222/

[10] J. S. Barlow, *The Electroencephalogram: Its Patterns and Origins*. MIT Press, 1993.

[11] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, "Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review," *Sensors*, vol. 21, no. 15, p. 5015, Jan. 2021, number: 15 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/21/15/5015

[12] F. Ekman, "Facial Action Coding System," *Environmental Psychology & Nonverbal Behavior*, 1978. [Online]. Available: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft27734-000

[13] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness," *Scientific Reports*, vol. 11, no. 1, p. 5214, Mar. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-021-84632-9

[14] R. Leung, "Subsurface Boundary Geometry Modeling: Applying Computational Physics, Computer Vision, and Signal Processing Techniques to Geoscience," *IEEE Access*, vol. 7, pp. 161 680–161 696, 2019, conference Name: IEEE Access.

[15] J. d. A. Dornelles, N. F. Ayala, and A. G. Frank, "Smart Working in Industry 4.0: How digital technologies enhance manufacturing workers' activities," *Computers & Industrial Engineering*, vol. 163, p. 107804, Jan. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835221007087

[16] Y. Cai, W. Zheng, T. Zhang, Q. Li, Z. Cui, and J. Ye, "Video based emotion recognition using cnn and brnn," in *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II 7*. Springer, 2016, pp. 679–691.

[17] E. C. Nook, K. A. Lindquist, and J. Zaki, "A new look at emotion perception: Concepts speed and shape facial emotion recognition," *Emotion (Washington, D.C.)*, vol. 15, no. 5, pp. 569–578, Oct. 2015.

[18] J. M. Leppänen and C. A. Nelson, "The development and neural bases of facial emotion recognition," in *Advances in Child Development and Behavior*, R. V. Kail, Ed. JAI, Jan. 2006, vol. 34, pp. 207–246. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S006524070680008X

[19] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, Oct. 1999, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[20] D. Mehta, M. F. H. Siddiqui, and A. Y. Javaid, "Recognition of Emotion Intensities Using Machine Learning Algorithms: A Comparative Study," *Sensors*, vol. 19, no. 8, p. 1897, Jan. 2019, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/19/8/1897

[21] S. Begaj, A. O. Topal, and M. Ali, "Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)," in *2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, Dec. 2020, pp. 58–63.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, arXiv:1409.1556 [cs]. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 10, 2015.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[25] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[26] M. Ali, A. A. Ali, A.-E. Taha, I. B. Dhaou, and T. N. Gia, "Intelligent autonomous elderly patient home monitoring system," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[27] B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, "Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings?" *BMJ global health*, vol. 3, no. 4, p. e000798, 2018.

[28] S. Guerreiro-Santalla, A. Mallo, T. Baamonde, and F. Bellas, "Smartphone-based game development to introduce k12 students in applied artificial intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 758–12 765.

[29] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized Convolutional Neural Networks for Mobile Devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4820–4828. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Wu_Quantized_Convolutional_Neural_CVPR_2016_paper.html

[30] 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," Apr. 2017, arXiv:1704.04861 [cs]. [Online]. Available: http://arxiv.org/abs/1704.04861

[31] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 6848–6856. [Online]. Available: https://ieeexplore.ieee.org/document/8578814/

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. [Online]. Available: https://ieeexplore.ieee.org/document/8578572/

[33] F. Saxen, P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with mobilenetv2 and nasnet-mobile," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2019, pp. 176–180.

[34] L. Torrey and J. Shavlik, "Transfer Learning," 2010, iSBN: 9781605667669 Pages: 242-264 Publisher: IGI Global. [Online]. Available: https://www.igi-global.com/chapter/transfer-learning/www.igi-global.com/chapter/transfer-learning/36988

[35] G. Laborde, *Learning TensorFlow. js*. " O'Reilly Media, Inc.", 2021.

[36] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[37] M. Koziarski, B. Kwolek, and B. Cyganek, "Convolutional Neural Network-Based Classification of Histopathological Images Affected by Data Imbalance," in *Video Analytics. Face and Facial Expression Recognition*, ser. Lecture Notes in Computer Science, X. Bai, Y. Fang, Y. Jia, M. Kan, S. Shan, C. Shen, J. Wang, G.-S. Xia, S. Yan, Z. Zhang, K. Nasrollahi, G. Hua, T. B. Moeslund, and Q. Ji, Eds. Cham: Springer International Publishing, 2019, pp. 1–11.

[38] B. A. Shawar and E. S. Atwell, "Using corpora in machine-learning chatbot systems," *International Journal of Corpus Linguistics*, vol. 10, no. 4, pp. 489–516, Jan. 2005, publisher: John Benjamins. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/ijcl.10.4.06sha

[39] J. R. Lee, L. Wang, and A. Wong, "Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition," *Frontiers in Artificial Intelligence*, vol. 3, p. 609673, 2021.

[40] J. Stilgoe, "Machine learning, social learning and the governance of self-driving cars," *Social studies of science*, vol. 48, no. 1, pp. 25–56, 2018.

[41] A. M. Pinto, L. F. Rocha, and A. Paulo Moreira, "Object recognition using laser range finder and machine learning techniques," *Robotics and Computer-Integrated Manufacturing*, vol. 29, no. 1, pp. 12–22, Feb. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584512000798

[42] S. Sharma, M. Bhatt, and P. Sharma, "Face Recognition System Using Machine Learning Algorithm," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Jun. 2020, pp. 1162–1168.

[43] S. Loussaief and A. Abdelkrim, "Machine learning framework for image classification," in *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Dec. 2016, pp. 58–61.

[44] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," Dec. 2016, arXiv:1612.02903 [cs]. [Online]. Available: http://arxiv.org/abs/1612.02903

[45] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, Feb. 2020. [Online]. Available: https://doi.org/10.1007/s42452-020-2234-1

[46] M. Coskun, A. Uçar, O. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," in *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, Nov. 2017, pp. 376–379.

[47] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, conference Name: Neural Computation.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: https://doi.org/10.1145/3065386

[49] Y. Tang, "Deep Learning using Linear Support Vector Machines," Feb. 2015, arXiv:1306.0239 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1306.0239

[50] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gulcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Cote, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, ser. ICMI '13. New York, NY, USA: Association for Computing Machinery, Dec. 2013, pp. 543–550. [Online]. Available: https://doi.org/10.1145/2522848.2531745

[51] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

[52] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally Parameterized Convolutions for Efficient Inference," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/f2201f5191c4e92cc5af043eebfd0946-Abstract.html

[53] S. Sharma and V. Kumar, "Performance evaluation of machine learning based face recognition techniques," *Wireless Personal Communications*, vol. 118, pp. 3403–3433, 2021.

[54] C. Qi, M. Li, Q. Wang, H. Zhang, J. Xing, Z. Gao, and H. Zhang, "Facial expressions recognition based on cognition and mapped binary patterns," *IEEE Access*, vol. 6, pp. 18 795–18 803, 2018.

[55] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 423–430.

[56] D. Amin, P. Chase, and K. Sinha, "Touchy feely: An emotion recognition challenge," *Palo alto: Stanford*, 2017.

[57] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2209–2214.

[58] T. Raksarikorn and T. Kangkachit, "Facial expression classification using deep extreme inception networks," in *2018 15th international joint conference on computer science and software engineering (JCSSE)*. IEEE, 2018, pp. 1–5.

[59] F. Radenović, G. Tolias, and O. Chum, "Fine-Tuning CNN Image Retrieval with No Human Annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[60] K. Alissa, R. Obeidat, S. Alqudah, R. Obeidat, and Q. Ismail, "Performance Evaluation of CNN-based Transfer Learning for COVID-19 Pneumonia Identification with Various Levels of Layer Partial Freezing," in *2022 International Conference on Engineering & MIS (ICEMIS)*, Jul. 2022, pp. 1–8.

[61] D. A. van Dyk and X.-L. Meng, "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, Mar. 2001, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/10618600152418584. [Online]. Available: https://doi.org/10.1198/10618600152418584

[62] T. U. Ahmed, S. Hossain, M. S. Hossain, R. ul Islam, and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2019, pp. 336–341.

[63] E. S. Hussein, U. Qidwai, and M. Al-Meer, "Emotional stability detection using convolutional neural networks," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE, 2020, pp. 136–140.

[64] M. Rao, R. Bao, and L. Dong, "Face emotion recognization using dataset augmentation based on neural network," in *Proceedings of the 6th International Conference on Graphics and Signal Processing*, 2022, pp. 1–6.

[65] D. Ammous, A. Chabbouh, A. Edhib, A. Chaari, F. Kammoun, N. Masmoudi *et al.*, "Designing an efficient system for emotion recognition using cnn," *Journal of Electrical and Computer Engineering*, vol. 2023, 2023.

[66] S. Porcu, A. Floris, and L. Atzori, "Evaluation of data augmentation techniques for facial expression recognition systems," *Electronics*, vol. 9, no. 11, p. 1892, 2020.

[67] X. Tang, X. Li, Y. Ding, M. Song, and Y. Bu, "The pace of artificial intelligence innovations: Speed, talent, and trial-and-error," *Journal of Informetrics*, vol. 14, no. 4, p. 101094, 2020.

[68] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," in *2020 6th International conference on advanced computing and communication Systems (ICACCS)*. IEEE, 2020, pp. 317–320.

[69] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[70] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. S. Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Materials Today: Proceedings*, vol. 80, pp. 3560–3564, 2023.

[71] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition," *Scientific Reports*, vol. 13, no. 1, p. 8425, 2023.

[72] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 217–225, 2019.

[73] C.-T. Yen and K.-H. Li, "Discussions of different deep transfer learning models for emotion recognitions," *IEEE Access*, vol. 10, pp. 102 860–102 875, 2022.

[74] T. Q. Ngo and S. Yoon, "Facial expression recognition on static images," in *Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings 6*. Springer, 2019, pp. 640–647.

[75] Y. Ma and C. Huang, "Facial expression recognition based on deep learning and attention mechanism," in *Proceedings of the 3rd International Conference on Advanced Information Science and System*, 2021, pp. 1–6.

[76] L. Mozaffari, M. M. Brekke, B. Gajaruban, D. Purba, and J. Zhang, "Facial expression recognition using deep neural network," in *2023 3rd International Conference on applied artificial intelligence (ICAPAI)*. IEEE, 2023, pp. 1–9.

[77] Y. Zhai and J. Liu, "Facial expression recognition based on transferring convolutional neural network," *Journal of Signal Processing*, vol. 34, pp. 729–738, 2018.

[78] S. Xie, M. Li, S. Liu, and X. Tang, "Resnet with attention mechanism and deformable convolution for facial expression recognition," in *2021 4th International Conference on Information Communication and Signal Processing (ICICSP)*. IEEE, 2021, pp. 389–393.

[79] J. Chandra, B. Annappa *et al.*, "Cross-database facial expression recognition using cnn with attention mechanism," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–7.

[80] A. A. H. Qutub and Y. Atay, "Deep learning approaches for classification of emotion recognition based on facial expressions," *Nexo Revista Científica*, vol. 35, no. 05, pp. 1–18, 2023.

[81] J. L. Gómez-Sirvent, F. López de la Rosa, M. T. López, and A. Fernández-Caballero, "Facial expression recognition in the wild for low-resolution images using voting residual network," *Electronics*, vol. 12, no. 18, p. 3837, 2023.

[82] G. Liang, S. Wang, and C. Wang, "Pose-Invariant Facial Expression Recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Dec. 2021, pp. 01–08.

[83] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, Sep. 2021, pp. 119–124, iSSN: 1949-0488.

[84] J. R. H. Lee and A. Wong, "Timeconvnets: A deep time windowed convolution neural network design for real-time video facial expression recognition," in *2020 17th Conference on Computer and Robot Vision (CRV)*. IEEE, 2020, pp. 9–16.

[85] G. Xu, H. Yin, and J. Yang, "Facial expression recognition based on convolutional neural networks and edge computing," in *2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*. IEEE, 2020, pp. 226–232.

[86] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.

[87] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[88] D. Gera, S. Balasubramanian, and A. Jami, "Cern: Compact facial expression recognition net," *Pattern Recognition Letters*, vol. 155, pp. 9–18, 2022.

[89] P. Barros, N. Churamani, and A. Sciutti, "The facechannel: a fast and furious deep neural network for facial expression recognition," *SN Computer Science*, vol. 1, pp. 1–10, 2020.

[90] Z. Ullah, M. I. Mohmand, S. U. Rehman, M. Zubair, M. Driss, W. Boulila, R. Sheikh, and I. Alwawi, "Emotion recognition from occluded facial images using deep ensemble model," *Cmc-Computers Materials & Continua*, vol. 73, no. 3, pp. 4465–4487, 2022.

[91] C. Hewitt and H. Gunes, "Cnn-based facial affect analysis on mobile devices," *arXiv preprint arXiv:1807.08775*, 2018.

[92] K. Krishnaveni *et al.*, "A novel framework using binary attention mechanism based deep convolution neural network for face emotion recognition," *Measurement: Sensors*, vol. 30, p. 100881, 2023.

[93] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th international conference on information and communication systems (ICICS)*. IEEE, 2020, pp. 243–248.

[94] H. Shahhoseini, E. Kandzi, and M. Mollajafari, "Nonflat surface level pyramid: A high connectivity multidimensional interconnection network," *The Journal of Supercomputing*, vol. 67, pp. 31–46, 01 2014.

[95] M. Mollajafari and M. Shojaeefard, "Tc3pop: a time-cost compromised workflow scheduling heuristic customized for cloud environments," *Cluster Computing*, vol. 24, pp. 1–18, 09 2021.

[96] M. Krumnikl and V. Maiwald, "Facial emotion recognition for mobile devices: A practical review," *IEEE Access*, 2024.

[97] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," *Soft Computing*, pp. 1–14, 2024.

MUSTAFA CAN GURSESLI received his master's degree in Clinical Psychology from Bergamo University, Italy, in 2020. He worked in Emerging Technologies Lab as an assistant researcher at Tampere University, Finland. He is pursuing his PhD in Information Engineering at the University of Firenze and working in the Virtual Human Dynamics Laboratory. Mustafa Can Gursesli is an IEEE Entertainment and Gaming (ENT) Technical Committee member. His research interests are Clinical Psychology, Game Studies, Social Psychology, Artificial Intelligence, and Computer Vision.

Sara Lombardi is a Biomedical Engineer and currently a PhD student at the Department of Information Engineering of the University of Florence. Her academic background includes a bachelor's degree in electronic and telecommunications engineering and a master's degree in biomedical engineering. In 2022, she worked as a research fellow in physiological signal modeling for microcirculation analysis at the Department of Information Engineering, University of Florence. Her research interests include biosignal analysis, bioimage processing, and Artificial Intelligence. She is a member of the IEEE and the Engineering in Medicine and Biology Society.

Mirko Duradoni Ph.D. obtained his title in Information Engineering in 2020, and he holds a master's degree in psychology from the Department of Education and Psychology, University of Florence, Italy. He is currently a member of the Virtual Human Dynamics Laboratory at the University of Florence, a contract professor in the Psychology of groups, and a teaching fellow in environmental psychology. He is particularly skilled in psychological dimensions assessment, research methodology, and data analysis. He taught psychometrics in the health sector and was enrolled in three EU-level projects (InSPIRES, Erasmus+ Restore, "PHOENIX).).

Leonardo Bocchi is currently an associate professor in Biomedical Engineering at the University of Florence, Italy. He is a member of the IEEE Technical Committee on Cardiovascular Systems, a project evaluator and reviewer for the EU and national projects; he acts as a referee for several international journals (among others, IEEE TMI, IEEE TBME, Pattern Recognition, Signal Processing, Biomedical Signal Processing, and Control) and Associate Editor in Frontier on Network Physiology. Under different roles, he participated in organizing various conferences (EuroGP & EvoWorkshops, ACIVS, MAVEBA, Interspeech). His research interests include medical image analysis and interpretation, modeling of physiological systems, noninvasive early diagnosis with Artificial Intelligence techniques (in particular on COVID and sepsis), and he is the author of about 120 indexed publications in the field of biomedical engineering.

Andrea Guazzini Ph.D. is an Italian researcher of the Department of Education, Literatures, Intercultural Studies, Languages and Psychology at the University of Florence, and he is an associate researcher of the Centre for the Study of Complex Dynamics (CSDC) of the University of Florence. AG is an experimental psychologist and received his Ph.D. in Complex Systems and Nonlinear Dynamics from the University of Florence (2008). Currently, his research interests range from the computational modeling of human cognition and (virtual) social dynamics to the study of the psychology of virtual environments. Finally, he is currently the supervisor of the Laboratory for the Study of Human Virtual Dynamics - VirtHuLab, of the University of Florence.

Antonio Lanata Ph.D. is currently an Associate Professor of Bioengineering at the Department of Information Engineering, University of Florence, Italy. His research interests include designing and implementing wearable systems for physiological monitoring and statistical and nonlinear biomedical signal processing. Applications of his research include the assessment of the autonomic nervous system on affective computing, assessment of mood and mental/neurological disorders, and human/animal/robot interaction. He is the author of more than 130 international scientific contributions. He is an Associate Editor of several International journals such as Frontiers Bioelectronics, MDPI Bioengineering, Bioelectronics, Biosensors, Algorithms, Electronics, and Animals. He is currently the head of the Laboratory B3Lab (Biosystems, Biosignals, and Bioimaging) of the University of Florence. A. Lanata is a member of the IEEE, the IEEE Circuits and Systems Society, Engineering in Medicine and Biology, Information Theory, and Signal Processing Society.