IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2023, accepted 19 January 2024, date of publication 24 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358207

RESEARCH ARTICLE

Using Diverse ConvNets to Classify Face Action Units in Dataset on Emotions Among Mexicans (DEM)

MARCO A. MORENO-ARMENDÁRIZ, ALBERTO ESPINOSA-JUAREZ, AND ESMERALDA GODINEZ-MONTERO

Computational Cognitive Sciences Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Mexico City 07738, Mexico

Corresponding author: Alberto Espinosa-Juarez (aespinosaj2021@cic.ipn.mx)

This work was supported in part by Instituto Politecnico Nacional through Instituto Politécnico Nacional, Secretaría de Investigación y Posgrado (IPN-SIP) Research under Grant SIP-2259 and Grant SIP-20240666, in part by IPN-Comisión de Operación y Fomento de Actividades Académicas (COFAA), in part by IPN- Estímulos al Desempeño de los Investigadores (EDI), and in part by Consejo Nacional de Humanidades Ciencias y Tecnologías, Sistema Nacional de Investigadoras e Investigadores (CONAHCYT-SNII).

ABSTRACT To understand how Convolutional Neural Networks (ConvNets) perceive the muscular movements of the human face, known as Action Units (AUs) in this work, we introduce a new dataset named Dataset on Emotions among Mexicans (DEM), consisting of 1557 images of Mexicans labeled with twenty-six AUs and seven emotions. As a benchmark, we used the comparison with DISFA+ labeled with 12 AUs. To address the task of detecting AUs in each image, six ConvNets were employed, and we evaluated their performance using the F1 Score. The two ConvNets with the best performance were VGG19 with 0.8180% (DEM), 0.9106 % (DISFA+), and ShuffleNetV2 with 0.7154% (DEM), 0.9440% (DISFA+). Subsequently, these ConvNets were analyzed using Grad-CAM and Grad-CAM++; this algorithms allows us to observe the areas of the face considered for prediction. In most cases, these areas consider the region of the labeled AU. Considering the F1 score and the visual study, we can conclude that using DEM as a dataset to classify AUs is promising since the experiments achieved performances similar to those of the current literature that only use ConvNets.

INDEX TERMS Action units, ConvNets, CAM analysis.

I. INTRODUCTION

Nonverbal Nonverbal communication is an essential aspect of human interaction, where facial expressions play a pivotal role in conveying emotions, intentions, and affective states. Since ancient times, humans have endeavored to comprehend and decode facial language, recognizing the significance of these visual cues in interpreting and understanding the intentions and emotions of others.

The study of facial expressions has evolved from subjective approaches based on intuition and observation to objective and systematic methodologies aimed at decomposing and encoding facial movements into specific units. In this context, the Facial Action Coding System (FACS) [1], developed by Paul Ekman and Wallace Friesen in 1978, has emerged as a central tool in analyzing facial expressions.

The Facial Action Coding System (FACS) serves as a framework for encoding and describing visually discernible muscular movements of the face, offering a precise and detailed representation of facial expressions. FACS introduces the concept of breaking down facial expressions into fundamental units known as Action Units (AU), each corresponding to specific muscle movements. Each AU is assigned a code, facilitating an objective and quantitative description of facial expressions.

FACS has proven to be a valuable tool across various fields of study and applications. FACS has been employed in psychology and neuroscience to investigate emotional responses, cognitive processes, and psychological disorders. For instance, in psychology, FACS has been used to explore cognitive and emotional processes [2], [3], [4], enabling an

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

objective and quantitative assessment of facial expressions in different contexts. It has also found utility in Artificial Intelligence (AI), where FACS has become a tool for automatic emotion recognition [5], [6].

Furthermore, FACS has found applications in fields such as medicine and therapy, where it has been employed for diagnosing and treating disorders related to facial expressions, such as Möbius syndrome [7] and facial paralysis [8].

One of the most well-known approaches in this field is using deep learning techniques, particularly Convolutional Neural Networks (ConvNets). These networks aim to extract relevant features from images and learn complex patterns. ConvNets have proven highly effective in image classification tasks, and their application to facial expression analysis has led to significant improvements in the accuracy and performance of facial recognition systems.

Studies attempt to classify AUs using the FACS as a basis for Deep Learning methods. These works include [9], which evaluates the visual transformers ViT and SWIN for AU classification using the DISFA+ dataset [10]. Besides normalizing the images, they perform facial alignment, horizontal flips, and rotations for preprocessing. The results recorded in their experiments show an average F1 score of 60% using SWIN and 54% using ViT.

There is also the work of [11], in which they propose an Attention-based Relationship Network (ABRNet) for AU classification. In this study, ABRNet utilizes multiple layers of relationship learning to capture different AU relationships automatically. These are then introduced into a self-attention fusion module to refine individual AU features with attention weights, thus enhancing feature robustness.

In [12], the authors investigated how to integrate the propagation of semantic relationships between AUs and a deep neural network to enhance the feature representation of facial regions. For this purpose, they constructed a structured knowledge graph of AUs. They integrated a Gated Graph Neural Network (GGNN) to propagate node information through the graph to generate an enhanced representation of AUs. The model uses two AU-labeled datasets: BP4D [13] (utilizing 12 AUs) and DISFA [14] (utilizing 8 AUs) for evaluation. In DISFA, since it is annotated with AU intensity ranging from 0 to 5, they took intensity two as a reference to indicate the presence of the AU.

Consequently, AUs are tagged as existing if the intensity is equal to or greater than two; otherwise, as absent. This work achieves an average F1 score of 62.9% on BP4D and 55.9% on DISFA.

We also have the work of [15], in which they employed a new Facial Action Units (FAU) correlation network based on a transformer encoder architecture to capture the relationships between the different AUs using the DISFA dataset. They detected facial regions for preprocessing and resized them to 224 × 224 grayscale pixels. Similarly to the previous work, they also decided to use intensity two as a discriminator for the appearance or absence of an AU. The results on DISFA showed a 61.5% F1 score.

TABLE 1. Total samples per emotion in DEM.

Emotion	Total
Contempt	171
Disgust	196
Anger	171
Happiness	258
Fear	130
Surprise	208
Sadness	138
Neutral	285

We also found works such as the one by [16] in which, in addition to classifying the AU, it also performs facial alignment, not as part of the preprocessing but as part of the learning of the neural model that uses an attention-learning module. Their evaluation in F1 score shows an average of 56%.

In brief, our main contributions are:

1) We introduce a novel dataset called Dataset on Emotions among Mexicans (DEM). DEM included 1557 images of Mexican individuals with the following labels: AUs, 0-5 intensity of the AU, and seven emotions.
2) Our experimental results using CAM algorithms show that ConvNets trained with DEM and DISFA+ correctly differentiate, in most cases, the visual areas where an action unit appears, even in images that do not belong to the training set.
3) In F1 score percentages, DEM performs similarly to those presented in the literature using only ConvNets.

II. MATERIALS AND METHODS

In the current literature, there are various datasets created to analyze Action Units, such as CK/CK+ [17], [5], BP4D [13], CASME [18], MMI [19], Bosphorous [20], among others, in this work, we use the following ones.

A. DATASETS

1) DEM

In this work, we present the Dataset on Emotions among Mexicans (DEM) for the first time, a dataset with Latin American faces labeled with AUs and emotions. It includes 246 participants, comprising students and professors from universities and research centers in Mexico City. Each volunteer had a maximum of eight photographs taken, resulting in a total collection of 1,942 images at a resolution of 6016 × 4016 pixels in RGB. We showed each participant pictures of eight emotions they had to portray, and we took photographs when the muscular movement resembled the ideal model. DEM includes labels for AUs with intensities ranging from 0-5, as well as labels for facial expressions that represent human emotions. Five experts performed the labeling of DEM. After individual labeling, we carried out bias reduction through a majority vote. Thus, if three individuals labeled an image the same way, it was considered

TABLE 2. Action units in DISFA+ and DEM for our experiments.

Action Unit	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU9	Nose Wrinkler
AU12	Lip Corner Pull
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU25	Lips Part
AU26	Jaw Drop

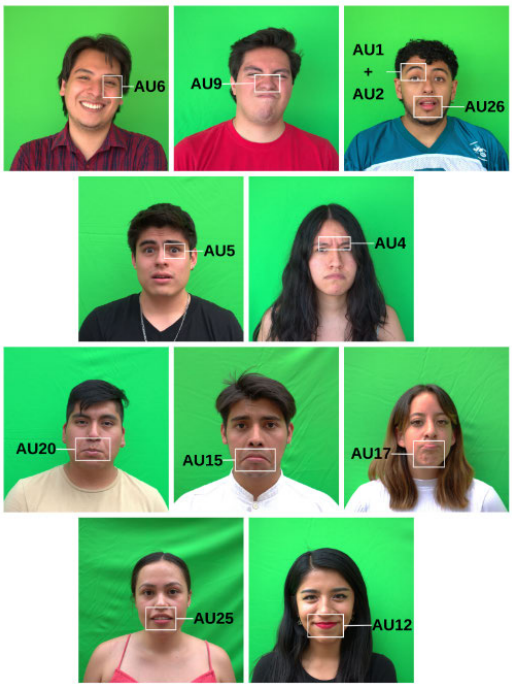

FIGURE 1. Example of AUs in DEM.

valid. This filtering discarded 20% of the images, resulting in 1557 images divided into seven emotions and a neutral face as can be seen in Table 1. To label the AUs it was necessary to remove the 285 neutral face images because no AUs are found in the neural faces.

DEM includes 26 labeled AUs; however, for comparison with DISFA+, only the 12 AUs that DISFA+ contains were used, listed in Table 2. To better exemplify the DEM distribution, Figure 1 shows some examples of images included in DEM that visually show the AUs used in this work. The nomenclature used is defined in FACS, aligning with the psychological and clinical literature.

It is important to note that the number of main AUs in the FACS varies between 28-30, and if we consider AUs with movement, the number reaches 58. These expressions spontaneously appear on the face, making it an extremely challenging task to capture the exact moment they occur and obtain a sufficient number of images for each, to the extent that, even today, there is no dataset in the literature that includes all of them.

2) DISFA+

We selected DISFA+ because it is a recently created dataset and a significant update to DISFA [14]. DISFA+ consists of nine individuals recorded in various contexts in a controlled environment to capture twelve AUs corresponding to the twelve classes to be classified. The dataset includes posed and non-posed action units. Each of the twelve AUs is also labeled by intensities ranging from 0-5, where zero signifies the absence of the AU, and 5 represents the maximum visual presence. It is essential to mention that while the dataset contains approximately 57,000 images, the total content of the dataset is around 96,000 AUs. This difference is because each image can have more than one AU labeled.

B. DEEP NEURAL NETWORKS

From the literature, we chose to conduct experiments with six Convolutional Neural Networks (ConvNets): VGG19 [21], ResNet101 [22], NASNet Mobile [23], MobileNetV2 [24], EfficientNetB0 [25], and ShuffleNetV2 [26]. We selected these neural networks as they have been employed for facial and emotion classification tasks, yielding favorable results, as observed with VGG19 in [27], ResNet101 in [28], and NASNet, MobileNetV2, EfficientNetB0, and ShuffleNetV2 in the works of [29], [30], [31], and [32] respectively. Therefore, we consider these ConvNets excellent candidates for classifying facial gestures and emotions, although this study will not perform the latter task. Below, we provide a brief description of each of the ConvNets used.

1) VGG19

This ConvNet was one of the earliest ConvNets with significant feature extraction power, maximizing 3×3 convolutions while maintaining a straightforward architecture (cascading convolutions with increasing filters in each convolutional block). Although its training is more time-consuming due to the number of filters and cascading convolutions, its effectiveness in classification tasks has made it one of the most widely used ConvNets in computer vision research.

2) RESNET101

On the other hand, ResNet101 introduced residual connections that help alleviate the vanishing gradient problem, a common issue in artificial neural networks with considerable depth. This problem prevents the weights in the initial and intermediate layers of the networks from updating during backpropagation because, as the weights of the last layers (the first to be updated during backpropagation) are updated, the gradients become progressively smaller, tending to zero in the initial layers. Another reason we chose ResNet101 is its use of bottlenecks, which are blocks employing 1×1 convolutional layers [33] acting as bottlenecks within the network. They reduce dimensions, then use 3×3 convolutions and again

1×1 convolutions to return to the original dimensions, thereby reducing the number of network parameters and matrix operations.

3) MOBILENETV2

MobileNetV2 introduces the inverted residual structure while employing Depthwise Separable Convolution implemented in MobileNetV1 [34]–convolutional layers widely used in contemporary convolutional architectures. MobileNetV2 utilizes two blocks: the first is a residual block with a stride of 1, and the second has a stride of 2 for size reduction. Each block starts with a 1×1 convolutional layer with ReLU6 activation function [34], followed by a depthwise convolution layer, and concludes with a 1×1 convolutional layer.

4) NASNET MOBILE

On the other hand, the authors constructed this Neural Network using the Neural Architecture Search (NAS) algorithm to find the most suitable network configuration for solving ImageNet. The objective was to achieve the highest accuracy while minimizing the number of parameters. It is essential to mention that NAS uses defined variations of convolutions to test and discover the best combination. Among the possible options is the use of 1×1 convolutions, 3×3 convolutions, dilated convolutions [35] (a type of convolution that inserts spaces between kernels, thus "skipping" pixels during convolution), depthwise separable convolutions (a convolution that divides the filter into two separate filters performing depthwise convolution and pointwise convolution, saving computational operations) introduced in MobileNetV1, as well as identity connections–a type of residual connection implemented in ResNet101. For this ConvNet, there is no figure illustrating the backbone of the architecture because, as it is a pseudo-randomly created architecture, there is no defined construction pattern. Additionally, the final architecture consists of 389 layers in depth.

5) EFFICENTNETB0

EfficientNetB0 can be used in classification, detection, segmentation, and even some natural language processing tasks. They achieved this by relying on two simple principles: efficiency and performance. ConvNet's architecture is based on compound scaling, which aims to balance the network's size, accuracy, and computational cost. This balance is achieved by scaling three essential dimensions of a ConvNet:

- Width: This dimension refers to the number of channels per convolutional layer. Increasing this dimension allows the capturing of increasingly abstract patterns.
- Depth: It refers to the number of layers a ConvNet can have. More layers imply the ability to represent more complex data. A ConvNet with fewer layers is more computationally efficient but may sacrifice precision.
- Resolution: Scaling the input images can lead to the loss or gain of helpful information in the ConvNet. For instance, higher-resolution input may provide more detailed information but requires more memory.

Therefore, the creation of this ConvNet involves balancing these three dimensions. Researchers used grid search to find the optimal combination of width, depth, and resolution to achieve this. This search is guided by a compound coefficient, denoted as "phi" (ϕ), which uniformly scales the model's dimensions. The user provides this coefficient at the beginning of the search. A significant ϕ results in a more practical but computationally more expensive model, while a small ϕ produces a lighter but more computationally efficient model.

6) SHUFFLENETV2

The researchers designed this network to achieve good results with lower computational cost, aiming for quick and efficient outcomes. To fulfill this objective, ConvNet proposes the following guidelines:

- Equal channel width minimizes Memory Access Cost (MAC), meaning maintaining a 1:1 ratio.
- Excessive group convolution increases MAC; increasing the number of groups leads to more computations, ultimately reducing speed.
- Network fragmentation reduces the degree of parallelism: Fragmentation is inversely proportional to parallel computation.
- Element-wise operations are not negligible: Runtime decomposition graphs show that simple element-wise operations impose overhead on speed.

7) CLASSIFIER

First, we add a global average max pooling layer to flatten the feature vector and immediately add a multilayer perceptron (MLP) to all the previously mentioned networks. This MLP consists of two fully connected layers: the first with 512 neurons and ReLU activation function, and the second with 12 neurons and Sigmoid activation function. The latter layer serves as the output of the classifier.

C. PREPROCESSING AND SEPARATION OF THE DATASET

1) INITIAL PREPROCESSING

All images in DEM and DISFA+ contain elements irrelevant to this task, such as the image capture date, additional subjects in the image, a significant amount of background, or even clothing. Examples can be seen in Figure 2, where Figure 2a shows a DISFA+ image with a person in the background and additional details like the capture date and time, and in Figure 2b, an example from DEM. To eliminate these unwanted elements and focus solely on the faces in the images, we decided to employ a Cascade Classifier that utilizes the Haar Cascade Filters [36]. This approach allowed us to crop the images in both datasets, leaving only the faces, as depicted in Figure 3. The final image sizes were 244×244 for DEM and 350×350 for DISFA+. However, in the case of DISFA+, the size was subsequently reduced to

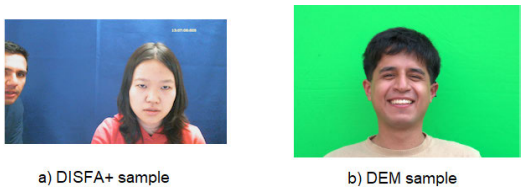

FIGURE 2. Sample image in DISFA+ and DEM.

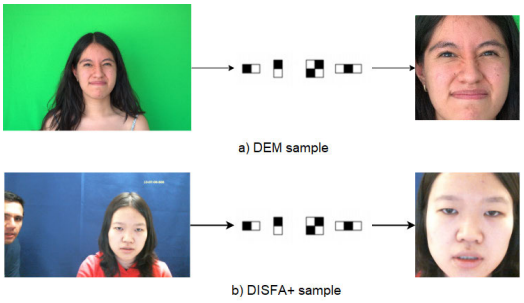

FIGURE 3. Diagram of image preprocessing in DISFA+ and DEM.

224 × 224 to align with the input dimensions of the selected ConvNets.

Figure 3 provides a general overview of selecting regions of interest and the final result. Figure 3a presents the step of passing a DEM image through the Cascade Classifiers, and in Figure 3b, the same procedure can be observed with DISFA+.

With the cropped faces, we create a tabular data structure to have the images and their labels in two simple columns. In the first column, we placed an image; in the second column, it is the list of labels representing a multi-label format; this is a significant difference in our approach compared to the state-of-the-art, where works typically use multi-class labeling.

DISFA+ and DEM initially consist of images labeled with the corresponding AU and their intensity (ranging from zero to five), where zero signifies no presence of the AU, and five indicates maximum presence. Following the same approach as state of the art, we decided to transform this problem into a binary classification task, using the intensity just as a threshold; so we assign the value one to AUs with an intensity of two or greater and the value zero to all AUs with an intensity of less than two. Also, we removed the images without labeled AUs. As shown in Tables 3-4, 70% of the total images in both datasets were used for training, 10% for validation, and 20% for testing.

In Table 4, it can be observed that DEM has few images for AU5, AU15, AU17, and AU20. To address the class imbalance issue, an artificial data augmentation was performed. The procedure begins with AU20, which has less data. We took the 14 images from the original set with AU20, and 10 artificial images were generated for each, resulting in 140 images. It is important to note that some images have more than one AU, so other AUs were also augmented. We repeated the process for AU17, which was sufficient to balance DEM. The Table 5 shows the final result.

TABLE 3. Total samples per partition in DISFA+.

AU	Total	Train	Validation	Test
AU1	7275	5107	635	1533
AU2	6087	4276	539	1272
AU4	9276	6508	807	1961
AU5	7021	4897	640	1484
AU6	7531	5224	708	1599
AU9	3049	2134	264	651
AU12	7389	5129	694	1566
AU15	2677	1858	252	567
AU17	4090	2887	382	821
AU20	3005	2115	270	620
AU25	9333	6491	885	1957
AU26	5417	3803	483	1131

TABLE 4. Total samples per partition in DEM.

AU	Total	Train	Validation	Test
AU1	261	173	25	63
AU2	229	149	19	61
AU4	279	193	31	55
AU5	64	46	6	12
AU6	261	189	28	44
AU9	100	71	8	21
AU12	459	317	45	97
AU15	76	51	1	24
AU17	39	29	2	8
AU20	14	12	1	1
AU25	327	230	33	64
AU26	187	128	19	40

TABLE 5. Total samples per partition in DEM augmented.

AU	Total	Train	Validation	Test
AU1	361	250	36	75
AU2	309	215	28	66
AU4	379	262	31	86
AU5	104	68	8	28
AU6	261	189	30	50
AU9	140	104	8	28
AU12	459	325	46	88
AU15	136	92	15	29
AU17	429	287	39	103
AU20	154	111	11	32
AU25	397	287	36	74
AU26	187	125	18	44

D. METHODOLOGY

Figure 4 illustrates the methodology followed for our experiments. Note that we have assigned a letter nomenclature to name each ConvNet to facilitate result explanations.The steps we followed are as follows:

1) We used the training and validation sets to realize individual training sessions for each ConvNet to observe the behavior of each architecture.
2) With the trained architectures, we used the test set to calculate evaluation metrics and measure the actual performance of the trained architectures.
3) Finally, using the same test set, we used GradCAM [37] and gradCAM++ [38] algorithms to realize a visual study of the previous results. Note that this calculation

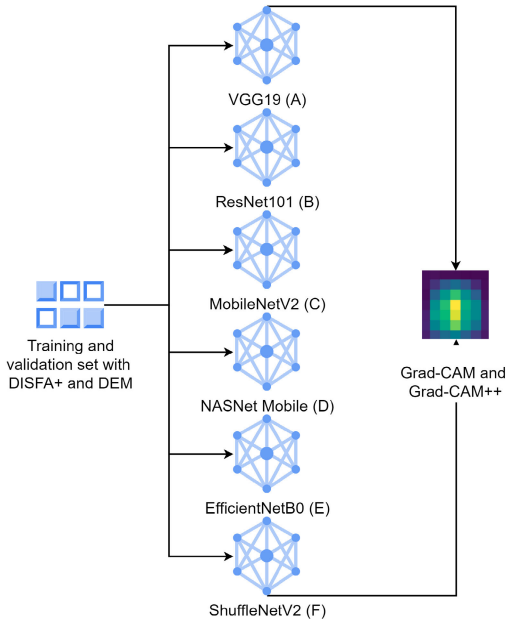

FIGURE 4. Proposed methodology.

TABLE 6. Hyperparameters.

Hyperparameter	Value	Network
Loss Function	Adam	All
Training epochs	50	VGG19, NASNet Mobile, MobileNetV2, ResNet101, EfficientNetB0 and ShuffleNetV2
Batch Size	48	NASNet Mobile, MobileNetV2, ResNet101, EfficientNetB0, ShuffleNetV2
	24	VGG19
Learning rate	0.0001	NASNet Mobile, MobileNetV2, Resnet101, EfficientNetB0
	0.001	VGG19 and ShuffleNetV2

was only done for VGG19 and ShuffleNetV2 because they were the models with the best F1 values.

Although the two CAM algorithms are similar, Grad-CAM sometimes fails when the image contains too many classes in a single image and does not always display the entire region used. Grad-CAM++ addresses this issue by employing a more advanced backpropagation method, solving the problems encountered by Grad-CAM.

Table 6 displays the configuration of the hyperparameters used during our training sessions. These hyperparameters are crucial because each architecture needs to adjust these values differently to achieve optimal performance. Additionally, given the nature of each architecture, even with Fine Tuning, it was necessary to adjust parameters for both DEM and DISFA+.

We conducted a total of twelve experiments: six experiments using DISFA+ and six experiments using DEM. We decided to use pre-trained architectures, except for ShuffleNetV2, as we can leverage the layers' weights that are already fine-tuned for similar tasks. In our case, we utilized pre-trained weights from ImageNet, which, while lacking a specific class for faces or similar, does contain examples where a human face may appear; this decision allows us to freeze the entire model except for the last convolutional block for fine-tuning.

TABLE 7. F1 score per class in DISFA+ trained models.

Action Unit	A	B	C	D	E	F
AU1	0.9309	0.9185	0.9095	0.9307	0.9155	**0.9633**
AU2	0.8811	0.8590	0.8319	0.8822	0.8323	**0.9334**
AU4	0.9478	0.9119	0.9245	0.9633	0.9436	**0.9742**
AU5	0.9427	0.8750	0.7505	**0.9504**	0.8268	0.9300
AU6	**0.9516**	0.7608	0.8114	0.8964	0.8313	0.9122
AU9	0.9652	0.9039	0.9062	**0.9663**	0.8955	0.9644
AU12	0.9761	0.8947	0.8977	0.9565	0.8816	**0.9780**
AU15	0.8908	0.6957	0.8800	0.9395	0.8511	**0.8942**
AU17	0.8957	0.6930	0.74686	0.8516	0.7742	**0.9216**
AU20	0.6828	0.2411	0.6480	0.6602	0.7314	**0.9289**
AU25	0.9687	0.8675	0.9073	0.9840	0.9126	**0.9879**
AU26	0.8937	0.7907	0.8024	0.9077	0.8480	**0.9403**
Average	0.9106	0.7843	0.8347	0.8715	0.8537	**0.9440**

III. RESULTS

A. EVALUATION METRICS FOR DISFA+

Following the same order as in state-of-the-art works, we used the F1 score as the primary metric for evaluating these experiments because the original dataset is imbalanced, so metrics like accuracy could yield misleading values at first glance. In Table 7, the F1 score for each AU is presented for each trained ConvNet. Numerically, we can observe that ShuffleNet outperforms all other ConvNets, except in the classes AU5, AU6, and AU9, where VGG19 and EfficientNetB0 is better.

Furthermore, in Figures 5 and 6, the confusion matrix for each class of the models with the highest performance (VGG19 and ShuffleNetV2) is visualized. It is easy to observe how both models achieve similar values in all classes, correctly classifying most test images. A particular case is AU20, where there is a significant difference in the AU20 class between VGG19 and ShuffleNetV2, where ShuffleNetV2 achieves nearly perfect classification while VGG19 does not.

Interestingly, ShuffleNetV2, with 4.5 million trainable parameters, outperformed VGG19 in most classes, even though VGG19 has 20.2 million trainable parameters, this indicates the optimization demonstrated in creating the ShuffleNetV2 architecture.

1) VISUAL STUDY FOR DISFA+

Although ShuffleNetV2 obtained better results than VGG19 numerically, we decided to perform a visual study using the Grad-CAM and Grad-CAM++ algorithms to examine the features that are taken into account by these ConvNets to predict the AUs; we decided because, on many occasions, when opting for purely numerical analysis, we can not know which areas of an image the ConvNets used to make

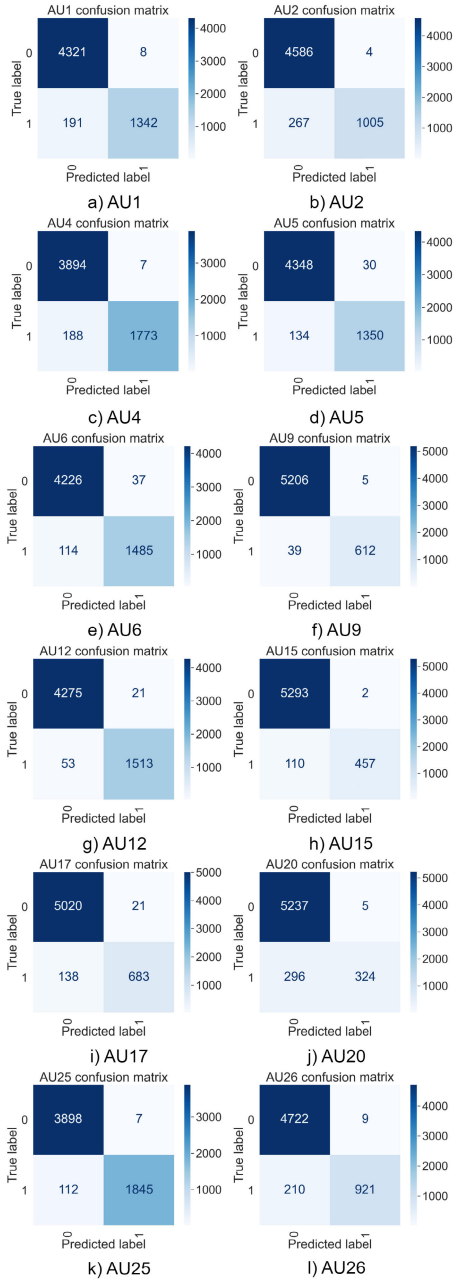

FIGURE 5. Confusion matrix of VGG19 trained with DISFA+.

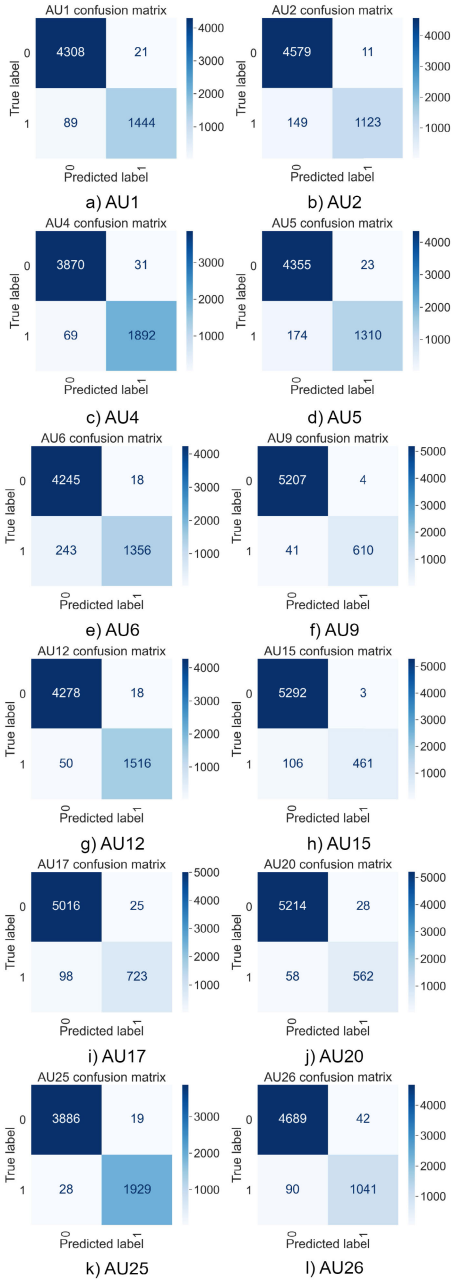

FIGURE 6. Confusion matrix of ShuffleNetV2 trained with DISFA+.

a prediction, falling into problems such as the areas not corresponding to the desired class.

Consider the image in Figure 7. The face on the left is AU4 (Brow Lowerer), and the face on the right is AU25 (Lips Part). In Figure 7a, we observed that the features taken by VGG19 for the prediction of AU4 take into consideration also part of the mouth (AU25); however, as can be noticed in the areas of higher intensity (red color), it still takes more features belonging to AU25, while AU4 differentiates it correctly and without overlapping with some other AU. On the other hand, in Figure 7b, the features of ShuffleNetV2 do not align with AU4, and in the case of AU25, it focuses more on the upper part of the face, considering only slightly some areas of AU25; this may be a sign that, although numerically, a UA is better identified visually, it may not correspond to that UA but to other features that appear in those movements.

To further validate these results, we use a DEM image, as shown in Figure 8, labeled AU4. Using an image that does not belong to DISFA+ may affect the network's performance but allows us to evaluate its knowledge generalization capabilities more broadly. In Figure 8a, note how VGG19 considers the AU4 region and other features such as nose and mouth; this situation could be explained as co-occurrence between AUs, an occurrence explained in the original DISFA+ article where the authors mention that often an AU obligatorily appears next to another one, however in Figure 8b

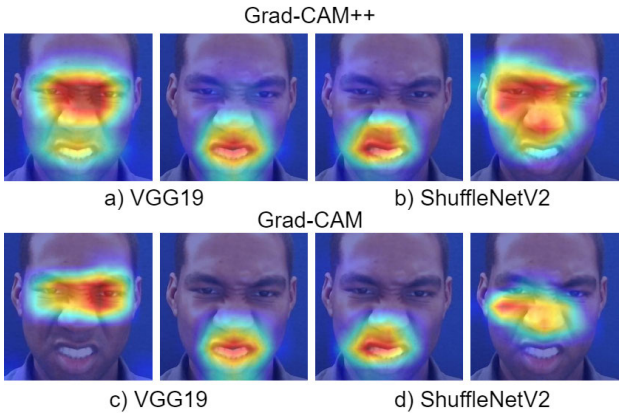

FIGURE 7. Grad-CAM++ and Grad-CAM using a DISFA+ image on VGG19 and ShuffleNetV2 trained with DISFA+.

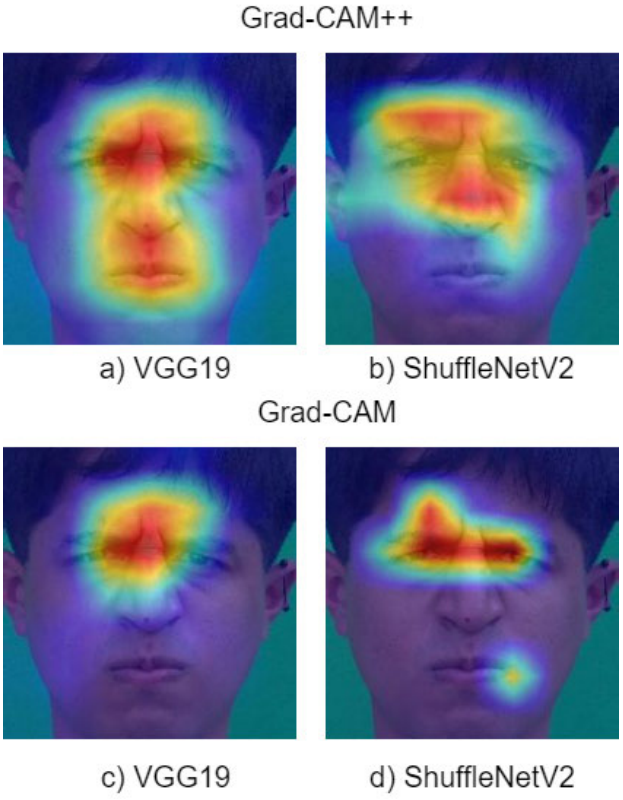

FIGURE 8. Grad-CAM++ and Grad-CAM using a external image on VGG19 and ShuffleNetV2 trained with DISFA+.

it fails in prediction. Now, considering only the Grad-CAM result shown in Figure 8c, it is evident that the area covered by the VGG19 features is considered a significant part of the upper region of the face; this area is on top of the AU4 area. Finally, in Figure 8d, a similar event is observed with ShuffleNetV2.

B. EVALUATION METRICS FOR DEM

Performing the same experiments as in the previous section with DEM, we obtained the following results (see Table 8). In this case, the architecture that achieved the best values

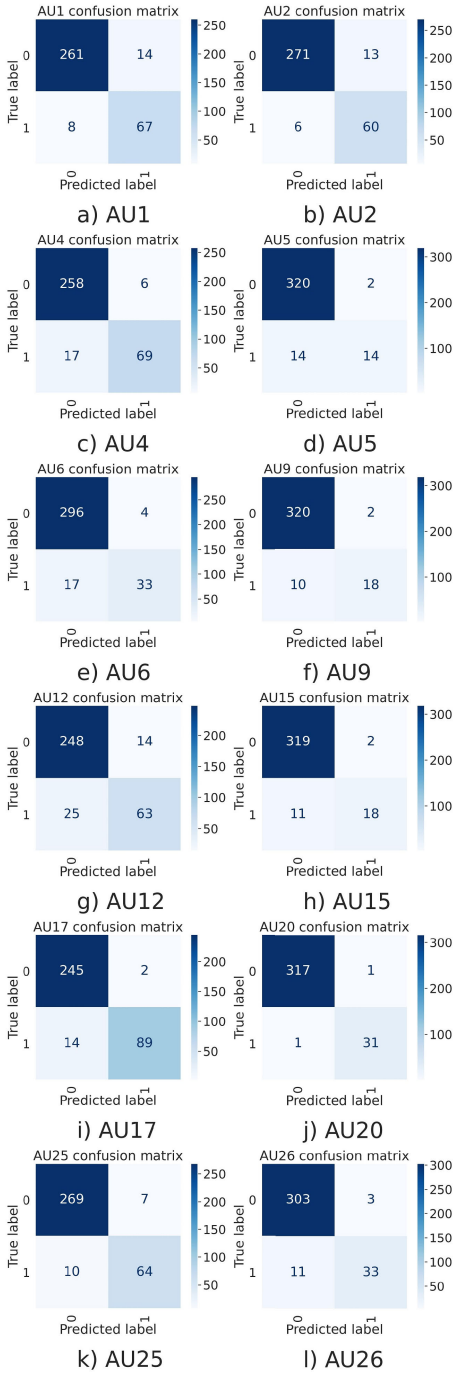

FIGURE 9. Confusion matrix of VGG19 trained with DEM.

was VGG19, followed by ShuffleNetV2, which performed better for AU17. Details of these results can be observed in Figures 9 - 10, where confusion matrices for each AU are displayed.

Regarding the DEM results, we can observe that despite some ConvNets managing to generalize knowledge, many others did not perform as well as with DISFA+; this could be attributed to the data available for training and validation before testing. Even though data augmentation was applied, and transfer learning with fine-tuning was performed, it is

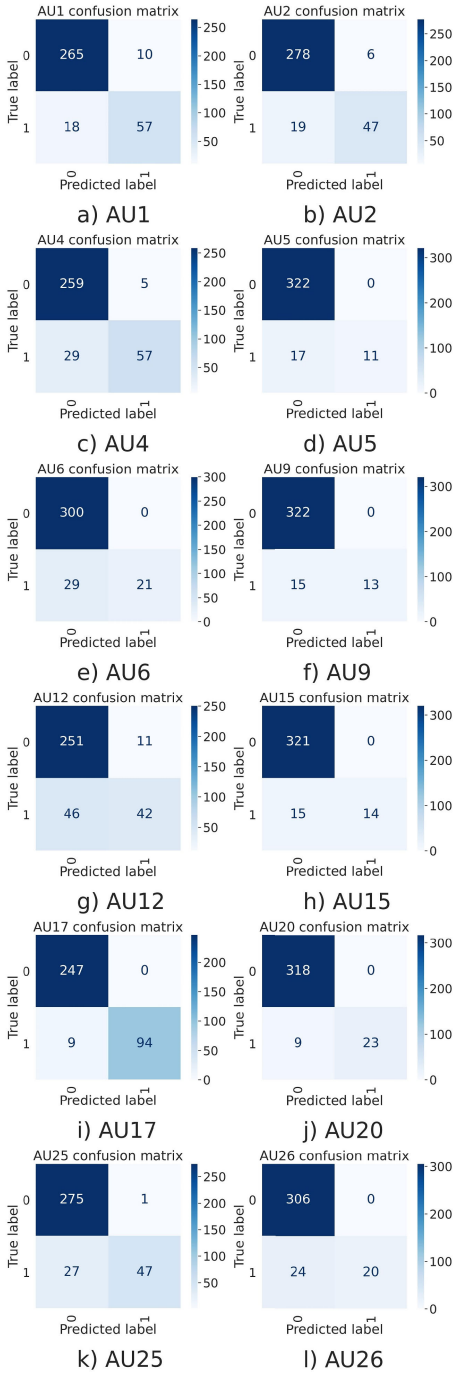

FIGURE 10. Confusion matrix of ShuffleNetV2 trained with DEM.

interesting to note that modern networks like NASNet Mobile did not generalize features well, a problem that ShuffleNetV2 could address, along with an older ConvNet like VGG19.

1) VISUAL STUDY FOR DEM

Repeating the procedure with DEM, we first used the Grad-CAM and Grad-CAM++ algorithms for the best-evaluated models (VGG19 and ShuffleNetV2) with images from the datasets. These results can be observed in Figure 11, where we can appreciate how in the case of

TABLE 8. F1 score per class in DEM trained models.

Action Unit	A	B	C	D	E	F
AU1	**0.8589**	0.6562	0	0.5966	0.1927	0.8028
AU2	**0.8633**	0.6727	0	0.7133	0.19178	0.7899
AU4	**0.8571**	0.7123	0.1276	0.0229	0	0.7702
AU5	**0.6363**	0.3684	0	0	0	0.5641
AU6	**0.7586**	0.5217	0.3278	0.5294	0.0769	0.5915
AU9	**0.75**	0.3030	0	0	0	0.6341
AU12	**0.7636**	0.5161	0.3898	0.3119	0.0444	0.5957
AU15	**0.7346**	0.3428	0	0.0666	0.1290	0.6511
AU17	0.9175	0.8808	0.6167	0.2393	0.6395	**0.9543**
AU20	**0.9687**	0.4761	0	0.4761	0	0.8363
AU25	**0.8827**	0.5636	0.1728	0.5283	0.0526	0.7704
AU26	**0.825**	0.5483	0	0.6021	0.0444	0.625
Average	**0.8180**	0.5468	0.1362	0.3405	0.1142	0.7154

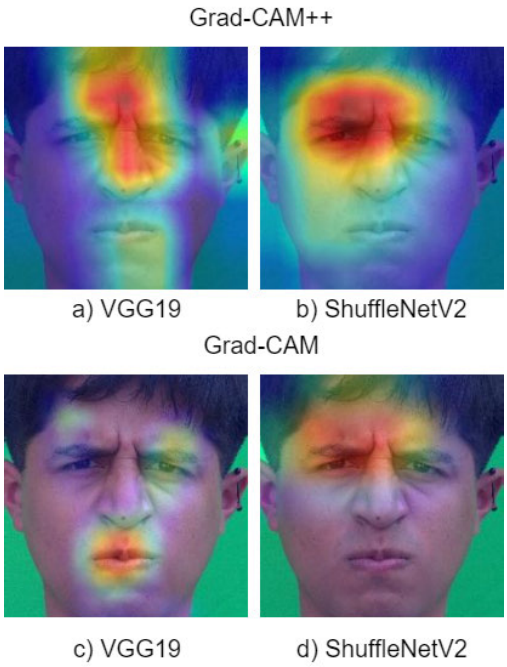

FIGURE 11. Grad-CAM++ and Grad-CAM using a DEM image on VGG19 and ShuffleNetV2 trained with DEM.

ShuffleNetV2 (Figures 11b-11d), AU4 does turn out to be the selected area by the ConvNet. However, in the case of VGG19 in Figure 11c, we observe that this Network does not use the AU4 region for it is prediction when in Figure 11d, we notice that while highlighting areas that do not belong to AU4, mainly considers the region of AU4.

On the other hand, using an image that does not belong to DISFA+, the results can be seen in Figure 12. In this figure, the AUs are part of the highlighted areas in both CAM algorithms. However, let us look at Figure 12a; we can observe AU25 in it. In the case of VGG19, it also captured parts that do not correspond to AU25 using the GradCAM++ algorithm. While using the GradCAM algorithm, this is correctly differentiated. In the case of ShuffleNetV2, we can observe that the ConvNets are more general than VGG19, considering that they cover much larger areas. However, there is still a distinction between those selecting AU4 and those selecting AU25.

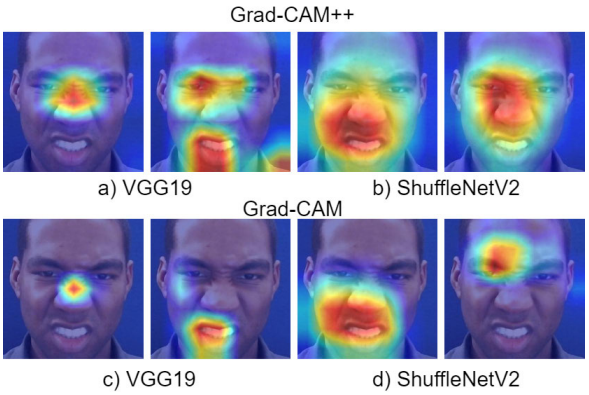

FIGURE 12. Grad-CAM++ and Grad-CAM using a external image on VGG19 and ShuffleNetV2 trained with DEM.

TABLE 9. F1 score comparison of our best models against the state of the art.

Action Unit	**DISFA+ Ours**	**DEM Ours**	DISFA [15]	DISFA [12]	DISFA [16]	DISFA+ [9]	DISFA+ [11]
AU1	**0.9633**	**0.8589**	0.461	0.457	0.437	0.72	0.764
AU2	**0.9334**	**0.8633**	0.486	0.478	0.462	0.68	0.723
AU4	**0.9742**	**0.8571**	0.728	0.596	0.560	0.68	0.835
AU5	**0.9300**	**0.6363**	-	-	-	0.67	-
AU6	**0.9122**	**0.7586**	0.567	0.471	0.414	0.73	0.772
AU9	**0.9644**	**0.75**	0.500	0.456	0.447	0.76	0.832
AU12	**0.9780**	**0.7636**	0.721	0.735	0.696	0.76	0.876
AU15	**0.8942**	**0.7346**	-	-	-	0.56	-
AU17	**0.9216**	**0.9175**	-	-	-	0.54	-
AU20	**0.9289**	**0.9687**	-	-	-	0.37	-
AU25	**0.9879**	**0.8827**	0.908	0.843	0.883	0.88	0.946
AU26	**0.9403**	**0.825**	0.554	0.436	0.584	0.75	0.743

C. STATE OF THE ART COMPARISION

Table 9 compares the metrics of our best models for both datasets.

This study includes state-of-the-art works that used DISFA or DISFA+. Although, in some cases, the authors do not report F1 values for specific AUs, all the considered works were conducted under similar conditions, making it feasible to establish a comparison. It is essential to highlight that we only included articles that considered the presence of the AU when its intensity is equal to or greater than two.

In this comparison, we only report our two best F1 score results. The "DISFA+ Ours" column was obtained with ShuffleNetV2, with an average value of 0.94%, and the "DEM Ours" column with VGG19, with an average value of 0.82%.

IV. DISCUSSION AND CONCLUSION

Despite being a classical ConvNet, the results of this work show that VGG19 remains a very efficient architecture for solving complex Computer Vision problems, in this case, the detection of AUs in the human face. When studying state-of-the-art works, we encountered the AU intensity label; however, for AU detection, we found that this value is not essential for the classification task. Binarizing the problem yielded suitable F1 values.

Regarding the datasets, while it is true that in Table 9 DISFA+ achieves higher F1 scores, this can be explained by the quantity and quality of the data. DISFA+ surpasses DISFA in the quality of captured images, not only in resolution but also in the equipment used and the scenario in which the photographs were taken (with better lighting conditions in DISFA+, for example).

While with DEM, we can observe that although its mean F1 value is lower than that obtained using DISFA+ in the experiments conducted, it is essential to consider that DISFA+ has only nine individuals. In contrast, DEM has 246 participants, which, because of the number of participants, adds more variability to the problem, and the distribution of characteristics changes significantly, so that for work that seeks to study these changes in the way individuals express the same AU it may be more valuable to have more test subjects. In addition, having such a diverse set of individuals in DEM enriches the dataset and, at the same time, makes it more challenging to model.

That said, we can affirm that DEM is a valuable dataset for detecting Action Units in human faces, in addition to providing labels on emotions, which may result in more complex work on the study of AUs and emotions alike, since as is known, a set of AUs can visually express an emotion on the face.

Using pre-trained ConvNets for transfer learning followed by fine-tuning was highly beneficial, as it reduced training time while achieving satisfactory F1 values for all the AUs to be classified; this is because pre-trained models have weights tuned for similar classification tasks, providing a good starting point for the specific classification problem addressed in this work.

Another exciting aspect is that we observed that models with few adjustable parameters, approximately 5 million, such as ShuffleNetV2, achieved similar or even better results in terms of F1 compared to VGG19, which has over 130 million parameters.

The use of gradient visualization techniques, Grad-CAM and Grad-CAM++, proved to be highly beneficial for examining the disparities in the selected features between ShuffleNetV2 and VGG19 (the two models that achieved the best F1 scores). This includes the features considered by both models for AU prediction, even in some cases taking into account features that do not necessarily represent a true AU. For example, selecting features from upper regions (part of the eyes and forehead) and lower regions (such as the chin and mouth) to represent a single AU. However, in general, these techniques demonstrated that the AUs are well-represented in the majority of cases. However, we find examples such as those presented in Figures 7a and 7c where substantial differences are evident when Grad-CAM++ is employed. Although it is clear that the model selects features that visually correspond to AUs and are correctly labeled in the image, there are cases where the features taken into account have a larger area than a typical AU. Another analogous case is observed in Figure 8a compared to Figure 8b, where a

discrepancy in the features considered by both visualization methods is apparent. Another similar scenario involves DEM in Figure 12a, where, using Grad-CAM++, certain features of the mouth are included that were not considered when using only Grad-CAM. The model incorporating features beyond those expected for predicting an outcome is standard. For AUs, the original DISFA+ article said that the appearance of one action unit may be linked to another simultaneously, which, according to our experiments, may cause a ConvNet to predict one AU based on the appearance of another.

When comparing results, we can observe that the ShuffleNetV2 model generalized the AUs better with DISFA+ than with DEM; this may be due to the number of images each data set has, DISFA+ being the largest. So, the more examples a model has to train, the better its generalization performance will be for this specific model. Since we did not use a pre-trained ShuffleNetV2 model, the number of images played an essential role during training, and the results were better with the more extensive data set.

On the other hand, we have the case of the VGG19 model, which generalized the AUs better with DEM than with DISFA+. In this case, we used a pre-trained VGG19 model for these experiments. However, the number of images is essential; their variability is more important, with DEM being the data set with the most significant amount of variability, which benefited the model to generalize better by having better (and more diversified) examples of the same case (AUs).

In this sense, the choice of architecture will depend entirely on what data and how much data is available. The deeper a network is, the more data it will need to generalize knowledge, but it will do better, which is why training a model from scratch is ideal if there is a large and coarse data set. On the other hand, if only have a small or medium data set, it is better to opt for pre-trained models (as in this example, VGG19), which will use the previously adjusted parameters to now generalize the new characteristics, being influential in both cases the variability of the data presented during the training phase. This work can be valuable in various scientific areas where the AUs can play a significant role. For example, the manual implementation of the FACS (without the use of machine learning) has been studied in clinical settings to detect pain in older adults with and without Alzheimer's disease [39]. In this experiment, the authors used 27 individuals diagnosed with Alzheimer's and 36 volunteers without cognitive issues. Considering a pain threshold obtained through the multiple random staircase technique, the participants' electric and mechanical stimuli were applied to record their facial reactions at each pain threshold. Subsequently, a FACS expert classified these reactions. The authors concluded that FACS can detect pain in individuals who cannot report their sensations. In such experiments, deep learning techniques like these would assist an expert.

Another work that already utilizes machine learning techniques, such as Random Forest, is found in [40]. In this study, the authors aimed to design an automatic AU classifier to detect stress in individuals through facial videos recorded. At the same time, people type or rest in front of a computer. The results of this work showed an effectiveness of 74% in subject-independent classification, reinforcing the theory that the proper use of deep learning techniques can improve even the results obtained by works applying classical machine learning classifiers.

The use of Deep Learning is functional for AU classification, and in turn, the use of AUs is valuable in different areas of knowledge, such as medicine. The next step in this work is to investigate AUs within the context of human emotions. Leveraging the existence of the DEM dataset, this could be useful for implementing emotion classifiers based solely on AUs. However, there is still a broad path to explore, where, in addition to implementing more modern ConvNet architectures, we can consider the actual intensities of each AU to add variables such as the sequence of occurrence.

ACKNOWLEDGMENT

The authors sincerely thank Carlos J. Morales Hernandez and Marco A. Ramirez Hidalgo for collaborating to create DEM.

REFERENCES

[1] P. Ekman and W. Friesen, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.*, Jan. 1978.
[2] P. Gosselin, M. Perron, and M. Beaupré, "The voluntary control of facial action units in adults," *Emotion*, vol. 10, no. 2, pp. 266–271, 2010.
[3] B. M. Waller, S.-J. Vick, L. A. Parr, K. A. Bard, M. C. S. Pasqualini, K. M. Gothard, and A. J. Fuglevand, "Intramuscular electrical stimulation of facial muscles in humans and chimpanzees: Duchenne revisited and extended," *Emotion*, vol. 6, no. 3, pp. 367–382, Aug. 2006.
[4] S. Polikovsky, Y. Kameda, and Y. Ohta, "Detection and measurement of facial micro-expression characteristics for psychological analysis," *Kameda's Publication*, vol. 110, pp. 57–64, Jun. 2010.
[5] 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.
[6] R. Zhi, M. Liu, and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *Vis. Comput.*, vol. 36, no. 5, pp. 1067–1093, May 2020.
[7] A. Calder, J. Keane, J. Cole, R. Campbell, and A. Young, "Facial expression recognition by people with Möbius syndrome," *Cognit. Neuropsychol.*, vol. 17, pp. 73–87, Sep. 2000.
[8] X. Ge, J. M. Jose, P. Wang, A. Iyer, X. Liu, and H. Han, "Automatic facial paralysis estimation with facial action units," 2022, *arXiv:2203.01800*.
[9] S. Nerella, K. Khezeli, A. Davidson, P. Tighe, A. Bihorac, and P. Rashidi, "End-to-end machine learning framework for facial AU detection in intensive care units," 2022, *arXiv:2211.06570*.
[10] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended DISFA dataset: Investigating posed and spontaneous facial expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 1–8.
[11] Y. Wei, H. Wang, M. Sun, and J. Liu, "Attention based relation network for facial action units recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[12] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8594–8601, doi: 10.1609/aaai.v33i01.33018594.
[13] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, pp. 1–6, Apr. 2013, doi: 10.1109/FG.2013.6553788.

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/T-AFFC.2013.4.

[15] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7680–7689.

[16] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 705–720.

[17] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001, doi: 10.1109/34.908962.

[18] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7, doi: 10.1109/FG.2013.6553799.

[19] M. Valstar and M. Pantic, "Induced disgust, happiness, and surprise: An addition to the MMI facial expression database," in *Proc. 3rd Int. Workshop EMOTION (Satell. LREC), Corpora Res. Emotion Affect*, vol. 10, 2010.

[20] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. 1st Eur. Workshop Biometrics Identity Manag.*, 2008, pp. 47–56.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern s (CVPR)*, Jun. 2018, pp. 8697–8710.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[26] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[27] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, Apr. 2021, doi: 10.3390/electronics10091036.

[28] A. Singh and D. Kumar, "Detection of stress, anxiety and depression (SAD) in video surveillance using ResNet-101," *Microprocessors Microsyst.*, vol. 95, Nov. 2022, Art. no. 104681, doi: 10.1016/j.micpro.2022.104681.

[29] F. Saxen, P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with MobileNetV2 and NasNet-mobile," in *Proc. 11th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2019, pp. 176–180, doi: 10.1109/ISPA.2019.8868585.

[30] F. Wang, R. Zheng, P. Li, H. Song, D. Du, and J. Sun, "Face recognition on raspberry Pi based on MobileNetV2," in *Proc. Int. Symp. Artif. Intell. Appl. Media (ISAIAM)*, May 2021, pp. 116–120, doi: 10.1109/ISAIAM53259.2021.00031.

[31] P. Metgud, N. D. Naik, S. M. Sukrutha, and A. S. Prasad, "Real-time student emotion and performance analysis," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2022, pp. 1–5, doi: 10.1109/CONECCT55679.2022.9865114.

[32] A. Ghofrani, R. M. Toroghi, and S. Ghanbari, "Realtime face-detection and emotion recognition using MTCNN and miniShuffleNet v2," in *Proc. 5th Conf. Knowl. Based Eng. Innov. (KBEI)*, Feb. 2019, pp. 817–821, doi: 10.1109/KBEI.2019.8734924.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[34] 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," 2017, *arXiv:1704.04861*.

[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[36] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[38] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[39] A. C. Lints-Martindale, T. Hadjistavropoulos, B. Barber, and S. J. Gibson, "A psychophysical investigation of the facial action coding system as an index of pain variability among older adults with and without Alzheimer's disease," *Pain Med.*, vol. 8, no. 8, pp. 678–689, Nov. 2007, doi: 10.1111/j.1526-4637.2007.00358.x.

[40] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, "Towards independent stress detection: A dependent model using facial action units," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6, doi: 10.1109/CBMI.2018.8516497.

MARCO A. MORENO-ARMENDÁRIZ received the B.S. degree from Universidad La Salle, Mexico, in 1998, and the M.S. and Ph.D. degrees in automatic control from Centro de Investigación y Estudios Avanzados (CINVESTAV) del Instituto Politécnico Nacional (IPN), Mexico, in 1999 and 2003, respectively.

From 2001 to 2006, he was a Researcher with Escuela de Ingeniería, Universidad La Salle. In April 2006, he joined Centro de Investigación en Computación (CIC), IPN. His current research interests include deep learning, machine learning, and cognitive sciences. He is the author of more than 80 publications in his research area. He is a member of the National System of Researchers, Level II.

ALBERTO ESPINOSA-JUAREZ received the bachelor's degree in computer engineering from Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), in 2020, and the master's degree (cum laude) in computer science from the Computational Cognitive Science Laboratory, Centro de Investigación en Computación, in 2023.

He is currently working on the development and research of computer vision systems. He is the author and coauthor of two publications related to deep learning. His current research interests include deep learning, focused on deep neural networks (convolutional and currently visual transformers).

ESMERALDA GODINEZ-MONTERO is currently pursuing the degree in computer systems engineering with Escuela Superior de Cómputo (ESCOM IPN).

In 2019, she graduated as an IT Technician at CECyT 13 "Ricardo Flores Magón." Since 2022, she has collaborated with the ESCOM Algorithm Club sharing knowledge on data structures and algorithms with fellow students. Also, she has been involved in projects related to facial emotion recognition using convolutional neural networks. Her current research interests include deep neural networks and attention mechanisms.