IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 February 2024, accepted 4 March 2024, date of publication 21 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380439

RESEARCH ARTICLE

Exploring the Impact of Partial Occlusion on Emotion Classification From Facial Expressions: A Comparative Study of XR Headsets and Face Masks

ALBERTO CASAS-ORTIZ[1,2], **JON ECHEVERRIA**[1,2], **NEREA JIMENEZ-TELLEZ**[3], **AND OLGA C. SANTOS**[1,2]

[1]Artificial Intelligence Department, Computer Science School, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
[2]PhyUM Research Center, UNED, 28040 Madrid, Spain
[3]Stanford Cardiovascular Institute, Stanford University School of Medicine, Palo Alto, CA 94304, USA

Corresponding author: Alberto Casas-Ortiz (alberto.casasortiz@dia.uned.es)

This work was supported in part by the Project ‘‘HumanAID-Sens: HUMan-Centered Assisted Intelligent Dynamic systems with SENSing Technologies’’ funded by the Spanish Ministry of Science and Innovation through Ministerio de Ciencia e Innovación (MCIN)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033 under Grant TED2021-129485B-C1, and in part by the European Union ‘‘NextGenerationEU’’/Plan de Recuperación, Transformación y Resiliencia (PRTR).

ABSTRACT This study provides of a comparative analysis of emotion estimation from facial expressions under partial occlusion caused by face masks and extended reality (XR) headsets. Unlike previous studies that have independently explored these two scenarios, this research compares and analyzes the statistical differences between them. In order to achieve this, the RAF-DB dataset has been used as a non-occluded baseline to construct two new datasets: i) a dataset formed by faces partially occluded by face masks, and ii) a dataset formed by faces partially occluded by XR headsets. To evaluate the impact of occlusion in emotion estimation, three deep learning models have been fine-tuned using transfer learning, and results from a random classifier have been used as a baseline. Seven different metrics were obtained per dataset, and a 2-way ANOVA test was performed on each metric. As expected, significant statistical differences are observed between the non-occluded faces (acc. 0.8780) and the faces partially occluded by face masks (acc. 0.7520) and XR headsets (acc. 0.7400) on all metrics. Notably, the comparison between the two partially occluded datasets revealed significant statistical differences in the metrics f1-score (macro), precision (macro) and recall (macro), which we attribute to different types of occlusion affecting different parts of the face that are key to some emotions. This research contributes to advancing emotion recognition systems by highlighting their robustness and effectiveness even in partial occlusion settings, and showing a full comparative analysis between two common types of occlusion.

INDEX TERMS Emotion classification, emotion recognition, facial expression analysis, partial occlusion, transfer learning, deep learning, extended reality, face masks, HMD, XR headset.

I. INTRODUCTION

Facial expressions are an incredibly complex and dynamic communication tool capable of expressing a wide range of emotions [1], [2]. Researchers have long been interested in accurately estimate emotions from facial expressions, as it has implications in various fields such as neuroscience, human-computer interaction, marketing, and psychomotor learning [3], [4], [5], [6], [7], [8].

In some scenarios, faces can be partially occluded, posing an increased challenge for facial emotion recognition algorithms. To overcome this, researchers can rely on alternative

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

approaches that do not depend only on facial expressions. These alternatives include estimation of emotions from body poses, variations in voice, physiological signals, and employing multimodal approaches combining different methods [9], [10], [11], [12], [13], [14]. Nevertheless, valuable information may still be found in the visible regions of the face unaffected by occlusion. This presents an opportunity for existing facial emotion recognition algorithms.

There are two scenarios of special interest in facial emotion recognition with partial occlusion. The first scenario is the use of extended reality (XR) headsets [15]. This scenario is interesting because of the growth of the XR industry [16], with available commercial devices like Meta Quest 2, 3 and Pro, Apple Vision Pro, or HTC Vive devices. It has application in fields like healthcare [17], [18], [19], [20], education and training [7], [17], [18], [21], [22], and gamification [17], [22], [23]. The second scenario is the use of face masks, which increased globally during the COVID-19 pandemic [24], [25], [26]. The use of face masks has been demonstrated to affect emotion recognition in humans [27], [28], [29], specially in individuals with conditions like autism spectrum disorder [30], [31], [32], as well as to affect performance on facial expression classification algorithms [27].

Both scenarios are complement each other: XR headsets occlude the upper part of the face (periocular area), while face masks occlude the lower part (orofacial region). However, existing literature lacks an exploration of deep learning algorithms performance differences between these two types of occlusion (Section II). Additionally, the intrinsic bias in emotion recognition datasets [33] is usually not adequately addressed, and the metrics used to compare the models are not sufficient. Moreover, techniques used to synthetically occlude the faces may inadvertently leave crucial areas of the face un-occluded, resulting in information leakage.

In this study, we hypothesize that comparing facial expressions without occlusion to those with partial occlusion caused by face masks or XR headsets will reveal significant statistical differences. However, we anticipate that there may not be significant statistical differences when comparing occlusion by face masks with occlusion by XR headsets. To test this hypothesis, three deep learning models are trained across three datasets: i) non-occluded faces, ii) faces partially occluded by face masks, and iii) faces partially occluded by XR headsets. The imbalance of the datasets is addressed by using a generative model for data augmentation, ensuring generated instances are sufficiently different to instances from the dataset. The synthetic occlusion was introduced considering the dimensions and inclination of the face to ensure there is no information leakage. Different metrics suitable for handling unbalanced datasets have then been selected to evaluate the performance of the models. These metrics are: accuracy, precision (macro and weighted), recall (macro and weighted), and f1-score (macro and weighted). Finally, a 2-way ANOVA test is employed to statistically compare the different groups of results and assessing the statistical differences.

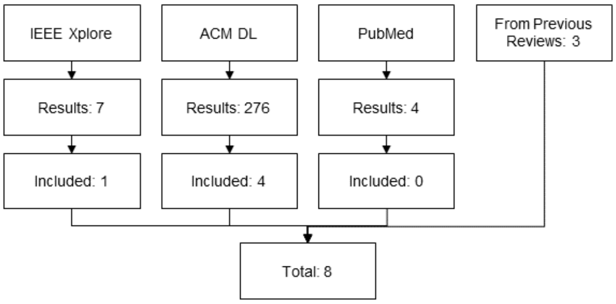

FIGURE 1. Workflow followed to perform the systematic review.

This paper is organized as follows. Section II reviews the related work in the field to provide context and background for our research. Section III outlines the materials and methods used, as well as the process and reasons followed to select them. Section IV presents our results. Section V discusses the implications of our findings, identifying the limitations of the study and suggesting directions for future research. Finally, Section VI shows the conclusions of this paper.

II. RELATED WORK

Bibliography on facial emotion classification involving partial occlusion caused by face masks and XR headsets using deep learning techniques is limited. This may be attributed to the availability of alternatives like the use of physiological sensors or voice analysis. To identify studies addressing this issue, we conducted a systematic literature review. Inclusion criteria for the systematic review were: i) the occlusion must be similar to that introduced either by face masks or XR headsets, ii) the study must use deep learning techniques, and iii) the study must try to recognize/classify basic emotions (e.g., happiness, sadness, contempt). Three databases (IEEE Xplore, ACM DL and PubMed) were selected, and the following search query was created: ("emotion") AND ("occlusion") AND ("face mask" OR "headset" OR "HMD"). This search query was executed on the search engines of the databases. We identified 5 papers, as well as two previous reviews on estimation of emotions with partial occlusion caused by the use of XR headsets [34], [35] from where we extracted three more papers. In total, our review includes 8 papers. The review workflow is illustrated in Fig. 1.

We conducted a depth analysis of the eight papers, examining aspects such as type of occlusion (periocular or orofacial), deep learning models used, datasets utilized, emotions assessed, or whether the occlusion employed was real or synthetic. The subsequent paragraphs provide a detailed examination of each paper. Table 1 presents a summary of the findings of our systematic review.

In Houshmand and Khan [36], two CNN architectures named VGG and ResNet are used to evaluate facial emotion classification in faces partially occluded by XR headsets. The datasets assessed were AffectNet, RAF-DB and FER+, with occlusion synthetically introduced by placing a black

TABLE 1. Information extracted from our systematic review.

Paper	Algorithms	Datasets	Emotions	Occlusion type	Real Occlusion	Synthetic occlusion	Bias correction	Transfer Learning	Metrics
Houshmand and Khan, 2020 [36]	VGG, ResNet	Affectnet, RAF-DB, FER+	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality, Comptempt	Periocular	No	Periocular Black patch	Data augmentation (Horizontal Flipping)	Yes	Accuracy
Yong et al., 2019 [37]	RestNet, Inception-RestNet-V2, DenseNet	RafD	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality, Comptempt	Periocular	No	Periocular Black patch	Not necessary	Yes	Precision, Recall, Accuracy
Georgescu and Ionescu [38]	VGG	FER+, AffectNet	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality, Comptempt	Periocular	No	Not Specified	Data augmentation (Horizontal Flipping)	Yes	Accuracy
Petrou et al., 2023 [39]	mini-Xception	Custom dataset	Happy, Sad, Neutral	Periocular	No	Periocular Black patch	Removed underrepresented classes	Yes	Accuracy, f1-score (macro and weighted)
Abate et al., 2023 [27]	Custom (CNNs), Residual Masking Networks, Amending Representation Modules	RAF-DB, FER2013	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality	Periocular, Orofacial	No	Periocular Black patch, Overlayed face mask	None	No	Accuracy, Precision, Recall, f1-score
Luo et al., 2022 [24]	ResNet, MobileNet	VIP-DB, RAF-DB, LFW	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality	Orofacial	Yes (VIP-DB)	Overlay Face Mask (RAF-DB, LFW)	Data augmentation (Horizontal Flipping)	No	Accuracy
Sola and Gera [25]	ResNet	MSD-E, MSD-PE, RAF-DB (only cross-dataset)	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality	Orofacial	Yes (MSD-E and MSD-PE)	Overlay Face Mask (RAF-DB)	Not necessary	No	Accuracy
Castellano et al., 2021 [26]	VGG	FER2013	Happiness, Sadness, Anger, Surprise, Fear, Disgust, Neutrality	Orofacial	No	Cropped eyes	None	Yes	Accuracy, Precision, Recall, f1-score

patch over the periocular area. To address dataset imbalance, horizontal image flipping is applied for data augmentation. The target emotions assessed were happiness, sadness, anger, surprise, fear, disgust, neutrality and contempt (contempt is only available on AffectNet and FER+). The models were trained both from scratch and using transfer learning on models pretrained on the VGG-Face dataset. Optimal results were achieved through transfer learning on the

RAF-DB and the FER+ datasets. However, the study has limitations, as relying only on the accuracy metric for result interpretation. This can potentially lead to bias in imbalanced datasets where the models can perform better on majority classes than on minority classes due to more information available. Additionally, the horizontal image flipping augmentation technique may introduce redundancy due to facial symmetry. It is notable how, even when optimal results are obtained on partially occluded faces demonstrating it is possible to do facial emotion recognition on them, this study does not compare results obtained on the partially occluded dataset against those on the original non-occluded dataset, lacking of a baseline for assessing the impact of occlusion in the results obtained.

In Yong et al. [37] three CNN architectures (ResNet, Inception-ResNet-V2 and DenseNet), are employed to evaluate facial emotion classification in faces partially occluded by XR headsets. The dataset used is Radbound Faces Dataset (RafD), a balanced dataset formed by images created in a laboratory setting, where participants pose the facial expressions. Occlusion was synthetically introduced by placing a black patch over the periocular area. The emotions assessed were happiness, sadness, anger, surprise, fear, disgust, contempt and neutrality. Transfer learning is used to train the models pretrained on the ImageNet dataset. The study reports optimal results with all three models, and the interpretation of the results relies on precision, recall and accuracy metrics for each individual class. Class activation maps are presented to illustrate how the CNNs focus attention on the orofacial region. A notable limitation of this study is that the dataset used is created in a controlled laboratory environment, potentially limiting its applications to real settings where different conditions such as lightning, image quality, and background noise may impact the model performance.

In Georgescu and Ionescu [38], two CNN architectures (VGG-f and VGG-face), are used to evaluate facial emotion classification in faces partially occluded by XR headsets. The datasets employed are FER+ and AffectNet, with synthetic occlusion introduced by a non-specified method. Horizontal image flipping is applied for data augmentation to address dataset imbalance. In the case of AffectNet, downsampling is also applied. The emotions assessed were happiness, sadness, anger, surprise, fear, disgust, contempt and neutrality. The models were training by joining the training sets of the AffectNet and FER+ datasets, and then testing on the separate test sets of both datasets. Transfer learning is employed, initially training the models on non-occluded faces and later on partially occluded faces. Both models were then tested on partially occluded faces. Favorable results are reported for models trained on partially occluded faces, and a comparison of results between partially occluded and non-occluded faces is presented. However, the interpretation relies only on the accuracy metric, potentially introducing bias in result analysis. The data augmentation technique, once again, involves flipping images, which may lead to redundancy. Furthermore, downsampling the majority classes is implemented to mitigate imbalance, but this approach could result in information loss for the majority classes.

In Petrou et al. [39], a lightweight and fast CNN architecture named mini-Xception is used to estimate emotions from faces partially occluded by XR headsets. This model was selected because it is suitable for mobile devices. The dataset used is a custom dataset created from internet images, initially created for a previous project. To address class imbalance, all of the minority classes were removed, leaving only three emotions: happiness, sadness and neutral. Synthetic occlusion was applied by overlying a black patch on the periocular area. Different transfer learning techniques were tested, revealing optimal results when updating all of the weights of the models instead of freezing layers. The evaluation metrics used were accuracy and f1-score (macro and weighted), more suitable for interpreting results on imbalanced datasets. A notable advantage of this study is the adaptability of the mini-Xception model to small devices. However, a drawback of the study is the downsampling method, since removing the minority classes reduces the variety of emotions the models can estimate.

In Abate et al. [27], three methods were selected (Residual Masking Networks, CNNs and Amending Representation Modules), based on their performance on non-occluded datasets and their availability online. The goal was to assess facial classification in faces partially occluded by face masks, and occlusion in the periocular area (similar to XR headsets). The datasets used were RAF-DB and FER2013. Occlusion was synthetically added by overlying a face mask, and by placing a horizontal black patch over the periocular area. The emotions assessed were happiness, sadness, anger, surprise, fear, disgust, and neutrality. The models were trained from scratch without using transfer learning. Notably, good results were achieved when training on partially occluded faces. The study concluded that a model trained on non-occluded faces performs worse on partially occluded faces than a model trained on partially occluded faces, indicating that occlusion impacts facial emotion recognition algorithms and that the characteristics of the training set are important for the network to focus on target areas of the face. The evaluation metrics used were accuracy, precision, recall and f1-score, suitable for imbalanced datasets. The authors acknowledged that different occlusions may affect the recognition of specific emotions more than others [40]. A limitation is that the periocular occlusion cannot be compared to that introduced by a XR headset, as the horizontal bar might not occlude both eyes and the eyebrows when the face is tilted. Moreover, even when the study uses as a baseline the results of the original studies over the non-occluded datasets, there is no further comparison between the two types of occlusion, and no method to correct the imbalance of the datasets is mentioned.

In Luo et al. [24], two models (ResNet and MobileNet with contrastive representation learning) are trained to estimate emotions from faces partially occluded by face masks. Three datasets are used: RAF-DB, LFW and VIP-DB. VIP-DB is a custom dataset composed of real-world images obtained from videos of people using face masks, while LFW is a face verification dataset categorized into positive, negative and neutral emotion. The emotions available in RAF-DB and VIP-DB are happiness, sadness, anger, surprise, fear, disgust, and neutrality. Despite VIP-DB having real occlusion, the dataset can be consideted small. Synthetic occlusion was inserted into the RAF-DB and LFW datasets by overlying a face mask over the faces. To address class imbalance, data augmentation involving horizontally flipped images was applied. Results were evaluated using the accuracy metric, revealing optimal results with contrastive representation learning on masked LFW and RAF-DB. However, VIP-DB did not obtain good results, likely due to its small size. A notable advantage of the study is the incorporation of real-world masked face images, although the size of the dataset is limited. Again, relying only on the accuracy metric to assess the results presents a limitation given the imbalance on face emotion recognition datasets, and other metrics should be used, like f1-score. Additionally, the use of horizontally flipped images to address imbalance may introduce redundancy in the dataset.

In Sola and Gera [25], the ResNet model with contrastive learning and knowledge distillation is used to estimate emotions from faces partially occluded by face masks. Two datasets are evaluated: MSD-E, a posed dataset of face images with and without face masks created by the authors and publicly available, and MSD-PE, a similar dataset containing pairs of images of the same person with and without a face mask. Both datasets are annotated with emotions: happiness, sadness, anger, surprise, fear, disgust, and neutrality. MSD-E is not highly imbalanced, so bias correction is not applied. The study achieves optimal results on non-masked images, particularly for positive emotions (surprise, happy and neutral) using accuracy as evaluation metric. As expected, partially occluded images show lower accuracy. Visualization tools like Grad-CAM plots for attention and t-SNE for feature visualization are utilized. A cross-dataset study was conducted using RAF-DB, where synthetic occlusion was introduced by overlying a face mask. A model trained on Masked RAF-DB was tested on MSD-E and a model trained on non masked RAF-DB was tested on non-masked images of MSD-E. Similar results were obtained for neutral, surprise, sadness and anger, misclassifyng most emotions as neutral in the masked case. The authors conclude that synthetically occluded images may not generalize to real-world situations, emphasizing the need for real-world datasets. The small size of MSD-E and MSD-PE could be a limitation, along with the use of just the accuracy metric for evaluation.

In Castellano et al. [26], the model VGG16 pretrained on the ImageNet dataset is used to estimate emotions from faces wearing face masks. The FER2013 dataset is used for emotion recognition, and an unnamed dataset consisting of unmasked and synthetically masked faces is used to detect presence of a mask. The emotions assessed were happiness, sadness, anger, surprise, fear, disgust, and neutrality. Three networks were trained: i) to detect if a face is wearing or not a mask using the unnamed dataset, ii) to detect emotions on unmasked faces using FER2013, and iii) network to detect emotions on cropped images of the periocular area using FER2013. Since only the periocular area is used as input, no synthetic occlusion is needed. Instead, unwanted parts of the image are eliminated. No bias correction technique is applied in this study. The evaluation metrics used were accuracy, f1-score, precision and recall, and optimal results were obtained for positive emotions like happiness, neutrality or surprise. This study introduces the concept of cropping the images, providing networks with only the essential information for emotion recognition. However, this approach may potentially remove useful information, such as head position or forehead details, which could contribute to a more accurate emotion classification.

In the eight reviewed studies, the applications of deep learning techniques has been proved to be effective in achieving optimal results estimating emotions from faces partially occluded by face masks or XR headsets. Further, studies using transfer learning report a faster convergence and even better results. However, certain limitations are noted across some of the studies. These include: giving insufficient details regarding the training process and hyperparameters used, absence of metrics more useful on evaluating models trained on imbalanced datasets, potential introduction of redundancy through bias correction techniques, and lack of strategies to address dataset imbalance. Additionally, only one study has addressed both types of occlusion, but no comprehensive comparison is conducted. To address these gaps, our paper aims to conduct a statistical analysis, assessing differences between various evaluation metrics obtained from models trained on both types of occlusion. This approach seeks to provide a more comprehensive understanding of the performance and limitations of emotion classifiers in scenarios involving face masks and XR headsets.

This study, specifically, focuses on the analysis of statistical differences among evaluation metrics obtained from training three CNN architectures (VGG-16, ResNet-50 and SE-ResNet-50) using transfer learning. The selection of these architectures is based on their suitability for transfer learning and the availability of pre-trained models from the keras_vggface python library. Addressing the limitations observed in existing studies, our focus is on comparing the occlusion effects induced by XR headsets with those caused by face masks. To accomplish this, in this study we conduct a statistical analysis of evaluation metrics obtained after training our models on three datasets: i) non-occluded, ii) partially occluded by face masks, and iii) partially occluded by XR headsets. Additionally, we include the metrics obtained from a random classifier as a baseline. The workflow for our study is outlined in Fig. 2, providing a visual representation of the methodology followed in our analysis.

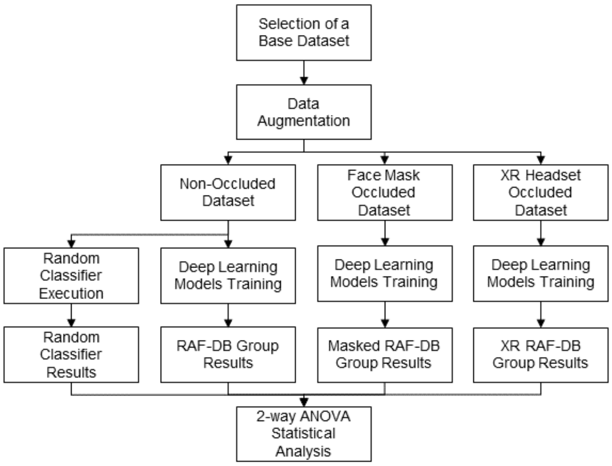
Selection of a Base Dataset
Data Augmentation
Non-Occluded Dataset
Face Mask Occluded Dataset
XR Headset Occluded Dataset
Random Classifier Execution
Deep Learning Models Training
Deep Learning Models Training
Deep Learning Models Training
Random Classifier Results
RAF-DB Group Results
Masked RAF-DB Group Results
XR RAF-DB Group Results
2-way ANOVA Statistical Analysis

FIGURE 2. Workflow followed in this study. First, a dataset was selected and augmented to address class imbalance. Then, two synthetic datasets introducing occlusion by face masks and XR headsets were created, resulting in three datasets (no occlusion, occlusion by face masks, and occlusion by XR headset). Deep learning models were trained on each dataset. The best results obtained were used to create three groups of metrics. Additionally, a random classifier was executed to obtain a baseline group. Finally, the four groups of metrics were compared using a 2-way ANOVA statistical analysis.

III. MATERIALS AND METHODS

A. DATASET SELECTION AND DATA AUGMENTATION

The first step of this study involved the meticulous selection of the datasets to train the models. In this process, we thoroughly examined facial emotion recognition (FER) datasets available on the web and identified in the preceding section. The following inclusion criteria were established: i) the dataset must consist of images, not videos, to simplify the selection of representative frames; ii) subjects must be real humans, excluding computer-generated avatars for greater generalizability to real-world scenarios; iii) images must not be posed to enhance applicability to real-world settings; iv) images must be in color, as color information can aid emotion classification and compensate for information loss introduced by partial occlusion; v) input size must be $224 \times 224 \times 3$ or greater, aligning with the input requirements of our models to avoid quality loss and deformation when resizing is needed; and vi) the total number of instances in the full dataset (combining training and test sets) must surpass 10,000 images to ensure achieving robust results. Table 2 provides a summary of the datasets explored, along with reasons for their inclusion or exclusion based on the established criteria.

After carefully explore available options, the only two datasets that met our criteria were RAF-DB [41] and AffectNet. However, AffectNet was discarded due to the unavailability of its test set. The RAF-DB dataset consists of 15,339 images (training set: 10,226; test set: 5,113) of faces captured in the wild from the web. The images are classified into seven emotions: happiness, sadness, anger, surprise, fear, disgust and neutrality, with each image labelled by at least 40 independent annotators.

TABLE 2. Datasets considered for this study.

Dataset	Reason discarded
RAF-DB	Comply with criteria. Dataset selected for evaluation.
AffectNet	Comply with criteria. Later discarded. *
4DFAB	Composed by video sequences and no images.
AFEW	Composed by video sequences and no images.
Aff-Wild2	Composed by video sequences and no images.
BP4D	Composed by video sequences and no images.
CK+	Composed by video sequences and no images.
CREMA-D	The facial expressions are posed.
DAiSEE	Composed by video sequences and no images.
DISFA	Composed by video sequences and no images.
FER2013/FER+	Images are monochromatic and image size is lower than required by our criteria.
FEAFA+	Composed by video sequences and no images.
FERG	The subjects are not real humans, but computer generated avatars.
JAFFE	The facial expressions are posed.
MAFW	Composed by video sequences and no images.
MH-FED	The subjects are not real humans, but computer generated avatars.
MMI	Composed by video sequences and no images.
MSD-E	The total number of instances of the datast is small.
Oulu-CASIA	The facial expressions are posed.
RaFD	The facial expressions are posed.
RAVDESS	The facial expressions are posed.
SAVEE	Composed by video sequences and no images.
SFEW	The total number of instances of the datast is small.

* The AffectNet dataset was initially selected for evaluation. However, due to the test set not being released at the moment this study was performed, we had to discard it.

The first analysis we performed was to calculate the class distribution of the dataset. Note that for the dataset to be balanced, the percentage of instances per class should be roughly 14.28% ($100\% / 7 = 14.2857\%$). Fig. 3 illustrates the class distribution of the dataset in the training set and the test set before any further manipulation. It is notable how the class distribution is roughly the same in the two sets. However, the dataset is imbalanced towards the happiness class, with the classes fear, anger, surprise and disgust underrepresented. It is crucial here to acknowledge that this imbalance is intrinsic to the nature of expressions in the real world and common in facial expression recognition datasets [33], and thus, it is difficult to overcome.

To address the uneven class distribution in the dataset, data augmentation techniques were employed. Specifically,

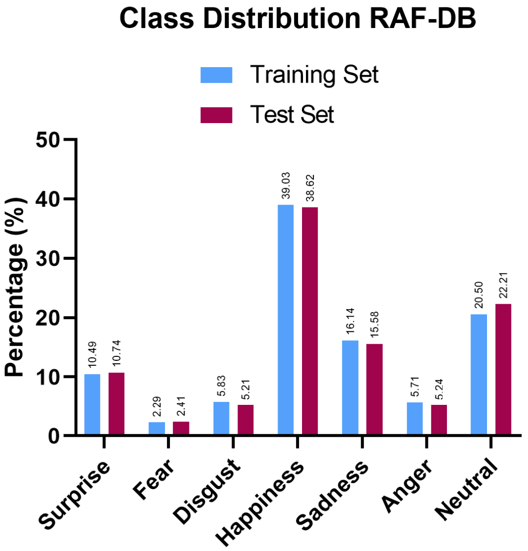

FIGURE 3. Class distribution of the RAF-DB dataset in the training set and the test set. The distributions of both datasets are similar, and highly imbalanced towards the happiness class, having the fear, anger, disgust, and surprise classes underrepresented.

CFR-GAN [42] was used to create new instances while preventing redundancy. The algorithm was used to frontalize and de-occlude the faces within the underrepresented classes (fear, anger, surprise and disgust) in the training set. To avoid biases in the results, the test set was not modified. Fig. 4 shows examples of the generated images. To ensure consistent facial proportions in the generated images, MediaPipe Face Detection [43] was used to detect the bounding boxes of the faces and crop them. Images where a face could not be detected were discarded. The new class distribution after augmenting the dataset is presented in Fig. 5. It can be noted how the imbalance is slightly corrected as the percentage of instances of the majority class (happiness) has decreased with respect to the percentage of instances of the minority classes (fear, anger, surprise and disgust). As a side effect, the percentage of instances of the sadness class was decreased too.

B. CREATION OF SYNTHETICALLY PARTIALLY OCCLUDED DATASETS

The following step was to create two synthetic datasets to simulate occlusion imposed by either a face mask or a XR headset. Mediapipe Face Landmark Detection was again used, but this time with the goal of identifying landmarks on the faces. Using the detected landmarks, face masks and XR headsets were overlayed on the faces. The overlay layers were resized and rotated to better align with the faces, ensuring that the faces were correctly partially occluded. This process aimed to realistically simulate the occlusion caused in real-world scenarios.

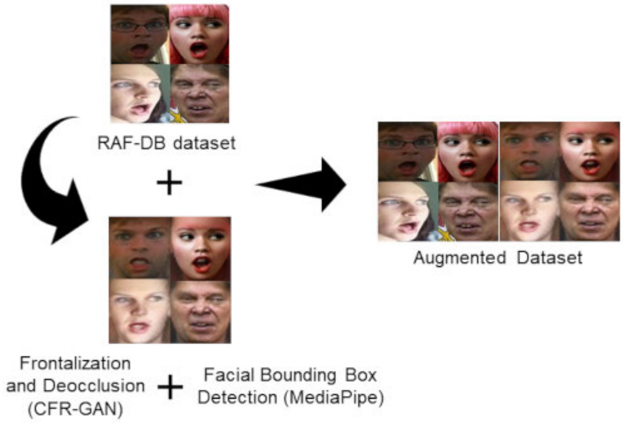
RAF-DB dataset
+
Augmented Dataset
Frontalization and Deocclusion (CFR-GAN)
+
Facial Bounding Box Detection (MediaPipe)

FIGURE 4. Data augmentation pipeline followed in this study. The images of the RAF-DB dataset belonging to the minority classes were frontalized and de-occluded using a generative method called CFR-GAN. These images were then cropped, by detecting the bounding box of the face using MediaPipe, to ensure consistent facial proportions in the generated images. The new images were added to the original dataset. Note how some facial characteristics of the person may change and elements like glasses are removed, while the general facial expression is kept.

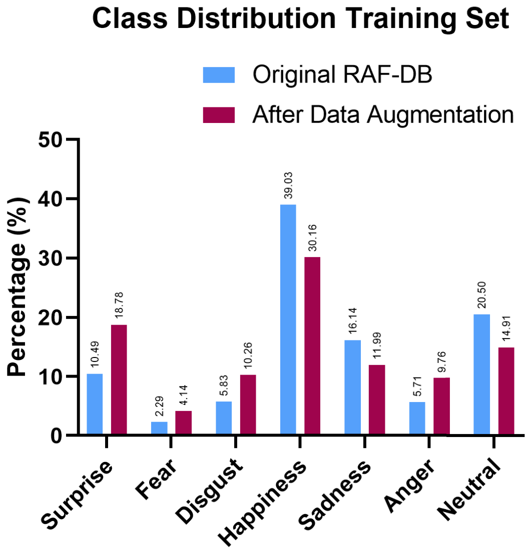

FIGURE 5. Class distribution in the training set of the RAF-DB dataset before and after augmenting the underrepresented classes (Fear, Disgust, and Anger). In the augmented dataset, the class distribution has been slightly corrected, though it is still imbalanced. Note that we have removed instances from the dataset not recognized by MediaPipe Face Detection, which may have also affected the class distribution.

This step resulted in three datasets: i) no-occlusion, ii) occlusion by face masks, and iii) occlusion by XR headsets. The class distribution in the training set remained unchanged, as depicted in Fig. 5. The reason of this is that we already removed images where faces were not detected in the previous step. However, the occlusion was added also to the test set, leading to the removal of images where faces were not detected from all three datasets. This resulted in a minimal variation in the class distribution of the test set, illustrated in Fig. 6. The pipeline used to generate these synthetic datasets is visually represented in Fig. 7.

C. SELECTION OF CLASSIFIERS AND HYPERPARAMETERS

To generate the necessary performance metrics for the statistical analysis, deep learning models must be selected, trained, and tuned. Following the trends observed in previous

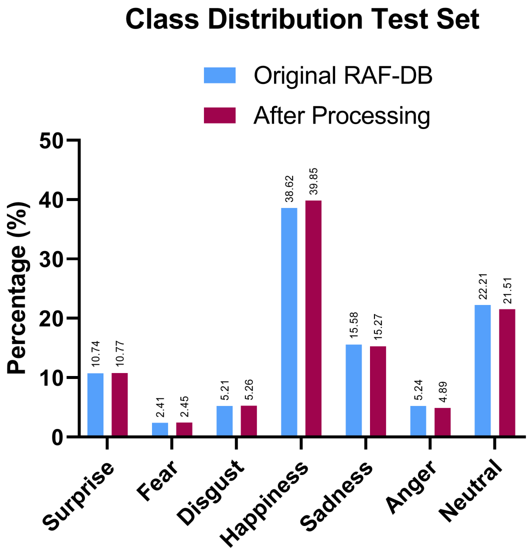

FIGURE 6. Class distribution in the test set of the RAF-DB dataset before and after introducing partial occlusion. The algorithm used to detect face landmarks did not detect faces in some of the instances, so they were removed causing a slight variation in the class distribution.

studies, transfer learning was employed. As described in [36], transfer learning serves as an efficient method to reuse existing trained models in a fast an easy way. Additionally, the original pretrained model may contain valuable information to mitigate the loss introduced by occlusion.

The keras_vggface python library was chosen for this study, primarily due to the availability of different pre-trained models and its ease of use for transfer learning tasks. The library provides access to models such as VGG-16 [44], ResNet-50 and Se-ResNet-50 (SeNet-50 for short) [45]. For each of these models, weights trained on the VGG-Face dataset [45] are downloaded and prepared for transfer learning. We followed the approach of updating all weights of the models that proven to be the best in [39].

The following step is the selection of an optimizer to train the models. This is an important decision as the optimizer plays a key role in adjusting the weights of the models by minimizing the loss function. Additionally, the choice of optimizer can impact the convergence of the models by avoiding local minima and facilitating faster convergence. In this study, seven different optimizers are evaluated: SGD (Stochastic Gradient Descent), RMSProp, Adagrad, Adadelta, Adam, Adamax, and Nadam.

The convergence of the models can be influenced by their hyperparameters, which is essential for achieving optimal results. Given the dataset size, the number of optimizers to test, and the computational resources required for training each configuration, a decision was made to focus on tuning the learning rate with the values: 0.01, 0.001 and 0.0001. The learning rate is a critical hyperparameter as it determines the step size taken to update the weights during training. A high learning rate may lead to overshooting, causing instability and bouncing during convergence, while a low learning rate may result in slow convergence. This focused approach allows for a balance between effective model training and efficient resource use.

TABLE 3. Summary of hyperparameters used in this study.

Name	Description	Values
Optimizer	Function that updates the weights of a model to minimize the loss function.	[SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax and Nadam]
Loss Function	Function that measures how well the model can predict the correct output of the model.	categorical crossentropy
Learning Rate	Step size taken by the optimizer during training, crucial for controlling convergence speed.	[0.01, 0.001, 0.0001]
Epochs	Number of times the training set is passed through the model during training.	20
Batch Size	Number of instances used in each iteration of the optimizer during training.	16
Folds	Number of partitions of data used in cross-validation.	5

The dataset labels were encoded using one-shot encoding, and the chosen loss function was categorical cross-entropy, which is well-suited for this type of encoding. To validate the models, stratified k-folds cross-validation [46] was selected. Stratified k-folds cross-validation addresses the class distribution imbalance by ensuring that the original class distribution is approximately preserved in each fold. A value of k=5 was used to ensure that the number of instances in each evaluation set is similar to the number of instances in the test set of the dataset.

The models were trained using the Keras library with TensorFlow backend. The hardware used was an Nvidia GeForce RTX 3080 Ti Laptop GPU with 16GB of dedicated memory, and 64GB of RAM memory. The batch size had to be lowered to 16 due to the weights of the models and images not fitting in memory. A maximum of 20 epochs was selected, with early stopping activated. The hyperparameters used in this study are summarized in Table 3.

D. EVALUATION METRICS AND COMPARATIVE ANALYSIS

The seven metrics chosen to evaluate the models were: accuracy, f1-score (macro and weighted), precision (macro and weighted) and recall (macro and weighted). The metrics were calculated using the scikit-learn python library. Note that the macro version of the metrics assigns equal weight or "relevance" to all classes in the dataset, providing an overall performance measure across all classes. On the other hand, the weighted version of the metrics assigns weight to each class based on the proportion of instances in the

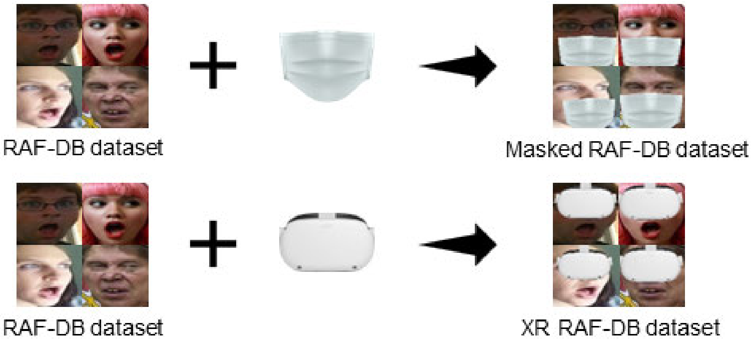

FIGURE 7. Pipeline followed to introduce partial occlusion on the instances of the non-occluded dataset. The faces were analyzed using MediaPipe to obtain face landmarks, and overlay images of a face mask and a XR headset were located in the faces, rotating and resizing according to face position and dimensions.

dataset, considering that the models could perform better on the majority classes than in the minority classes and that it could affect results [47].

To analyze and compare the results obtained from training the models on the different datasets, a 2-way ANOVA test with Tukey's post hoc analysis for multiple comparisons was used. This statistical test was employed to address if there are significant statistical differences between the different groups of performance metrics. Additionally, metrics were calculated from a random classifier over the non-occluded dataset. The random classifier was exclusively used on the non-occluded dataset because it assigns classifications randomly without learning characteristics from the image. Consequently, the results from the random classifier are expected to be independent of the dataset used. The random classifier serves as a baseline because it represents a classifier that knows nothing about the datasets. This makes it adequate since we can compare the results obtained by the rest of the models to assess if they are learning at all.

IV. RESULTS

The three classifiers were trained on the three datasets to obtain metrics for comparing model performance across different types of occlusion. A total of 189 result tables were generated during this training process (3 datasets x 3 models x 7 optimizers x 3 learning rates = 189 result tables). Each table contains the results of training the models through stratified k-folds cross-validation using 5 folds, resulting in 945 models trained under different conditions (189 tables x 5 folds = 945 times a model was trained). The total number of individual values obtained was 6,615 values (945 models trained x 7 metrics obtained per model = 6,615 individual values). In this section, only the best sets of results obtained for each dataset is presented. For a more in-depth analysis, including all tables and values, refer to the Supplementary Material (S1: Selection of Best Classifier per Dataset and S2: Result Tables Grouped).

After an in-depth analysis of the metrics (S1: Selection of Best Classifier per Dataset), the optimal outcomes for each dataset were identified. Specifically, for the non-occluded dataset, the best results were achieved using SeNet-50 with the Adadelta optimizer and a learning rate of 0.01. For the dataset with XR headset occlusion, the best results were achieved using SeNet-50 with the Adagrad optimizer and a learning rate of 0.001. For the dataset with face mask occlusion, the best results were achieved using ResNet-50 with the Adagrad optimizer and a learning rate of 0.001. To facilitate explanations the results obtained on the non-occluded dataset will be referred to as the RAF-DB group, those from the dataset partially occluded by XR headsets as the XR RAF-DB group, those from the dataset partially occluded by face masks as the Masked RAF-DB group, and results from the random classifier as the random classifier group. Table 4 provides a more detailed overview of the best results obtained for each dataset.

A 2-way ANOVA test using GraphPad Prism 8 was conducted, and the results are presented in Fig. 8. The statistical analysis revealed that the three groups (RAF-DB, XR RAF-DB and Masked RAF-DB) exhibit significant statistical differences when compared against the random classifier group. This indicates that all three groups successfully learned valuable information to classify emotions. Additionally, when comparing the non-occluded group (RAF-DB) against the two partially occluded groups (XR RAF-DB and Masked RAF-DB), significant statistical differences are also observed. This aligns with our expectations, as even though the two partially occluded groups (XR RAF-DB and Masked RAF-DB) are learning when compared against the random classifier, occlusion leads to loss of information when compared to the non-occluded dataset (RAF-DB).

The most interesting findings emerged when comparing the two partially occluded groups (XR RAF-DB and Masked RAF-DB). Notably, there were no significant statistical differences in terms of accuracy, recall (weighted) and f1-score (weighted). However, significant statistical differences were observed in terms of precision (macro), recall (macro) and f1-score (macro). The disparities in the macro metrics can be attributed to the imbalanced nature of the dataset. In imbalanced datasets, models tend to favor the majority class due to the availability of more information (instances)

TABLE 4. **Summary of the hyperparameters used and the metrics obtained from the best models per each dataset. The metric values in this table are the mean (with standard deviation) of the five folds calculated per each model. The tables with all of the obtained metric values per fold and a detailed description of the results can be seen in the Supplementary Material (S2: Result Tables Grouped).**

Random Classifier		
Metric	Mean	Std
Accuracy	0.1420	0.0110
F1-score (macro)	0.1220	0.0084
Precision (macro)	0.1400	0.0071
Recall (macro)	0.1480	0.0130
F1-score (weighted)	0.1620	0.0110
Precision (weighted)	0.2420	0.0110
Recall (weighted)	0.1420	0.0110
RAF-DB Group		
Classifier	SeNet-50	
Optimizer	Adadelta	
Learning Rate	0.01	
Metric	Mean	Std
Accuracy	0.8780	0.0110
F1-score (macro)	0.8260	0.0207
Precision (macro)	0.8340	0.0313
Recall (macro)	0.8260	0.0207
F1-score (weighted)	0.8800	0.0071
Precision (weighted)	0.8820	0.0045
Recall (weighted)	0.8780	0.0110
XR RAF-DB Group		
Classifier	SeNet-50	
Optimizer	Adagrad	
Learning Rate	0.001	
Metric	Mean	Std
Accuracy	0.7400	0.0100
F1-score (macro)	0.6340	0.0167
Precision (macro)	0.6400	0.0292
Recall (macro)	0.6320	0.0045
F1-score (weighted)	0.7380	0.0084
Precision (weighted)	0.7420	0.0084
Recall (weighted)	0.7400	0.0100
Masked RAF-DB Group		
Classifier	ResNet-50	
Optimizer	Adagrad	
Learning Rate	0.001	
Metric	Mean	Std
Accuracy	0.7520	0.0148
F1-score (macro)	0.6940	0.0167
Precision (macro)	0.7100	0.0212
Recall (macro)	0.6820	0.0205
F1-score (weighted)	0.7540	0.0152
Precision (weighted)	0.7580	0.0084
Recall (weighted)	0.7520	0.0148

about it. Consequently, the models may become better at distinguish the majority class from the other classes. The weighted metrics consider the number of instances in each class when calculating the results, reflecting this bias. On the other hand, macro metrics weight all of the classes equally when calculated, not reflecting this bias. Hence, the observed discrepancy may be attributed to the minority classes performing worse in the XR RAF-DB group than in the Masked RAF-DB group. This differences may be attributed to the majority classes performing better in the XR RAF-DB group than the minority classes. This aligns with the findings of [40], that concluded that the mouth is more informative for neutral, happiness and anger emotions than the eyes (happiness and neutral are majority classes in our study), and the eyes are more informative for surprise. This can also be observed in the metrics obtained per each class in the supplementary material (S1: Selection of Best Classifier per Dataset), where the only emotion not aligning with the finding of [40] is anger. Thus, the majority classes performing better in the XR RAF-DB group than the minority classes, while keeping the overall performance the same, could explain the significant statistical differences in the macro metrics. Significant statistical differences were also observed in terms of precision (weighted). After examining the results, the conclusion was that the significance is statistical but not practical, since the differences between the means of the two groups were low when performing the 2-way ANOVA test.

V. DISCUSSION

This study focuses on investigating the statistical differences between the absence of occlusion and two types of occlusion (occlusion introduced by a XR headset, and occlusion introduced by a face mask) during the classification of facial expression. The evaluation revolves around the statistical variances in performance metrics obtained from training deep learning classifiers on the datasets, each featuring one of the types of occlusion. The occlusion was introduced synthetically by overlying images of an XR headset or a face mask onto the faces within the RAF-DB dataset.

To generate performance metrics, three CNN models (ResNet-50, SeNet-50 and VGG-16) were trained using transfer learning. The models were fine-tuned by adjusting the optimizer and learning rate. The top-performing models for each dataset were selected (refer to S1: Selection of Best Classifier per Dataset), and metrics from these models were then used to create three evaluation groups: i) RAF-DB group: results from the best-performing model on the non-occluded dataset, ii) XR RAF-DB group: results from best-performing model on the dataset partially occluded by XR headset-s, and iii) Masked RAF-DB group: results from the best-performing model on the dataset partially occluded by face masks. Additionally, a fourth group named Random Classifier, consisting of results from a random classifier applied to the non-occluded dataset, was included as a baseline.

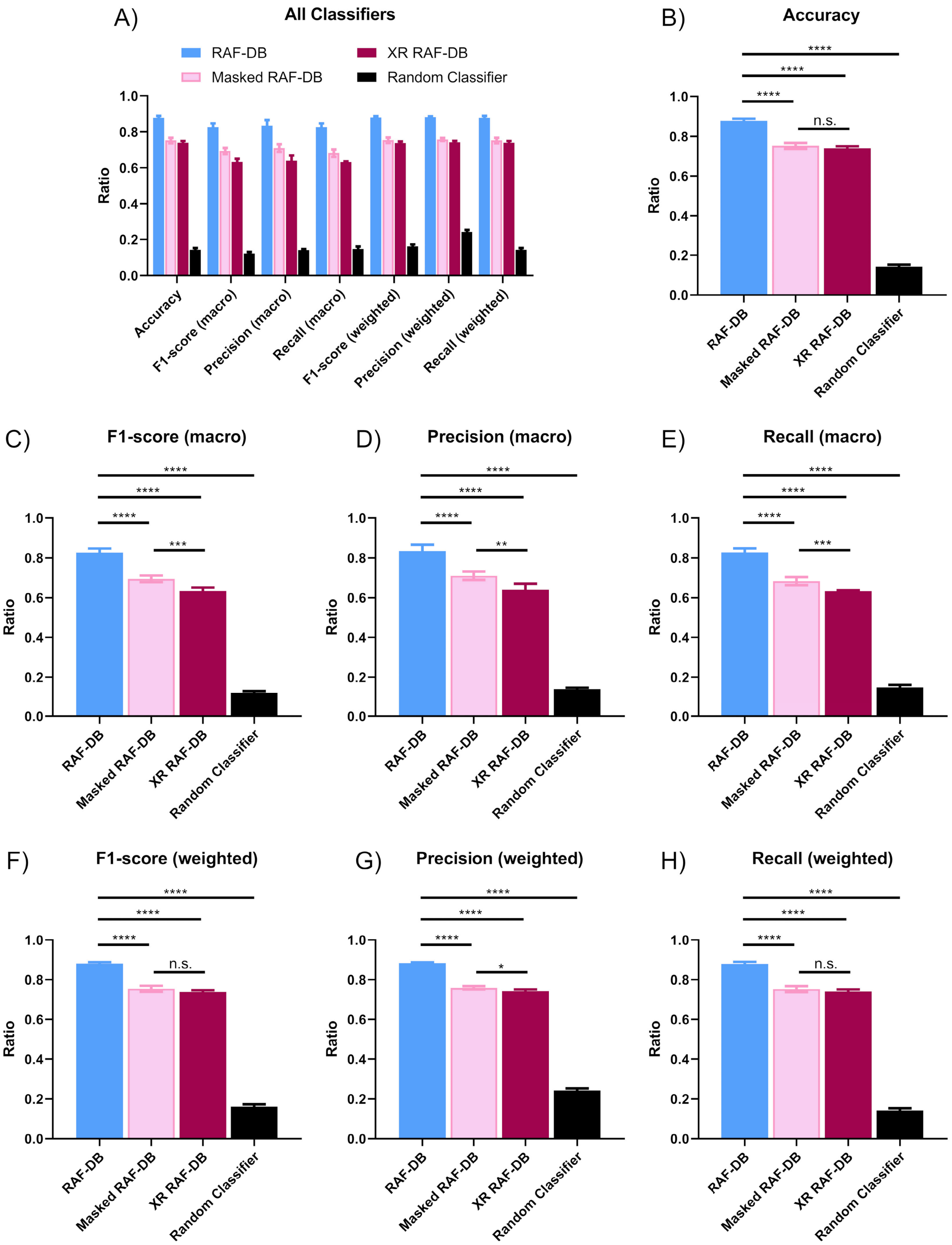

FIGURE 8. Statistical analysis by 2-way ANOVA test. Representation of (A) evaluation metric values obtained by the classifiers, (B) accuracy, (C) f1-score (macro), (D) precision (macro), (E) recall (macro), (F) f1-score (weighted), (G) precision (weighted), (H) recall (weighted). Bars indicate mean ± SD (B) F (3, 16) = 3888, **** $p < 0.0001$, (C) F (3, 16) = 1797, **** $p < 0.0001$, *** $p = 0.0001$, (D) F (3, 16) = 797.2, **** $p < 0.0001$, ** $p = 0.0016$, (E) F (3, 16) = 1667, **** $p < 0.0001$, *** $p = 0.0008$, (F) F (3, 16) = 4376, **** $p < 0.0001$, (G) F (3, 16) = 5721, **** $p < 0.0001$, * $p = 0.0365$, (H) F (3, 16) = 3888, **** $p < 0.0001$ using two-way ANOVA with Tukey's post hoc analysis for multiple comparisons. n=5 for every group (one value per fold). According to the analysis, there are significant statistical differences between the Random Group and the other groups (RAF-DB, Masked RAF-DB, and XR RAF-DB). This finding suggests that all three groups have successfully learned valuable information from the faces. Furthermore, when the two partially occluded groups are compared (Masked RAF-DB and XR RAF-DB) with the non-occluded group (RAF-DB), there are also significant differences. This outcome is expected, since occlusion implies loss of information. Significant statistical differences were also observed in the macro metrics between the two partially occluded datasets. This discrepancy may be attributed to the majority classes performing better in the XR RAF-DB group than in the Masked RAF-DB group. Finally, significant statistical differences are also found on the weighted precision metric, but we consider that it has no practical significance.

The four groups were compared using a 2-way ANOVA test. The test results (Fig. 8) revealed significant statistical differences between the random group and the other three groups. This finding indicates that, independently of the occlusion introduced, all three groups are learning valuable information for facial expression classification. Significant statistical also emerged when comparing the non-occluded group against the partially occluded groups, implying that the non-occluded group acquired more information, as expected due to occlusion causing information loss. In the comparison between the two partially occluded groups (XR RAF-DB and Masked RAF-DB), statistical differences were observed only in the macro metrics. These differences may stem from the minority classes of the XR RAF-DB group performing worse than those in the Masked RAF-DB group due to the distinct impact of occlusion types on different emotions.

The findings align with the first part of our hypothesis (“comparing facial expressions without occlusion to those with partial occlusion caused by face masks or XR headsets will reveal significant statistical differences”), indicating that there are significant statistical differences between the results obtained on facial expressions with no occlusion and those with occlusion caused by face masks or extended reality (XR) headsets. Interestingly, the second part of our hypothesis (“there may not be significant statistical differences when comparing occlusion by face masks with occlusion by XR headsets”) is contradicted by our results due to the fact that significant statistical differences in some metrics highlights the impact of different occlusion types on emotion recognition tasks.

The full source code for this study is accessible in the GitHub repository under an open-source license: https://github.com/Physical-User-Modeling-PhyUM/XR-FaceMask-EmoClass. Although the code is tailored to the RAF-DB dataset structure, it can be readily adjusted for use with other datasets through minor modifications.

A. LIMITATIONS

This study shows that even when faces are partially covered by XR headsets or face masks, meaningful information can be extracted for facial expression classification, as revealed by our statistical analysis. However, it is crucial to recognize the limitations of our research, which should be taken into account when interpreting the findings and planning future studies:

1) The dataset employed in this study is highly imbalanced. Although we used various strategies to address the issue, like data augmentation, the use of informative metrics and the implementation of stratified k-folds cross-validation, it is important to acknowledge that the class imbalance may still impact the results of our study.
2) The synthetic nature of the added partial occlusion may not fully replicate the complexity and variability present in real-world scenarios. While the use of synthetic occlusion is justified for the specific goals of this study, we acknowledge that future research should emphasize the application of these methods in authentic real-world settings.
3) Despite the results obtained with partial occlusion being satisfactory, it is important to note that the occlusion is still decreasing the performance of the models. This is expected as loss of information is inherent to occlusion in computer vision problems.
4) This study and related works focus only on basic emotions, sometimes incorporating additional emotions like neutrality or contempt. This narrow focus could be limiting, especially in scenarios such as educational settings, where the assessment of more complex emotions like boredom, engagement, or frustration may be necessary. One dataset suitable for this task is the DAiSEE dataset [48], [49], although it is formed by videos and not images. For instance, in psychomotor learning, which involves the interconnection of cognitive, psychomotor, and affective aspects of learning [50], the assessment of emotions associated with educational settings can be beneficial. Some studies, such as those reviewed in [51] utilize XR headsets for educational settings, particularly in learning medical skills, where learners may need to wear face masks as per protocol.

B. FUTURE LINES OF RESEARCH

Future lines of research should focus on overcome the limitations here mentioned, as well as in pushing the state of the art of this study and the studies mentioned in the related work:

1) As existing datasets become outdated over time, future research should prioritize the creation of new datasets or the adaptation of existing ones to meet modern needs. One primary limitation involves the inherent imbalance in facial expression recognition datasets, as evident in RAF-DB [41], Affectnet [33] and FER+ [52], [53]. Different approaches can be considered to address this issue. Some may involve labor-intensive efforts, such as expanding datasets with new instances, while others may be more sophisticated, such as leveraging generative models to generate new instances for underrepresented classes. Moreover, certain datasets like FER+ may consist of images with dimensions too small for modern architectures, necessitating image resizing and consequent quality loss. The application of super-resolution techniques [54] could offer a solution to mitigate this problem.
2) As mentioned before, certain scenarios may require the evaluation of emotions not conventionally explored in facial expression recognition research, such as boredom or engagement within educational settings [55]. Investigating how existing models perform when applied to these new emotions can prove valuable, particularly in contexts like psychomotor learning, where the affective domain is intricately linked to the

cognitive and psychomotor domains. In addition, how to provide affective feedback to the learners in an effective way [56] while using multisensorial channels that emerge when the learning environment extends the traditional computer-based scenarios [57].

3) To address the decline in performance resulting from partial occlusion, multimodal approaches can be used. These approaches involve collecting a broader set of signals that can be analyzed in synchrony to increase performance. For instance, some psychomotor learning systems may already have the required resources for this. As an example, the KUMITRON system [58] is an example that integrates inputs from physiological sensors. This system that uses cameras and physiological sensors can be easily adapted to estimate emotions using a multimodal approach, offering a potential solution to the challenges posed by facial occlusion.
4) The high availability and affordability of smartphones and smartwatches makes them ideal tools for emotion estimation. These devices typically feature cameras and inertial sensors that can be leveraged to analyze facial images and motion for emotion estimation. Moreover, newer smartphones may even incorporate physiological sensors like heart rate monitors or skin conductivity sensors. Expanding on this notion, psychomotor learning systems that utilize smartphones as a complete infrastructure, such as KSAS [59] could serve as a foundational framework for advancing research in this direction, by capturing information using the smartphone as a wearable, and executing lightweight and fast models directly on the device as it is done in [39], highlighting the potential of these ubiquitous devices in emotion estimation research.

VI. CONCLUSION

This study undertakes a statistical analysis to evaluate the performance of deep learning models in facial emotion recognition under partial occlusion. The partial occlusion types included in this study are the use of face masks and the use of XR headsets, two scenarios that are becoming more and more common in our lives. To do this analysis, a real-world facial emotion recognition dataset was selected and processed, addressing class imbalance and introducing synthetic occlusion through overlaying face masks and XR headsets images on the instances of the dataset. This resulted in three datasets: i) non-occuded dataset, ii) dataset with occlusion introduced by face masks, and iii) dataset with occlusion introduced by XR headsets. Three deep learning models (VGG-16, ResNet-50 and SeNet-50) were fine-tuned, and the best results obtained for each datasets were analyzed using a 2-way ANOVA statistical test.

This study highlights the potential of deep learning models with transfer learning for recognizing facial emotions in static faces, even with partial occlusion. Despite the inherent imbalance on facial emotion recognition datasets and the challenges posed by partial occlusion, the models demonstrated the ability to learn information, although with a significant decrease in performance when the faces where partially occluded. Notably, we observed no significant statistical differences between the two types of occlusion, with exceptions for certain emotions where the occluded parts of the face could play a more prominent role.

Our findings and our systematic review underscore the need to address the limitations of existing facial emotion recognition datasets, including the lack of non-basic emotions like confusion, engagement or boredom. Additionally, the imbalance of classes is still a issue that, although inherent to the problem, should be addressed. We also encourage the use of multimodal approaches to overcome the effects of partial occlusion and the use of mobile devices for a more accessible emotion estimation. These efforts could be crucial for enhancing the practical applicability and robustness of facial expression recognition models in real-world scenarios, regardless of the presence of occlusion.

In summary, this research contributes valuable insights into the challenges and opportunities associated with facial emotion recognition under partial occlusion, paving the way for future advances in the field and the development of more reliable, versatile and accesible emotion recognition systems.

ACKNOWLEDGMENT

The authors would like to thank Shan Li, Weihong Deng, and JunPing Du for creating the RAF-DB dataset used in this study and for generously granting them with access to it. UNED funding has been obtained for open access publishing.

REFERENCES

[1] H. Hwang and D. Matsumoto, "Evidence for the universality of facial expressions of emotion," in *Understanding Facial Expressions in Communication: Cross-Cultural and Multidisciplinary Perspectives*. New York, NY, USA: Springer, 2015, pp. 41–56.

[2] R. E. Jack and P. G. Schyns, "The human face as a dynamic tool for social communication," *Current Biol.*, vol. 25, no. 14, pp. R621–R634, Jul. 2015.

[3] P. Xu, S. Peng, Y.-J. Luo, and G. Gong, "Facial expression recognition: A meta-analytic review of theoretical models and neuroimaging evidence," *Neurosci. Biobehavioral Rev.*, vol. 127, pp. 820–836, Aug. 2021.

[4] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, Apr. 2021.

[5] H.-H. Wang and J.-W. Gu, "The applications of facial expression recognition in human-computer interaction," in *Proc. IEEE Int. Conf. Adv. Manuf. (ICAM)*, Nov. 2018, pp. 288–291.

[6] M. A. Espinoza Mina and D. D. P. Gallegos Barzola, "Neuromarketing and facial recognition: A systematic literature review," in *Technologies and Innovation* (Communications in Computer and Information Science), R. Valencia-García, G. Alcaraz-Mármol, J. D. Cioppo-Morstadt, N. Vera-Lucio, and M. Bucaram-Leverone, Eds. Cham, Switzerland: Springer, 2018, pp. 214–228.

[7] O. C. Santos, "Psychomotor learning in martial arts: An opportunity for user modeling, adaptation and personalization," in *Proc. Adjunct Publication 25th Conf. User Modeling, Adaptation Personalization*. New York, NY, USA: Association for Computing Machinery, Jul. 2017, pp. 89–92.

[8] O. S. Ekundayo and S. Viriri, "Facial expression recognition: A review of trends and techniques," *IEEE Access*, vol. 9, pp. 136944–136973, 2021.

[9] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan, "Emotion in a century: A review of emotion recognition," in *Proc. 10th Int. Conf. Adv. Inf. Technology*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–8.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[11] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, May 2019.

[12] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[13] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, Jun. 2012.

[14] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020.

[15] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion recognition in immersive virtual reality: From statistics to affective computing," *Sensors*, vol. 20, no. 18, p. 5163, Sep. 2020.

[16] J. Y. Lee, "A study on metaverse hype for sustainable growth," *Int. J. Adv. smart Converg.*, vol. 10, no. 3, pp. 72–80, 2021.

[17] L. A. Cárdenas-Robledo, Ó. Hernández-Uribe, C. Reta, and J. A. Cantoral-Ceballos, "Extended reality applications in industry 4.0.—A systematic literature review," *Telematics Informat.*, vol. 73, Sep. 2022, Art. no. 101863.

[18] V. R. Curran, X. Xu, M. Y. Aydin, and O. Meruvia-Pastor, "Use of extended reality in medical education: An integrative review," *Med. Sci. Educator*, vol. 33, no. 1, pp. 275–286, Dec. 2022.

[19] A. Herur-Raman, N. D. Almeida, W. Greenleaf, D. Williams, A. Karshenas, and J. H. Sherman, "Next-generation simulation—Integrating extended reality technology into medical education," *Frontiers Virtual Reality*, vol. 2, pp. 1–14, Sep. 2021.

[20] E. G. Muñoz, R. Fabregat, J. Bacca-Acosta, N. Duque-Méndez, and C. Avila-Garzon, "Augmented reality, virtual reality, and game technologies in ophthalmology training," *Information*, vol. 13, no. 5, p. 222, Apr. 2022.

[21] T. Liu, J. Wang, B. Yang, and X. Wang, "Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom," *Infr. Phys. Technol.*, vol. 112, Jan. 2021, Art. no. 103594.

[22] S. Doukianou, D. Daylamani-Zad, and I. Paraskevopoulos, "Beyond virtual museums: Adopting serious games and extended reality (XR) for user-centred cultural experiences," in *Visual Computing for Cultural Heritage* (Series on Cultural Computing), F. Liarokapis, A. Voulodimos, N. Doulamis, and A. Doulamis, Eds. Cham, Switzerland: Springer, 2020, pp. 283–299.

[23] D. Checa and A. Bustillo, "A review of immersive virtual reality serious games to enhance learning and training," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 5501–5527, Mar. 2020.

[24] F. Luo, L. Zhao, Y. Wang, and J. Kato, "Contrastive representation learning for expression recognition from masked face images," in *Proc. 1st Workshop User-Centric Narrative Summarization Long Videos (NarSUM)*, New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 23–29.

[25] S. Sola and D. Gera, "Masked student dataset of expressions," in *Proc. 13th Indian Conf. Comput. Vis., Graph. Image (ICVGIP)*. New York, NY, USA: Association for Computing Machinery, May 2023, pp. 1–9.

[26] G. Castellano, B. De Carolis, and N. Macchiarulo, "Automatic emotion recognition from facial expressions when wearing a mask," in *Proc. 14th Biannual Conf. Italian SIGCHI Chapter (CHItaly)*. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 1–5.

[27] A. F. Abate, L. Cimmino, B.-C. Mocanu, F. Narducci, and F. Pop, "The limitations for expression recognition in computer vision introduced by facial masks," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11305–11319, Mar. 2023.

[28] S. D. McCrackin, F. Capozzi, F. Mayrand, and J. Ristic, "Face masks impair basic emotion recognition: Group effects and individual variability," *Social Psychol.*, vol. 54, pp. 4–15, Dec. 2023.

[29] M. Marini, A. Ansani, F. Paglieri, F. Caruana, and M. Viola, "The impact of facemasks on emotion recognition, trust attribution and re-identification," *Sci. Rep.*, vol. 11, no. 1, p. 5577, Mar. 2021.

[30] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai, "Visual attention analysis and prediction on human faces for children with autism spectrum disorder," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3s, pp. 1–23, Oct. 2019.

[31] F. Pazhoohi, L. Forby, and A. Kingstone, "Facial masks affect emotion recognition in the general population and individuals with autistic traits," *PLoS ONE*, vol. 16, no. 9, Sep. 2021, Art. no. e0257740.

[32] M. Ventura, F. Innamorato, A. Palmisano, G. Cicinelli, E. Nobile, V. Manippa, R. Keller, and D. Rivolta, "Investigating the impact of disposable surgical face-masks on face identity and emotion recognition in adults with autism spectrum disorder," *Autism Res.*, vol. 16, no. 5, pp. 1063–1077, May 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aur.2922

[33] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[34] D. Pamod, C. Joseph, V. Palanisamy, and S. Lekamge, "Emotion analysis of occluded facial expressions—A review of literature," in *Proc. ASU Int. Conf. Emerg. Technol. Sustainability Intell. Syst. (ICETSIS)*, Jun. 2022, pp. 423–429.

[35] T. Ortmann, Q. Wang, and L. Putzar, "Facial emotion recognition in immersive virtual reality: A systematic literature review," in *Proc. 16th Int. Conf. PErvasive Technol. Rel. Assistive Environments (PETRA)*. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 77–82.

[36] B. Houshmand and N. M. Khan, "Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2020, pp. 70–75.

[37] H. Yong, J. Lee, and J. Choi, "Emotion recognition in gamers wearing head-mounted display," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Mar. 2019, pp. 1251–1252.

[38] M.-I. Georgescu and R. T. Ionescu, "Recognizing facial expressions of occluded faces using convolutional neural networks," 2019, *arXiv:1911.04852*.

[39] N. Petrou, G. Christodoulou, K. Avgerinakis, and P. Kosmides, "Lightweight mood estimation algorithm for faces under partial occlusion," in *Proc. 16th Int. Conf. PErvasive Technol. Related Assistive Environments (PETRA)*. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 402–407.

[40] Z. Lian, Y. Li, J.-H. Tao, J. Huang, and M.-Y. Niu, "Expression analysis based on face regions in real-world conditions," *Int. J. Autom. Comput.*, vol. 17, no. 1, pp. 96–107, Feb. 2020.

[41] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[42] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee, "Complete face recovery GAN: Unsupervised joint face rotation and de-occlusion from a single-view image," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1173–1183.

[43] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond neural face detection on mobile GPUs," 2019, *arXiv:1907.05047*.

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

[45] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[46] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2. San Francisco, CA, USA: Morgan Kaufmann, 1995, pp. 1137–1143.

[47] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.

[48] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[49] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[50] O. C. Santos, "Beyond cognitive and affective issues: Designing smart learning environments for psychomotor personalized learning," in *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy* M. J. Spector, B. B. Lockee, and M. D. Childress, Eds. Cham, Switzerland: Springer, 2016, pp. 1–24.

[51] A. Casas-Ortiz, J. Echeverria, and O. C. Santos, "Intelligent systems for psychomotor learning: A systematic review and two cases of study," in *Handbook of Artificial Intelligence in Education*. Cheltenham, U.K.: Edward Elgar, Apr. 2023, pp. 390–421.

[52] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 279–283.

[53] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing* (Lecture Notes in Computer Science), M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Germany: Springer, 2013, pp. 117–124.

[54] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: A brief review," *Inf. Fusion*, vol. 79, pp. 124–145, Mar. 2022.

[55] M. Saneiro, O. C. Santos, S. Salmeron-Majadas, and J. G. Boticario, "Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches," *Sci. World J.*, vol. 2014, 2014, Art. no. e484873.

[56] R. Cabestrero, P. Quirós, O. C. Santos, S. Salmeron-Majadas, R. Uria-Rivas, J. G. Boticario, D. Arnau, M. Arevalillo-Herráez, and F. J. Ferri, "Some insights into the impact of affective information when delivering feedback to students," *Behaviour Inf. Technol.*, vol. 37, no. 12, pp. 1252–1263, Dec. 2018, doi: 10.1080/0144929X.2018.1499803.

[57] O. C. Santos, R. Uria-Rivas, M. C. Rodriguez-Sanchez, and J. G. Boticario, "An open sensing and acting platform for context-aware affective support in ambient intelligent educational settings," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3865–3874, May 2016.

[58] J. Echeverria and O. C. Santos, "KUMITRON: Artificial intelligence system to monitor karate fights that synchronize aerial images with physiological and inertial signals," in *Proc. 26th Int. Conf. Intell. User Interfaces Companion (IUI)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 37–39.

[59] A. C. Ortiz and O. C. Santos, "KSAS: A mobile app with neural networks to guide the learning of motor skills," in *Proc. XIX Conferencia de la Asociación Española Para la Inteligencia Artificial (CAEPIA20/21)*, Sep. 2021, pp. 997–1000.

ALBERTO CASAS-ORTIZ received the B.S. degree in computer science engineering from Universidad Complutense de Madrid, Spain, in 2016, and the M.S. degree in artificial intelligence from Universidad Nacional de Educación a Distancia, Spain, in 2020, where he is currently pursuing the Ph.D. degree in intelligent systems. He is a member of the PhyUM Research Center, Universidad Nacional de Educación a Distancia.

From 2015 to 2018, he was a Software Engineer at Hospital Central de la Defensa Gómez Ulla and Tecnitia Technologies SL, focusing on the applications of computer sciences to the health sector. After completing his M.S. degree, he has focused his career on the planning, design, and development of software for scientific scenarios. This includes the use of high-performance advanced linear algebra techniques and machine learning in the field of population genetics (University of Calgary, 2020–2022); the creation of intelligent systems to learn psychomotor skills using artificial intelligence techniques (Universidad Nacional de Educación a Distancia, since 2018); and the creation of biomechanics tools from the OpenSim team, including opensim-core, opensim-gui, opencap, addbiomechanics, and sit2stand (Stanford University, since 2022).

JON ECHEVERRIA was born in Elgoibar, in 1974. He received the degree in business administration from Gestio Eskola, in 1998, the Specialization degree in total quality from Euskalit, in 2004, the degree in computer engineering from UNED, in 2020, and the master's degree in artificial intelligence from UNED, in 2022. He is currently pursuing the Ph.D. degree in intelligent systems.

He is a highly skilled and passionate technology and business professional. He boasts a strong background in management and engineering. His extensive knowledge and experience in these fields make him a prominent leader in the business world. Professionally, he has co-founded and served as the Financial Director for Loetxe. He is also the Business Manager with the PhyUM Research Center. He has also made contributions to scientific research, particularly in artificial intelligence and machine learning applied to karate, with publications and presentations in international conferences. He is developing the KUMITRON psychomotor system (based on patent WO/2022/008768) to train karate combats, and which has already received entrepreneurship funding (Eikentzaile program) and been finalist in several entrepreneurship awards.

NEREA JIMENEZ-TELLEZ received the B.S. degree in biochemistry from Universidad Complutense de Madrid, Spain, in 2016, the M.S. degree in genetics and cell biology from Universidad de Alcalá, Spain, in 2017, and the Ph.D. degree in biochemistry and molecular biology from the Dr. Naweed Syed's Laboratory, University of Calgary, Canada, in 2022, studying the cellular and molecular mechanisms underlying anesthetic-induced cytotoxicity and their impact on learning and memory.

During her B.S. degree in 2015, she participated in the TASSEP Exchange Program, at the University of Saskatchewan, Canada, where she completed an Honors Thesis project on the Regulation of the Metastasis Suppressor Protein CREB3L1. During the M.S. degree, she worked on the Role of p27Kip1 in the Pluripotency and Differentiation of Dopaminergic Neurons. Currently, she is a Postdoctoral Scholar at Dr. Joseph Wu's Laboratory, at Stanford University, where she is working on the Toxicoepigenetic Effects of e-Cigarette Exposure Using Human iPSC-derived Organoids.

OLGA C. SANTOS is currently an Associate Professor with the Artificial Intelligence Department, Computer Science School, UNED, and Spanish Open University; and coordinates the "Master in Research in Artificial Intelligence." She is a main Researcher in the HumanAID Project, the President of the IAIED Society, and the Head of the PhyUM Research Center, among other responsibilities. She has participated in 16 research projects (UE and national) and published more than 200 articles. In her pre-doctoral research, she developed a data-driven user-centred methodology called TORMES to elicit the personalized support to be provided in interactive learning systems. During her postdoctoral research, she took into account the affective computing to enrich the personalization support to be offered through the AICARP platform. Her current research interests focus on combining artificial intelligence with sensing technologies to support personalized affective psychomotor learning in diverse domains, such as martial arts and sports, rehabilitation and active aging, following a human-centric approach.

Her awards and honors include the Best Doctoral Thesis Award from the IEEE Spanish Chapter of the Education Society and the 2014 Young Researcher Award from the IEEE Technical Committee on Learning Technology. She supervises the research on psychomotor systems of Alberto Casas-Ortiz (since 2018) and Jon Echeverria (since 2019), among other students.