IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2023, accepted 21 October 2023, date of publication 27 October 2023, date of current version 2 November 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3328210*

RESEARCH ARTICLE

# Deep Learning-Based Gender Classification by Training With Fake Data

**MOHAMED OULAD-KADDOUR[1], HAMID HADDADOU[1], CRISTINA CONDE VILDA[2], DANIEL PALACIOS-ALONSO[2], KARIMA BENATCHBA[3], AND ENRIQUE CABELLO[2], (Member, IEEE)**

[1]Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supèrieure d'Informatique, Oued-Smar, Algiers 16309, Algeria
[2]Escuela Tècnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, Campus de Mostoles, 28933 Madrid, Spain
[3]Laboratoire des Méthodes de Conception des Systèmes, Ecole Nationale Supèrieure d'Informatique, Oued-Smar, Algiers 16309, Algeria

Corresponding author: Mohamed Oulad-Kaddour (m_ouled_kaddour@esi.dz)

This work was supported in part by Spanish Ministerio de Ciencia e Innovación under Grant PID2021-124176OB-I00 MEDEA project.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the NIST Color FERET Team.

**ABSTRACT** Gender classification of human faces is a trending topic and a remarkable biometric task. This research area has useful applications in several fields, such as automated border control (ABC) and forensic work. There are many approaches to gender classification in the literature; the classical approaches usually use real faces. Although good performances have been achieved, data collection remains a problem. Additionally, the privacy of individuals must be included in many existing works. These drawbacks can be overcome by using fake faces. Recently, the creation of a robust fake face corpus using machine learning has become possible. Our main contribution in the present paper is to experimentally investigate the ability of an artificial deepfake corpus to be a substitute for real corpora in facial gender classification tasks. We propose a deep learning-based approach using convolutional neural networks trained with fake faces and tested on real faces. By exploiting artificial faces, data collection obstacles are resolved for the training step, and privacy is highly preserved. Four classifiers based on popular convolutional neural network architectures were implemented. In the test phase, we used faces of real identities extracted from well-known experimental databases such as Face Recognition Technology (FERET), Faculdade de Engenharia Industrial (FEI) faces, Face Recognition and Artificial Vision (FRAV) and Labeled Faces in the Wild (LFW). The results achieved are very promising. We obtained high accuracy rates and low EER scores. They are similar to those of research works using real faces. As a result of this work, we propose a gender-labeled deepfake facial dataset containing more than 200k deepfake corpora that we will make available upon request for research purposes.

**INDEX TERMS** Adversarial neural networks, convolutional neural networks, deep learning, fake faces, gender classification.

## I. INTRODUCTION

Gender classification (GC) is a biometric task of categorization. It is a binary classification problem for a permanent human attribute. It has been studied with various biometric modalities, such as fingerprints [1], hand [2], face [3], ears [4], periocular region [5], full-body [6], and oral regions [7]. The research community has been focusing on gender classification since 1990 [8]. It is one of the most active research areas in biometrics [9], [10], [11]. It allows us to exploit vital permanent information on human beings in various fields, such as trade, robotics, and demographic data collection [12], [13].

As in any biometric system, some criteria must be considered during the deployment of a gender classification approach, in particular [14], [15], [16]: performance endorsed on confident measures, durability for computed features, acceptability from the target population, preserving the

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

privacy of people's personal data and universality guaranteeing calculability of used features for every subject in the target population. In practice, facial modality is the most studied in the literature for gender classification tasks. It meets many of the cited criteria, especially in terms of universality, acceptability and performance [13], [14].

Deep learning is becoming the key to the state-of-the-art for gender classification from human traces captured in two-dimensional image form, specifically for face modality [3], [17]. Several approaches concerning various contexts have been proposed for gender classification. Some use hybridization with classical approaches such as support vector machine (SVM) [18] and adaptive boosting (AdaBoost) [3]. Nevertheless, deep learning-based approaches greatly outperform existing works in terms of performance [19]. However, deep learning requires, in its training phase, a large amount of data. In most of the works, a large facial database is built using internet images without the interested parties' agreement, thus violating people's privacy [20].

To overcome this problem, we investigate the use of fake data in the learning phase of gender classification. In recent years, learning-based approaches that allow the creation of fake identities for living beings have emerged [21], [22]. Especially in the case of human identity, there are recent approaches performing human-face retouch [22], [23], [24] and fake identity generation [25], [26], [27], [28]. In a brilliant work, [26] proposed generative adversarial neural networks (GANs) to generate faces of fake human identities. For more generated data realism, [27] improved the quality of created faces. A recent empirical study [29] perceived that fake faces are indistinguishable from real faces and more trustworthy. Furthermore, for machines, real versus fake face separation requires special attention, and the task is attracting researchers' interest [30], [31]. The question that arises is whether fake identities can be an alternative to real identities in biometric systems.

In light of the fact that the majority of state-of-the-art gender classification approaches exclusively utilize real human faces, this paper explores gender classification using artificial faces generated by generative adversarial networks (GANs) during the training phase. Convolutional neural networks (CNNs) were trained exclusively with GAN-generated faces [25], and the testing phase involved real human faces. This paper makes the following key contributions:

- Introduction of a novel facial gender classification approach that incorporates deepfake faces of non-existent identities as training data, prioritizing individuals' privacy. This approach not only addresses privacy concerns but also opens new possibilities for data augmentation in machine learning.
- Empirical assessment of the potential of deepfake faces to substitute real ones through an extensive experimental comparison of well-known CNN architectures for gender classification, evaluated on four real datasets. Our findings shed light on the robustness and adaptability of gender classifiers to diverse facial data sources.
- Performance evaluation of CNN-based gender classifiers trained with various deepfake datasets generated under controlled and uncontrolled facial variations. This analysis not only assesses the classifiers' accuracy but also provides insights into their generalization capabilities in real-world scenarios.
- In addition to the above, we have compiled and meticulously gender-labeled an extensive dataset of over 200,000 deepfake faces. This dataset is made available upon request to facilitate further research and foster advancements in the field of computer vision and privacy-preserving machine learning.

The present paper is organized as follows. Section II is a short review of the state-of-the-art on gender classification and generative adversarial networks. Section III describes the proposed approach. Section IV presents the experimentation and results. Conclusions and perspectives are given in Section V.

## II. RELATED WORKS

In this section, we briefly review the state-of-the-art for gender classification and generative adversarial networks' artificial faces.

### A. GENDER CLASSIFICATION

Human-face gender classification is widely studied, and the literature is rich in proposed works [12], [13], [25], [32]. Gender classification is performed in three phases. First, preprocessing includes principally facial bounding box computing. Second, feature extraction is performed in a discriminative vector. Finally, classification is performed for decision-making [12]. After facial region of interest (ROI) determination and based on the principle of input preprocessing, gender classification approaches can be qualified as global, local or hybrid [12]. In global approaches, the whole face is processed without segmentation. In local approaches, the information derived from the small face's regions is combined, such as the facial subregion and its landmarks. In hybrid approaches, both global and local methods are combined with eventual score fusion [12].

In the experimentation phase, gender classification approaches use databases dedicated to human-face analysis and recognition. FERET [41], FRAV2D [42] and FEI [43] are facial databases settled by experts in a controlled context. They have, in general, acceptable quality. LFW [44], GROUPS-Faces [45], CelebFaces [20], CASIA-WebFace [20], and MORPH [46] are more challenging facial databases. Those databases were designed to evaluate classification performances in an uncontrolled context. Occlusion, image quality and face poses are the most challenging variations in real-world databases.

In classical approaches, support vector machine (SVM) [47], local binary pattern (LBP) [48], Gabor filter [49], artificial neural network (ANN) [50], principal component analysis (PCA) [49], local directional pattern (LDP) [51] and AdaBoost [3], [52] are examples of tools that were

**TABLE 1.** State-of-the-art analysis.

| Approach | Type | Principle | Observation |
|---|---|---|---|
| Jian [10] | Global | CNN | +Improve GC accuracy for real-world context.<br>-Collection of a large web dataset without privacy terms clarification. |
| Ming [11] | Global | CNN+ ELM | +Introduce a hybrid approach using CNN and ELM (extreme learning machine).<br>-Eventual data overlapping (data division unclear). |
| Ayesh [33] | Hybrid | VSM+ CNN | +Make evidence visual saliency maps to improve performance.<br>+Multilevel CNN for attributes classification.<br>-Test on small sets on LFW and eventual data overlapping. |
| Rai [34] | Global | 2DCPA+ Gabor+ SVM | + Optimized features for robustness to illumination, facial expressions and noise.<br>-Low accuracy in real-world context/ |
| Afifi [3] | Local | CNN+ AdaBoost | + Introduce foggy face for GC.<br>+Well records for closed contexts.<br>+Training with isolated facial regions.<br>-Eventually, train/test overlapping subjects. |
| Van [18] | Global | CNNs+ SVM | + Investigation of CNNs fine tuning for GC.<br>+Robust face detection and augmenting detected face bounding box to improve accuracy.<br>-Eventual train/test sets overlapping subjects.<br>-Low accuracy for real-world context.<br>-Best records only for closed scenarios. |
| Hyperface [35] | Global | CNNs | +Exploiting CNNs for multi-attribute face categorization.<br>+Outperforms previous methods on LFW.<br>-Test-set size unclear. |
| Lee [9] | Local | CNN | + Train CNN with separate facial regions.<br>+Majority voting inter sub-scores.<br>+Outperforms various works on LFW.<br>-Train/test data overlapping. |
| Geeta [36] | Hybrid | LDP+ LBP+ SVM | + High accuracy for closed controlled context.<br>-Unchallenging and limited dataset.<br>-Eventually, train/test overlapping data. |
| Khan [17] | Local | CRF+ RDF | + Introduce novel facial method for GC.<br>+No overlapping data.<br>+Perfect accuracy on FERET dataset.<br>-Low accuracy on FEI dataset. |
| Tiago [6] | Global | CNN | + GC in the wild under challenging variations.<br>- Eventual train/test subjects overlapping. |
| Lu [37] | Global | CNN | + Exploit data augmentation for real-life GC.<br>-Train and test set sizes not clearly declared. |
| Sheikh [38] | Global | CNNs | + Introduce central difference for GC. |
| Mozhdeh [39] | Local | PixelHop +PCA + SVM/ LR/RF | +Introduce successive subspace learning for GC.<br>+ Acceptable acc on low-resolution context.<br>-Small test set on LFW with eventual overlapping. |
| Kimmo [40] | Global | CNN | +Propose and exploit a new balanced facial dataset for bias measurement and mitigation. |

largely exploited for feature extraction and classification steps. Detailed surveys and in-depth experimental studies are available in [12], [13], [32], and [37]. These approaches can be criticized primarily for their use of very small image sets in non real-world contexts during the testing phase.

Recent studies are almost all based on machine learning and, in particular, on convolutional neural networks (CNNs). They focus mainly on feature extraction or both feature extraction and classification. Although there are no standard rules for comparing state-of-the-art methods, Table 1 summarizes analyses of recent works by describing their type, exploited techniques for feature extraction and classification steps. Some advantages and critiques are highlighted. All existing approaches are deployed using one or more real databases among those described previously, and some of them use random image collections drawn from the web. Referring to Table 1, we note that existing approaches can be criticized for not respecting privacy where many used datasets collected from the web or legal processes in the collection protocol were not detailed, resulting in overlapping data. There were also some approaches that divided the dataset between training and testing phases without affirming non-duplication of identities between them (for a person with multi-face image acquisition), and the test size for some approaches that did not clearly declare the size of the used subset in the test step.

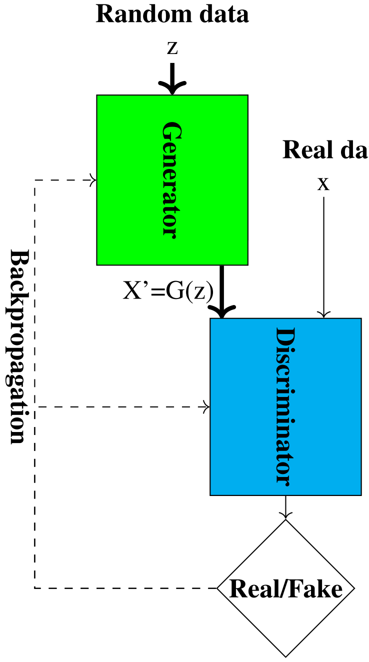


**FIGURE 1.** General overview of a generative adversarial network.

## B. GENERATIVE ADVERSARIAL NETWORK

In artificial intelligence, image generation is a task that aims to synthesize and translate images with the objective of generating novel realistic images. Generative adversarial networks or GANs are powerful tools for image generation [12], [22], [31], [53]. This kind of neural network was introduced by Goodfellow in 2014 [54].

As shown in Fig. 1, a generative adversarial network is a deep network structured principally in two subnetworks, both based on deeper networks: a generator G that generates synthetic (fake) data $x'$ from random input data z ($x' = G(z)$) and a discriminator D whose role is to distinguish (classify) the generated fake $x'$ from real data x. Called forger and expert, the generator and discriminator are competitive networks, respectively. The forger attempts to mislead the expert by creating realistic images emulating the nature of the real data. The expert ensures the perfect separation of fake/real data [48]. Performed simultaneously for its two subnetworks, the training of a GAN network is established based on loss backpropagation. For this, GAN's setting is optimized through the determination of the equilibrium (G*, D*), assuring satisfaction for both the forger, by decreasing the classifier's accuracy, and the expert, by increasing its accuracy. The equilibrium corresponds to the argument of the following optimization expression [54]:

$$(G^{*}, D^{*}) = Min_{G}Max_{D}[E(Log(D(x)) + E(Log(1 - D(x'))] \tag{1}$$

where x designates real samples browsing a real dataset and $x' = G(z)$ designates the fake-generated image for noise data z browsing a random dataset. D(.) estimates the expert behaviour in probability (1 for data predicted as real and 0 for data predicted as fake).

Since their introduction, faces generated by GANs have been adopted in various contexts for human-face analysis. Permanent and temporal human attribute manipulation, entire face synthetic creating faces of non real person, deepfake face-swap exchanging the objective face in a video by another target face and facial expression swapping are the most common emerging GAN face manipulations. In return, to limit the misuse of fake data for malicious purposes, researchers have also studied fake image detection [25].

In [55], an interesting survey of data augmentation for actual face recognition systems is presented. This shows that GAN-generated face synthesis increases the size of the training set and improves recognition performance. For face analysis with poor-quality images, Li et al. [56] exploited GAN faces to improve the quality of images for face recognition in the wild. Concerning security aspects, [21], [30], [31] proposed an antispoofing approach to discriminate between real and fake faces that can be exploited for antispoofing scenarios. For the same purpose, [57] studied the analysis of GAN-generated fingerprints. Reference [58] studied the generation of GAN faces by preserving human gender. Mescheder et al. studied the local convergence properties of GAN methods [59]. With the goal of mitigating gender classification bias across race groups and especially for women and dark-skinned persons, Ramachandran and Rattani [60] retrained a state-of-the-art GAN network to realize data augmentation by synthesizing view creation for existing identities. They reported accuracy enhancement and bias across gender-racial group reduction via experimental validation.

However, generating non-existing identity faces is one of the most salient tasks in face applications. In one of the greatest iconic GAN architectures [26], Karras proposed a generative adversarial neural network allowing the creation of deepfake faces. The proposed approach generates artificial faces with the improvement of fake semantic quality. The created faces are for non-existing identities and are extremely realistic (see Fig. 2) [27], [28].

## III. PROPOSED APPROACH

In this section, the proposed method is presented. By presenting the overview of our approach, the face detection principle and the convolutional neural networks.

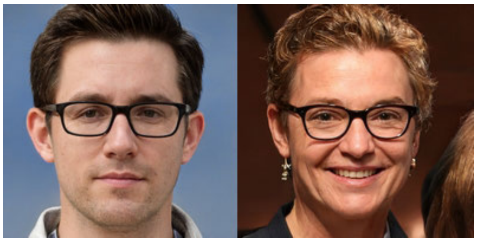


**FIGURE 2. Examples of realistic artificial faces of non-existing fake identities.**

As mentioned above, existing gender classification systems are usually performed using the real faces of existing persons for both the training and testing steps. The goal of our research study is the exploitation of deepfakes to perform a deep learning-based gender classification system.

The overview of our proposed method is illustrated in Fig. 3.

- **Training:** In this step, a fake dataset composed of artificial faces is exploited. After preprocessing, the fake images are used to train the convolutional neural network. We obtain a fake dataset's trained model. To tune the CNN's hyperparameters and avoid overfitting, a small subset is used for the validation of the trained CNN. The objective of this step is to perform the full setting of the convolutional neural network by using exclusively fake data.
- **Testing:** In this step, faces of real identities are exploited to assess the performances of the fake dataset's trained convolutional neural network. After face bounding box computing, gender prediction is performed by using the trained CNN. In our experimentation for this step, several datasets were used. The details and characteristics of these datasets are shown in the next section.

The used deepfake faces were generated by using the StyleGAN network [26] that was trained using a recent and high-quality facial dataset independent of those we exploited in the test. We applied dlib (http://dlib.net) person recognition tools to check semantic similarity between real and deepfake faces with the goal of ensuring that there are no real identities inadvertently incorporated in the deepfake dataset.

Next, we summarize some of the advantages of the proposed approach:

- The privacy constraint's problem is limited: Subjects' identities do not exist, and the training dataset can be shared or made in public access without a person's privacy violation [28].
- Data collection problems resolved for the training step: Deep learning requests a large quantity of data that is almost exploited in training. The use of automatic tools allows the generation of large fake datasets, as needed.
- Overlapping avoidance: By exploiting the face of artificial identities in the training phase, we avoid the data overlapping phenomenon. It reassures that there are

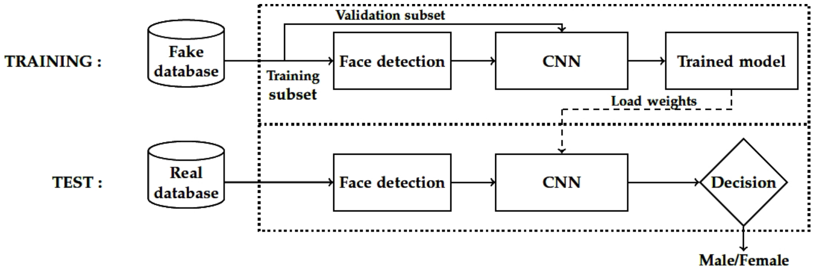


**FIGURE 3.** Overview of the proposed approach.

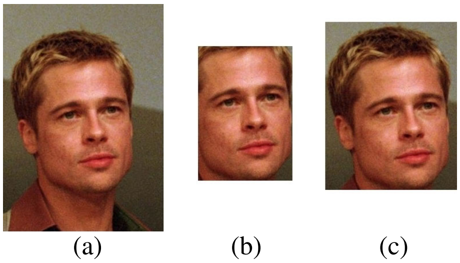


**FIGURE 4.** Examples of detected face bounding box augmentation (a): Input image (b): Detected face (c): Augmented bounding box.

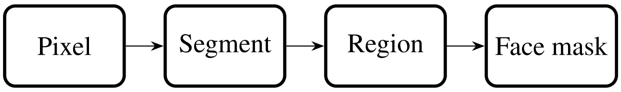


**FIGURE 5.** CNNs create implicit features at the pixel level for facial masks.

no duplicated identities between the training and testing sets.

- According to the variation in the fake dataset, it gives the opportunity to perform tests on the maximal subset of the experimental dataset without selective elimination or distribution of images between the training and testing subsets.

### A. FACE DETECTION

Face detection is an indispensable preprocessing step for facial image analysis, specifically for images with poor quality. In our work, we used a well-known and robust face detection method [61], the Region-CNN (R-CNN) algorithm [62]. Our preliminary experimentations showed that the internal face box returned by R-CNN had less discriminative information.

As shown in Fig. 4, after face detection and to inject more facial information, the detected face's bounding box is augmented with a portion of 0.3 (30%) for the returned size width.

### B. CONVOLUTIONAL NEURAL NETWORKS

Convolutional neuronal networks or CNNs are deep neural networks. They are generally powerful techniques for image classification [63], [64]. Taking the case of facial data and from a naive pixel level, CNNs can learn characteristics hierarchically, passing through small segments to interpretable regions and face masks [65](see Fig. 5).

Based on the philosophy of connectionism, in general, a CNN is composed of a set of layers, where each layer takes as input the output of the direct predecessors (priors). The principal types of CNN layers are [63], [64]: convolutional, pooling, rectified linear unit, fully connected, dropout and output layers. The convolutional layer is a fundamental layer aiming to learn feature maps so that the presence will be detectable in future subjects. Mathematics convolution operators are applied over an input matrix. A convolution filter is characterized by its kernel sizes. Frequently placed just after one or successive convolutional layers, with smaller sizes and with the objective of reducing the resolution of the feature map, a pooling operator is involved in the calculated feature maps. Average pooling (AvgPool) and maximum pooling (MaxPool) corresponding to a grid (submatrix) to the average and maximum values, respectively, are most commonly used. The rectified linear unit (ReLU) layer is a cell of the neural network in which a simple activation function is applied over an input vector to eliminate some rejected information. The basic formula used for an input vector X is:

$$\mathrm{RelLU}(X) = \max(0, X).$$

The fully connected layer, also known as the multilayer perceptron, connects all neurons of the prior layer to every neuron of its own layer. It is also called a dense layer when it implements a linear operation. Fully connected layers are generally placed at the end of deep networks. Addressing the overfitting problem and basically used on the fully connected layer, the key idea of the dropout layer is the random dropping of CNN neuron units during training. An output layer of a CNN designed for a k-classification problem is a vector formed with a list of estimated class probabilities. For each class, probabilistic information is stocked in the corresponding case on this vector, and the argument for the final decision class is related to the higher one. Formally, it can be defined as follows:

$$Prediction(X) = Arg[Max(OutPut(i), 1 \leq i \leq k] \quad (2)$$

where X is an input processed by the CNN and k is the class number.

For a given architecture, the configuration of the convolutional neural network is performed via the training

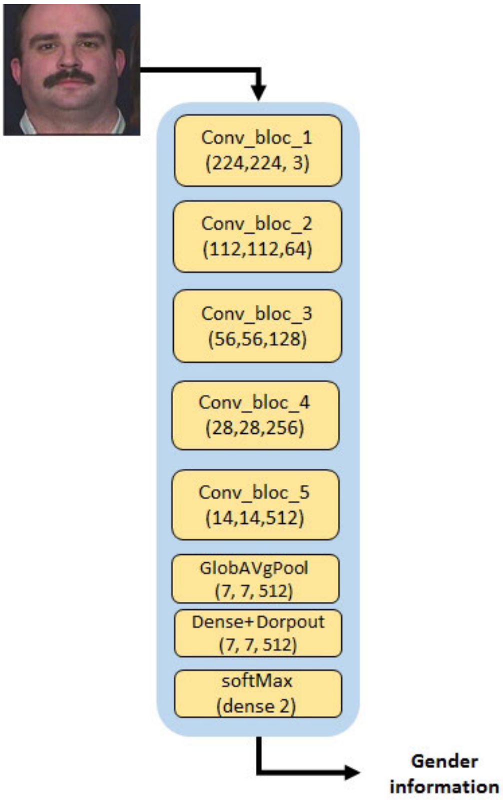


**FIGURE 6.** **Overview of adopted VGG16-inspired gender classifier.**

step, and its goal is the determination of the large number of CNN hyperparameters (on average, a few dozen of millions) produced by convolutional and fully connected layers. Training a CNN requires specific hardware, such as GPU machines with very large datasets [63], [64]. In practice, convolutional neuronal network architecture performance is benchmarked on very large datasets such as ImageNet [64].

In our work, in transfer learning, we fine-tuned random data pretrained convolutional neural networks. We adopt a very deep convolutional neural network (VGG16) [64] inspired gender classifier. For domain adaptation, we replaced the last fully connected layers with personalized layers, namely, the average pooling layer, dropout layer and ReLU layer. A softmax activation function is used as the final binary classifier for gender prediction. Fig.6 shows an overview of the adopted gender classifier architecture. We also compared the classifier with other well-known convolutional neural network architecture-based networks, namely, Inception-V3 (GoogLeNet version-3), ResNet (deep residual neural network) and MobileNet (efficient convolutional neural networks for mobile vision applications) [64].

## IV. EXPERIMENTAL PROTOCOL

In this section, we describe the experimental context of our approach. Artificial faces, real faces datasets and evaluation metrics are presented.

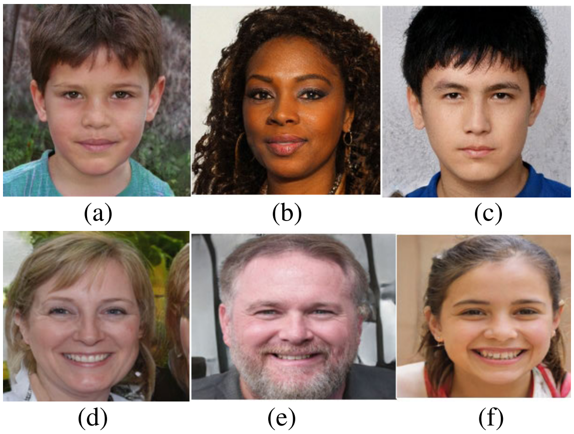


**FIGURE 7.** **100k-generated-images artificial dataset's variation examples (a): Young male (b): Black female (c): Asian male (d): Smiled old female (f): Bearded old male (f): Smiled young female.**

### A. ARTIFICIAL FACES DATASET

For the training step, we used the 100k-generated-images dataset. This is a gender unlabeled artificial facial dataset whose high-quality images were automatically generated by the StyleGAN model proposed by Karras et al. [25], [26]. The background of the 100k generated images dataset is the FFHQ (Flickr-Faces-High-Quality) dataset exploited for the training of the StyleGAN model. FFHQ contains 70k real facial images of high resolution, collected under permissive licences from the Flicker platform. FFHQ objective faces were detected and preprocessed using dlib library tools.

Fig. 7 illustrates some samples of the 100k-generated-images dataset.

It has multiple variations, principally:

- Ethnic: Black, White, Asiatic, Indian.
- Age: Young age, middle age, old age.
- Facial expressions: Happy, yelling, surprised, laughing, sad, etc.
- Natural and synthetic accessories: Beard, moustache, glasses, hats, etc.
- Background: Random background.
- Face pose: Frontal, semi-profile.

In addition, as a facial dataset, the 100k generated images can be qualified with the following advantageous properties: traditional deployment for the image acquisition process is not needed, the collection does not take time, and it is rich in terms of facial variations.

To perform our experimentation, a gender-balanced subset composed of 60k artificial images from the 100k-generated image dataset was exploited (see Table 2). To reduce the raw artificial face labeling cost, we defined a semiautomatic process for image gender labeling. First, we manually labeled a small set of 2000 images that was used to perform a fake gender classifier. Then, the fake gender classifier was applied to realize pseudo-labeling of the rest of the whole subset of 60k. Finally, the pseudo-labeled images were manually checked to eliminate falsely affected images in each class.

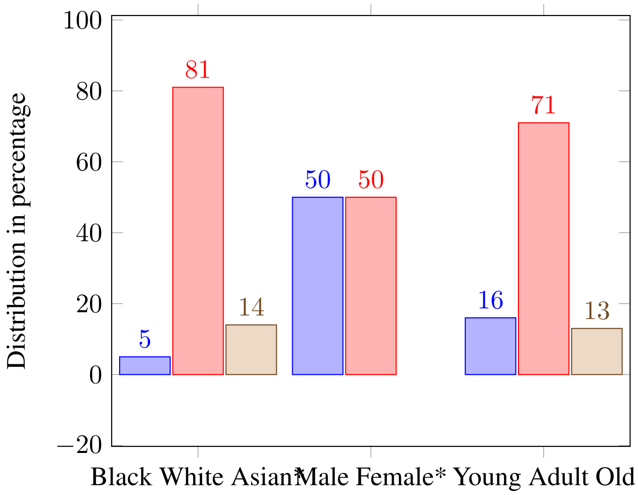


**FIGURE 8.** Training deepfake dataset attributes distribution.

**TABLE 2.** Details of the artificial training set.

| Gender | size in subject number |
|---|---|
| Man | 30k |
| Woman | 30k |

Fig. 8 shows the training deepfake dataset distribution in terms of race, gender and age attributes.

## B. REAL DATASETS

For the test task, we used the following experimental real facial datasets: FERET, FRAV2D, FEI and LFW.

- **FERET dataset:** FERET is a well-known database for human-face recognition and analysis. The images were collected in indoor lab conditions. It is a comparably simple dataset. FERET contains 1199 subjects and 14,126 images produced by random multiple acquisitions per subject. Faces were captured with multiple variations in face pose, facial expression and illumination [41]. In our case, we used the coloured part of frontal and semi-profile images.
- **FRAV2D DATASET:** FRAV2D is a colour Spanish dataset collected in the FRAV laboratory of URJC University. This dataset contains 3488 images of 320*240 resolution and acceptable quality. The acquisition was fairly performed for the 109 subjects (75 men and 34 women) by capturing 32 facial images per subject. The principal variations are facial expressions, lighting conditions, background and face pose (frontal and semi-profile). The database is delivered for free exclusively for research purposes [42].
- **FEI dataset:** Faculdade de Engenharia Industrial or FEI faces is a Brazilian dataset. This dataset contains 2800 facial images that were fairly captured for 200 subjects. The age range of the subjects 19 to 40. For the gender classification task, FEI is a balanced dataset. The principal variations are facial expression, face poses, image quality and background [43].
- **LFW dataset:** Labeled faces in the wild is a large challenging dataset collected in the wild. With random multi-faces per subject, the LFW is an unbalanced dataset in terms of gender attributes. It contains 13,233 images for 5479 subjects. LFW was originally created for face identification, but it is also used for face categorization tasks. The images of this dataset are of very poor quality and contain many variations, such as pose, illumination, occlusion, and facial expressions [44]. In our work, the whole part of the frontal and semi-profile face is used.

**TABLE 3.** Used real datasets FE: Facial Expression, I: Illumination, P: Face Pose, BG: Background, Q; Image Quality, W: Face in the Wild, E: Ethnic, O: Face Occlusion.

| Real dataset | Subject number | Image number | Variations |
|---|---|---|---|
| FERET | 1199 | 14k | FE, I, P, BG, E |
| FEI | 200 | 2800 | Q, I, FE, P |
| FRAV2D | 109 | 3488 | FE, P, O |
| LFW | 5749 | 13233 | FE, I, P, BG, Q, W, R, O |

**TABLE 4.** Details of used real subset for the test step.

| Real dataset | Used set | Size |
|---|---|---|
| FERET | Frontal, semi-profile | 6233 |
| FRAV2D | Whole dataset | 3488 |
| FEI | Whole dataset | 2800 |
| LFW | Frontal, semi-profile (Almost completely) | 13164 |
| Total tested images | - | 25685 |
| Total tested persons | - | 7257 |

Table 3 summarizes the characteristics of the described face datasets with their principal variations. According to the variations of the training artificial dataset, all variations of the used real datasets were considered at the test step. With the exception of the face poses presented in the FERET and LFW datasets, colour images with frontal and semi-profile face poses were selected. Fig. 9 shows samples from real used datasets. The details of the sets used for the evaluation of the trained models are shown in Table 4.

## C. EXECUTION CONTEXT AND EVALUATION METRICS

As specified previously, the fine-tuned CNNs were trained using 50 epochs with artificial faces generated by GAN and tested with real faces. To accelerate the processing time, all executions were performed by exploiting a workstation integrating GPU (graphics processing unit) memory for parallel computing. The details of the used GPU are summarized in Table 5. For each experiment, we used model checkpoints to call back the best performances obtained in intermediary epochs. The checkpoints were based on accuracy improvement during the epoch's execution.

To enrich variations and expand the size of the training set, we adopted real-time data augmentation by applying the Keras deep learning framework's (https://keras.io/api/

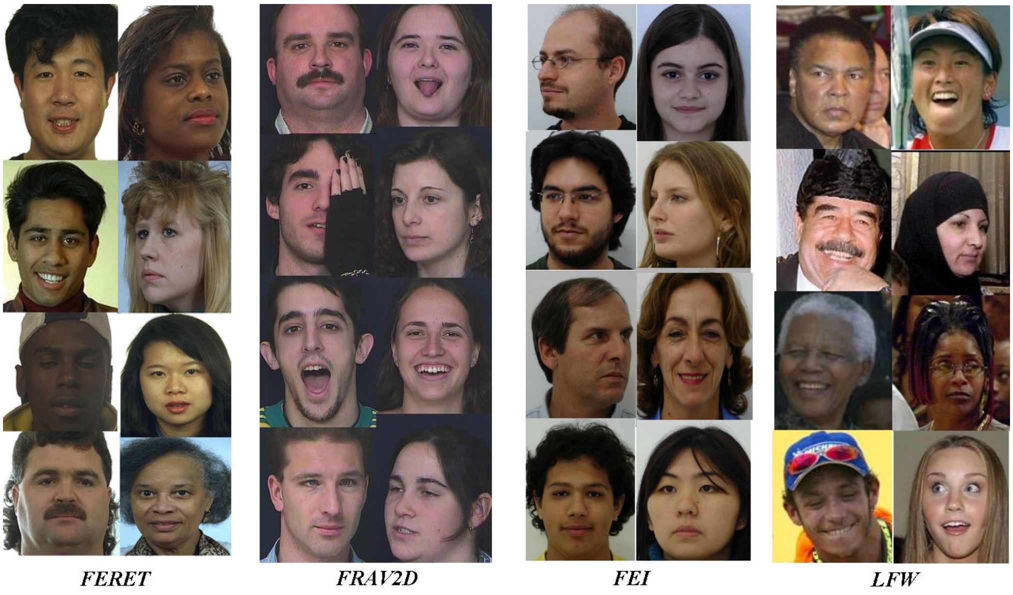


**FIGURE 9.** Samples of real faces of experimentally used datasets.

**TABLE 5.** Used GPU memory characteristics.

| GPU | size GPU | type Specs |
|---|---|---|
| 12 GB | Nvidia | Tesla K80 |

**TABLE 6.** Confusion matrix.

| | Man | Women |
|---|---|---|
| Predict. as man | True positive (TP) | False positive (TN) |
| Predict. as woman | False negative (FN) | True negative (TN) |

preprocessing/image/) tools to generate the training batch tensor of images. In particular, the following procedures were performed: width and height shifting with a probability of 0.1, image rotation into 45°, horizontal flipping with the nearest fill mode and image thumbnail with a target size of 100*100.

To experimentally evaluate the classification result, we used the confusion matrix shown in Table 6. The rows and columns describe, respectively, the predicted and original class for a given subject. The parameters TP and TN count the number of subjects that are correctly classified. The parameters FP and FN count the number of subjects that are misclassified.

The first metric is accuracy, which computes the percentage of the subjects correctly classified in the whole set of tested instances [66]. The accuracy formula is:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

The second metrics are the receiver operating characteristic (ROC) and the area under the curve (AUC). They are two equivalent metrics. The ROC allows a graphical interpretation of the evolution of the false acceptance rate (FAR) against the false rejection rate (FRR) for given threshold values. The AUC quantitatively estimates the surface under the ROC curve. It was theoretically and empirically proven that ROC and AUC are more powerful metrics for binary classifier performance evaluation [66]. The formulas for both FAR and FRR submetrics are:

$$FAR = \frac{FP}{FP + TN} \quad ; \quad FRR = \frac{FN}{TP + FN} \tag{4}$$

The last metric is the equal error rate (EER). It is a biometric system security metric [14]. It is a widely used metric. EER is used to predetermine the threshold, minimizing the FAR and FRR. As long as the error is minimized and close to zero, the system is safe. By obtaining the optimal values FAR opt and FRR opt to allow the minimization of the absolute difference between both rates, the EER is computed as follows [67]:

$$ERR = \frac{FAR_{opt} + FRR_{opt}}{2} \tag{5}$$

It can also be deduced graphically from the point where the FAR and FRR curves intersect.

## V. RESULTS, DISCUSSION AND BASELINE COMPARISON

In this section, the obtained results for the described metrics with implemented CNNs, VGG16, Inception V3, ResNet50 and MobileNet, are discussed, and a baseline comparison is performed.

### A. RESULTS DISCUSSION

- **Accuracy**

In the first part of the experimentation, we evaluate the accuracy values. Table 7 summarizes the obtained accuracy for gender prediction CNNs trained with fake data and tested

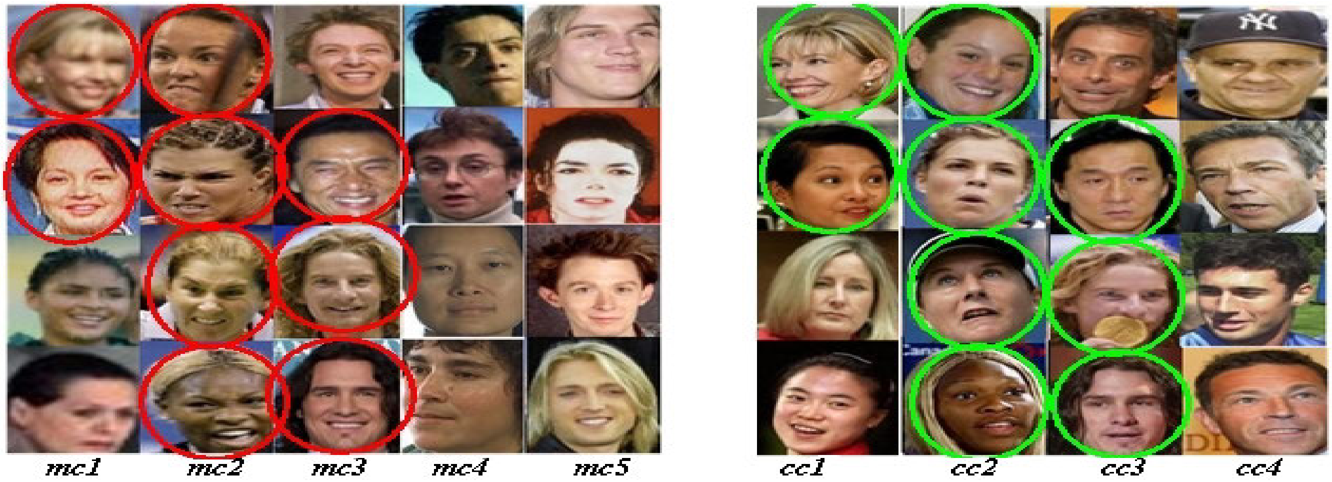


**FIGURE 10.** Example of misclassified ($mc_i$) and correctly classified ($cc_i$) subjects from the FERET and LFW datasets (mc1): female images with very poor quality, (mc2): female face with poor quality and aggressive reaction, (mc2): smiling male faces with poor quality, (mc4) challenging female subject, (mc5): challenging male faces, (cc1, cc2): correctly classified female, (cc3, cc4): correctly classified male.

on real faces. By analysing the obtained results, we qualified them as encouraging and promising. Acceptable accuracy was achieved for gender classification in various contexts. The best accuracy of 98.2 and 96.87 was obtained with VGG16 and ResNet tested on 6233 images for 1199 FERET subjects of various ages, races, illumination, face poses and facial expressions. In a perfectly gender-balanced context (100 subjects per gender) of white race with multiple variations in the face poses and facial expression, the best accuracies of 97.93 and 97.21 were obtained with VGG16 and ResNet, respectively, tested on the FEI dataset. Similarly, in a context with acceptable image quality containing multiple variations (occlusion especially), accuracies of 98.05 and 96.16 were obtained with VGG16 and Inception-V3 tested on 3488 faces of FRAV-database subjects. The result obtained with the balanced FEI and FRAV2D datasets shows that in an acceptable quality context, occlusion, illumination changes, facial expressions and face pose (frontal, semi-profile) do not affect the gender prediction accuracy performance. Table 7 also shows the accuracies with the challenging real-world dataset LFW (Face Labeled in the Wild). The best accuracies of 94.93 and 96.97 were returned for Inception-V3 and VGG16 tested on 13164 images for more than 5k subjects. For this last database containing images with poor quality, the augmentation of the face bounding box returned by the R-CNN face detector is an indispensable task. The worst accuracies of 92.55 and 92.90 were returned by MobileNet tested on occluded faces of the FRAV2D dataset and real-world faces of the LFW dataset.

In contrast, by looking at the misclassified subjects in previous datasets, we notice some factors affecting gender classification performance: quality degradation, facial expressions and semantically challenging subjects. Indeed, image quality is the factor that most affects performance. We justify this by the fact that in images with poor quality, much discriminative information is lost. The second remarkable factor is the facial expressions, where it can be observed that the trained CNNs will be more sensitive in front of facial expression variation in images of poor quality. Especially for images with poor quality, experiments show that smiling faces are reserved for the female gender, while aggressive faces are more reserved for the male gender. For the last factor, like any automatic classification system, naturally, there is an error merge reserved for more challenging subjects. In the case of gender attributes, there are semantically challenging human faces for which gender prediction is not as obvious for machines as for human beings. Fig. 10 illustrates some examples of correctly classified and misclassified subjects, such as the discussed scenarios (contoured faces).

**TABLE 7.** Obtained accuracies (in Perc.) for CNNs trained with fake data and tested on real data.

| | Inception-V3 | ResNet | VGG16 | MobileNet |
|---|---|---|---|---|
| FERET | 96.54 | 96.87 | 98.2 | 96.04 |
| FEI | 96.29 | 97.21 | 97.93 | 96.56 |
| FRAV2D | 96.16 | 94.44 | 98.05 | 92.55 |
| LFW | 94.93 | 94.23 | 96.97 | 92.90 |

- **Receiver operating characteristic and area under the curve**

To affirm the viewed accuracies, and compare the trained CNNs and more readable detection of the better classifier, we traced for each dataset the ROC curves of tested CNNs, and we also computed the AUC scores. Fig. 11 shows the ROC curves comparing the performance evolution for real-data-tested CNNs. Table 8 summarizes the corresponding probabilistic AUC scores for each test.

By analysing the ROC curves, we can observe their coherence with the viewed accuracy and confirm the obtained results. In the context without occlusion, as in the FERET and

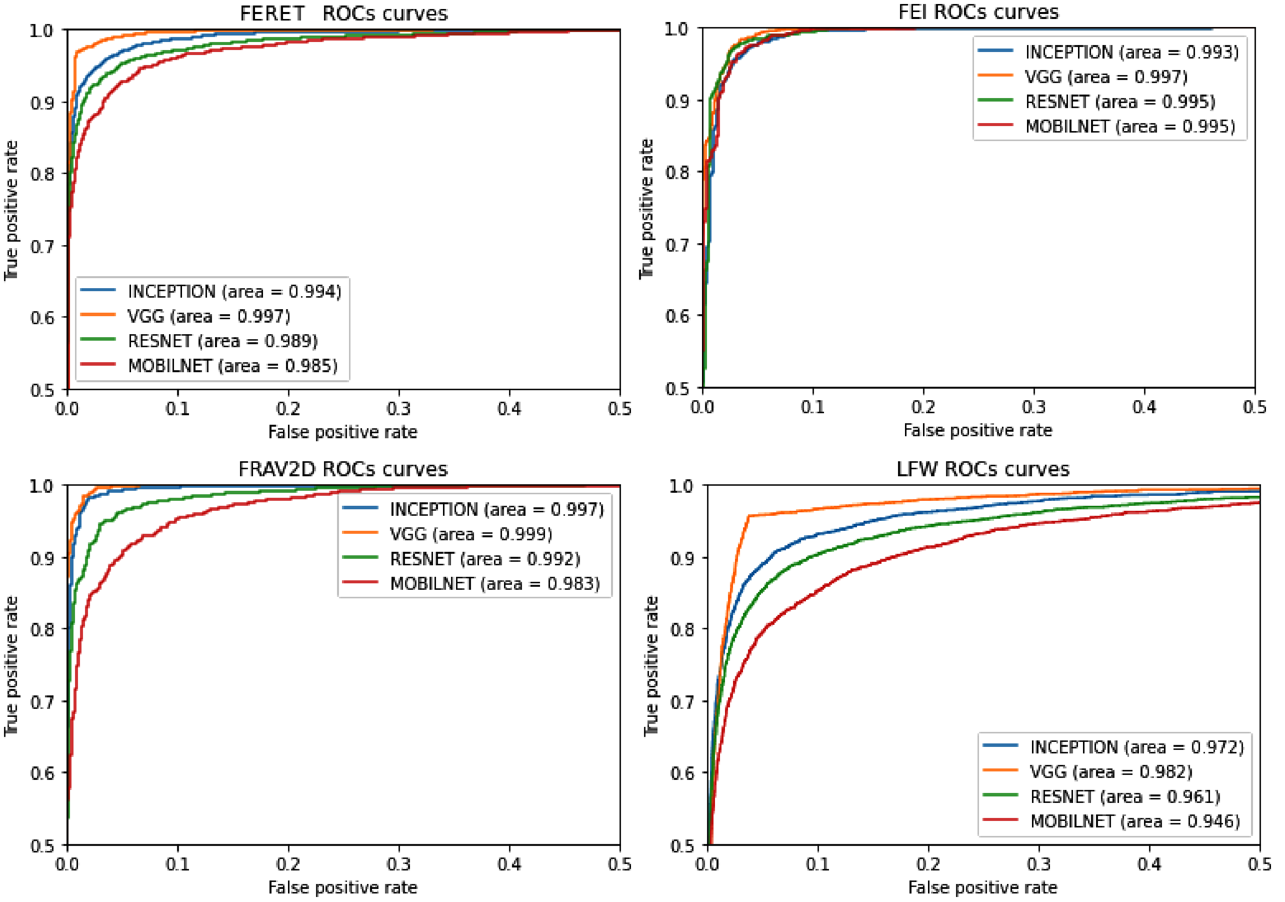


**FIGURE 11.** Comparison between the ROC curves of personalized CNNs (Inception-V3, VGG16, ResNet and MobileNet) trained with fake data and tested on the four real databases.

**TABLE 8.** Comparative table between the fake-data-trained CNNs by computing the AUC (as proba).

| | Inception-V3 | ResNet | VGG16 | MobileNet |
|---|---|---|---|---|
| FERET | 0.994 | 0.989 | 0.997 | 0.985 |
| FEI | 0.993 | 0.995 | 0.997 | 0.995 |
| FRAV2D | 0.997 | 0.992 | 0.999 | 0.983 |
| LFW | 0.972 | 0.961 | 0.982 | 0.946 |

FEI databases, all tested CNNs are more or less closers and converge normally as very good classifiers. The best AUC score of 0.997 was returned by VGG16 for both the FERET and FEI datasets, 0.012 higher for the worst AUC score obtained with MobileNet. For the occluded context presented in the FRAV dataset, the FRAV2D ROC curves clearly show that ResNet and MobileNet's performances decreased while VGG16 and Inception-V3 remained more stable in front of occluded faces. For challenging contexts and as shown in the last LFW ROC curves, VGG16 remains more stable in front-of-face images with poor quality and uncontrolled environments in comparison with other classifiers for which performances were remarkably degraded. The VGG16 allows an AUC score of 0.982, 0.01, higher than the second-best score obtained with Inception-V3.

- **Equal Error Rate**

In this subsection, the gender classification performances of fake-data-trained CNN classifiers are estimated in terms of the EER, which is a very interesting error for biometric system evaluation [67]. It was deduced after the determination of the FAR and FRR's optimal values by varying the threshold in a unit interval. Table 9 summarizes the obtained EERs for all performed tests. As seen in this last table, the obtained EERs for various contexts are reasonable. It allows for affirming the sufficient adjacency between the real testing data and artificial fake data exploited in training phases for the human gender prediction task. It also allows us to validate our proposed approach as a biometric task. The best EER of 0.021 for the FERET dataset obtained with VGG16 was 0.08 higher than the second EER returned with ResNet. For the FEI dataset, the two best closer EERs of 0.029 and 0.032 were returned by ResNet and VGG16, respectively. Similarly, for the FRAV2D dataset, the two best closer EERs of 0.019 and 0.022 were obtained with Inception-V3 and VGG16, respectively. The ResNet EER decreased with occluded data. For the uncontrolled context's

**TABLE 9.** Obtained ERRs.

| | Inception-V3 | ResNet | VGG16 | MobileNet |
|---|---|---|---|---|
| FERET | 0.031 | 0.029 | 0.021 | 0.041 |
| FEI | 0.038 | 0.029 | 0.032 | 0.036 |
| FRAV2D | 0.019 | 0.043 | 0.022 | 0.071 |
| LFW | 0.085 | 0.085 | 0.080 | 0.088 |

**TABLE 10.** Obtained GC accuracy (in Perc.) by training with StyleGAN, OpenForensics and StarGAN-v2 deepfakes datasets.

| Training<br>Test | StyleGAN (deepfake) | StarGAN-v2 (deepfake) | OpenForensics (deepfake) |
|---|---|---|---|
| FERET | 98,20 | 97.16 | 97.79 |
| FEI | 97.93 | 97.14 | 97.40 |
| FRAV2D | 98.05 | 96.96 | 97.60 |
| LFW | 96.97 | 96.10 | 97.11 |

LFW, the best EER of 0.08 is obtained with VGG16. The worst error rates were obtained with MobileNet.

Based on the discussed metrics ACC, ROC-AUC and EER, we notice in terms of classifier quality that globally, the best metric values in various contexts have been obtained by VGG16. It is the most adequate and efficient for accomplishing the gender classification task. Moreover, it seems to be affirmed that human gender prediction is an achievable task by exploiting the artificial faces generated by GAN. Especially for an acceptable quality context, GAN-generated faces can be exploited as an alternative for real data.

### B. STATE-OF-THE-ART AND BASELINE COMPARISON

- **Deepfakes dataset comparison**

After evaluating gender classification using StyleGAN-generated artificial faces and with the goal of performing facial gender classification by exploiting other deepfake faces, we also performed an experimental assessment of the VGG16-based classifier by training with other deepfake datasets. The first dataset was collected by exploiting the StarGAN-V2 Model [68]. This model allows human-face manipulation through image-to-image translations on multiple domains. As proposed by Yunjey et al., StarGAN-V2 was trained using the CelebA dataset. The second dataset is the OpenForensics deepfake dataset. It was collected by Le et al. [69] for face forgery detection and segmentation in the wild. Authors inspired by the StyleGAN model. Synthetic faces in the OpenForensics dataset were generated with uncontrolled real-world and challenging facial conditions, such as wild face poses, poor image quality, and occlusions. After we performed gender labeling for both StarGAN-V2 and OpenForensics deepfakes datasets, equitable sets of 60k images were exploited in the experiments.

Table 10 recapitulates the obtained accuracy as a percentage of the four private real datasets by training using StyleGAN, StartGan-v2 and OpenForensics deepfake. The achieved classification rates are comparable within the same order of magnitude. We notice that training with the StyleGAN artificial faces allows us to obtain results that exceed the case of using StarGAN-v2 faces. This can be justified by the fact that the StarGAN collected deepfakes dataset contains some disfigured synthetic faces. Except for the StarGAN-v2 dataset, the StyleGAN-generated faces are almost all realistic. For the OpenForensics dataset, we observe that its use as training data enhances the classification rate in an uncontrolled context. This is due to its richness in terms of real-world variations in comparison with the StyleGAN and StarGAN datasets, whose images have been generated with acceptable and high quality under condoled face poses and limited occlusion.

**TABLE 11.** Comparison of obtained results with reference ones of the state-of-the-art.

| Approach | Dataset | Acc | Test set size | Overlapping | Train data |
|---|---|---|---|---|---|
| Driven PCA [70] | FERET | 84.00 | Unclear | Eventual | Real |
| Van [18] | FERET | 97.30 | 2291 | Eventual | Real |
| Rai [34] | FERET | 98.18 | Unclear | Eventual | Real |
| **Proposed** | **FERET** | **98.2** | **6192** | **Avoided** | **Fake** |
| Tapia et al. [71] | FERET | 99,10 | - | Eventual | Real |
| Afifi [3] | FERET | 99.49 | 2286 | Eventual | Real |
| Khan [17] | FERET | 100 | 1412 | Avoided | Real |
| Khan [17] | FEI | 93.70 | 270 | Avoided | Real |
| Rai [34] | FEI | 96.61 | 200 | Eventual | Real |
| **Proposed** | **FEI** | **97.93** | **whole** | **Avoided** | **Fake** |
| Driven PCA [70] | FEI | 99.00 | Unclear | Eventual | Real |
| Geetha [36] | FEI | 99.00 | 100 | Eventual | Real |
| Sheikh [38] | FEI | 99.10 | Whole | No | Real |
| Rai et al. [34] | LFW | 88.34 | 6505 | Eventual | Real |
| FairFace [40] | LFW | 92.12 | Whole | No | Real |
| Khan [17] | LFW | 93.90 | 1323 | Avoided | Real |
| FaceTracer [72] | LFW | 94.00 | - | Eventual | Real |
| HyperFace [35] | LFW | 94.00 | whole | Avoided | Real |
| FaceHop [39] | LFW | 94.63 | 2647 | Eventual | Real |
| Afifi [3] | LFW | 95.98 | 10283 | Eventual | Real |
| VEGAC [33] | LFW | 96.80 | 3739 | Eventual | Real |
| **Proposed** | **LFW** | **96.97** | **13164** | **Avoided** | **Fake** |
| Sheikh [38] | LFW | 97.79 | 11483 | No | Real |
| Tapia [71] | LFW | 98.01 | - | Eventual | Real |
| Lee [9] | LFW | 98.45 | 11029 | Eventual | Real |
| Jia [10] | LFW | 98.69 | 13061 | Eventual | Real |
| PANDA [73] | LFW | 99.00 | Unclear | Eventual | Real |

- **Facial gender classification baseline comparison**

Table 11 presents a comparison between our proposed method for gender classification by using fake GAN faces and some reference works in the literature. The most commonly used databases in the literature are referenced in this table, especially the FERET, FEI and LFW databases. We performed our comparison on the common metric of accuracy describing the rate of correctly classified subjects from the test set. For each work, the comparison was performed by making the best gender prediction percentage obtained with the maximum test set for each database, the size of the used test set, the eventual data overlapping and the nature of exploited data for the training step. In addition, Fig. 12 illustrates a more readable baseline comparison for the challenging LFW dataset by grouping both the accuracy and size of the used test set as percentages.

By analysing the comparative table above and Fig. 12, it can be noticed that artificial faces generated by GAN allow obtaining results of the same order as those obtained

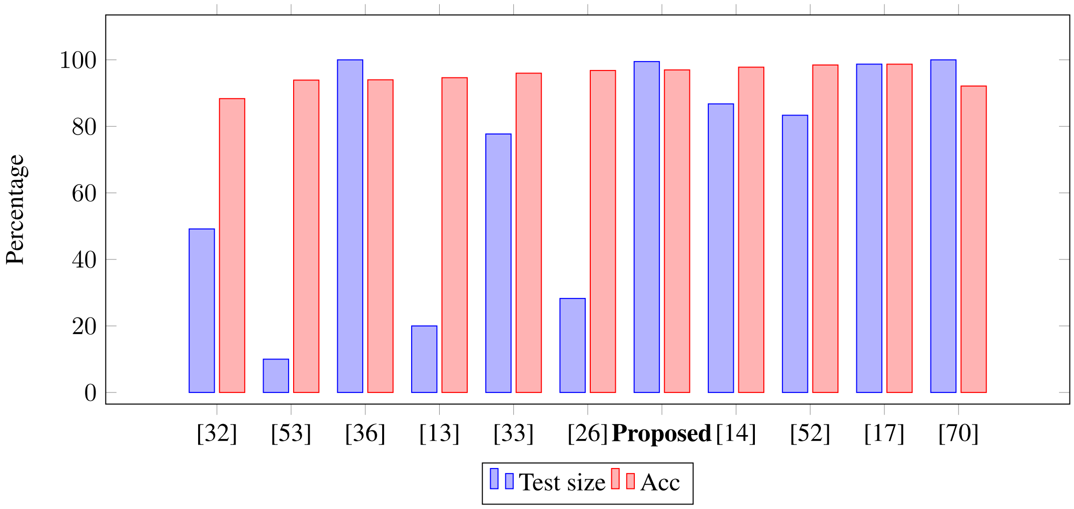


**FIGURE 12.** Baseline comparison on real-world LFW dataset.

by training with real data with the advantage of allowing tests on larger subsets. It should be noted that many existing approaches realize train-test data splitting with eventual major critical phenomena of data overlapping, as the faces of certain identities can be watched early by the neural network during its training. Additionally, there is no clear information allowing deducting the size of the used set at the test task.

Finally, we note that as a result of our work, a large facial gender-labeled deepfake dataset was collected. It will be made available upon request for scientific research uses concerning deepfake detection, human-face analysis and categorization.

## VI. CONCLUSION

In this paper, we proposed a deep learning-based approach for human gender classification from face modality. We investigated the use of convolutional neural networks by training with artificial GAN-generated faces. The trained CNNs were tested on multiple real datasets frequently used in the literature. CNN evaluations were performed with various solid metrics recommended for biometric systems. We obtained very encouraging and promising results validating our approach. The best performances were returned with the fake-trained VGG16-based classifiers. As an overall assessment, we assert that for the human gender classification task, the GAN-generated faces of artificial identities can be exploited as an alternative for real identity faces in training or both testing/training steps. Moreover, it allows more credibility in terms of people's privacy, which is indirectly violated in the state-of-the-art approaches and gives the opportunity to perform tests on a maximum subset of existing legal datasets. Finally, our goal for future work will be as follows:

- Feature visualization by exploiting hidden layer behaviours of fake-data-trained CNNs in front of real data.
- Fake-data-trained CNNs versus real-data-trained CNNs performances comparison for gender classification by realizing test on common context.
- Exploiting fake data to investigate other facial categorization tasks for more features such as age, race, facial expression and face accessories (glasses, beard, moustache, etc.).
- Investigate the human gender classification from full-body by employing deepfakes.

## REFERENCES

[1] S. Tarare, A. Anjikar, and H. Turkar, "Fingerprint based gender classification using DWT transform," in *Proc. Int. Conf. Comput. Commun. Control Autom.*, Pune, Feb. 2015, pp. 689–693.

[2] M. Afifi, "11K hands: Gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools Appl.*, vol. 78, no. 15, pp. 20835–20854, Aug. 2019, doi: 10.1007/S11042-019-7424-8.

[3] M. Afifi and A. Abdelhamed, "AFIF$^4$: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 77–86, Jul. 2019.

[4] D. Yaman, F. I. Eyiokur, N. Sezgin, and H. K. Ekenel, "Age and gender classification from ear images," in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Jun. 2018, pp. 1–7.

[5] S. Khellat-Kihel, J. Muhammad, Z. Sun, and M. Tistarelli, "Gender and ethnicity recognition based on visual attention-driven deep architectures," *J. Vis. Commun. Image Represent.*, vol. 88, Oct. 2022, Art. no. 103627.

[6] T. Roxo and H. Proença, "YinYang-net: Complementing face and body information for wild gender recognition," *IEEE Access*, vol. 10, pp. 28122–28132, 2022.

[7] M. Oulad-Kaddour, H. Haddadou, C. Conde, D. Palacios-Alonso, and E. Cabello, "Real-world human gender classification from oral region using convolutional neural netwrok," *ADCAIJ, Adv. Distrib. Comput. Artif. Intell. J.*, vol. 11, no. 3, pp. 249–261, Jan. 2023.

[8] B. Golomb, D. Lawrence, and T. Sejnowski, "SEXNET: A neural network identifies sex from human faces," in *Advances in Neural Information Processing Systems*, vol. 3, R. Lippmann, J. Moody, and D. Touretzky, Eds. Burlington, MA, USA: Morgan-Kaufmann, 1990. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1990/file/bbcbff5c1f1ded46c25d28119a85c6c2-Paper.pdf

[9] B. Lee, S. Z. Gilani, G. M. Hassan, and A. Mian, "Facial gender classification—Analysis using convolutional neural networks," in *Proc. Digital Image Comput., Techn. Appl. (DICTA)*, Dec. 2019, pp. 1–8.

[10] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Gender classification by deep learning on millions of weakly labelled images," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Barcelona, Dec. 2016, pp. 462–467.

[11] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, Jan. 2018.

[12] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, "A review of facial gender recognition," *Pattern Anal. Appl.*, vol. 18, no. 4, pp. 739–755, Jul. 2015.

[13] F. Lin, Y. Wu, Y. Zhuang, X. Long, and W. Xu, "Human gender classification: A review," *Int. J. Biometrics*, vol. 8, nos. 3–4, p. 275, 2016.

[14] M. El-Abed, C. Charrier, C. Rosenberger, M. El-Abed, C. Charrier, and C. Rosenberger, "Evaluation of biometric systems," in *New Trends and Developments in Biometrics*. London, U.K.: IntechOpen, Nov. 2012. [Online]. Available: https://www.intechopen.com/chapters/41062

[15] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 125–143, Jun. 2006.

[16] M. E. Abed, R. Giot, B. Hemery, and C. Rosenberger, "Evaluation of biometric systems: A study of users' acceptance and satisfaction," *Int. J. Biometrics*, vol. 4, no. 3, p. 265, 2012.

[17] K. Khan, M. Attique, I. Syed, and A. Gul, "Automatic gender classification through face segmentation," *Symmetry*, vol. 11, no. 6, p. 770, Jun. 2019.

[18] J. V. D. Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 188–195.

[19] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 34–42.

[20] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, "The devil of face recognition is in the noise," in *Proc. 15th Eur. Conf.* Munich, Germany: Springer, Sep. 2018, pp. 780–795.

[21] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2019, pp. 1–9.

[22] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440.

[23] S. M. Albright, "Mccloskeysource generator attribution via inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2019, pp. 96–103.

[24] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva, "SpoC: Spoofing camera fingerprints," 2019, *arXiv:1911.12069*.

[25] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4396–4405.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2019, pp. 8110–8119.

[28] (2020). [Online]. Available: https://www.thispersondoesnotexist.com

[29] S. J. Nightingale and H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 8, Feb. 2022, Art. no. e2120481119.

[30] L. Nataraj, T. M. Mohammed, B. S. Manjunath, S. Chandrasekaran, A. Flenner, J. H. Bappy, and A. K. Roy-Chowdhury, "Detecting GAN generated fake images using co-occurrence matrices," *Electron. Imag.*, vol. 31, no. 5, p. 532, Jan. 2019.

[31] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1038–1048, Aug. 2019.

[32] J. Lem, L. Abdul-WahidSami, A. W. BanikRazvan, and A. Andonie, "Comparison of recent machine learning techniques for gender recognition from facial images," in *Proc. 27th Modern Artif. Intell. Cogn. Sci. Conf.*, Dayton, OH, USA, 2016, pp. 1–6.

[33] A. Gurnani, K. Shah, V. Gajjar, V. Mavani, and Y. Khandhediya, "VEGAC: Visual saliency-based age, gender, and facial expression classification using convolutional neural networks," 2018, *arXiv:1803.05719*.

[34] P. Rai and P. Khanna, "An illumination, expression, and noise invariant gender classifier using two-directional 2DPCA on real Gabor space," *J. Vis. Lang. Comput.*, vol. 26, pp. 15–28, Feb. 2015.

[35] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[36] A. Geetha, M. Sundaram, and B. Vijayakumari, "Gender classification from face images by mixing the classifier outcome of prime, distinct descriptors," *Soft Comput.*, vol. 23, no. 8, pp. 2525–2535, Dec. 2018.

[37] L. E. Lin and C. H. Lin, "Data augmentation with occluded facial features for age and gender estimation," *IET Biometrics*, vol. 10, no. 6, pp. 640–653, Apr. 2021.

[38] M. S. Fathollahi and R. Heidari, "Gender classification from face images using central difference convolutional networks," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 4, pp. 695–703, Sep. 2022.

[39] M. Rouhsedaghat, Y. Wang, X. Ge, S. Hu, S. You, and C. C. J. Kuo, "Face-Hop: A light-weight low-resolution face gender classification method," in *Proc. Int. Workshops Challenges*. Cham, Switzerland: Springer, 2021, pp. 169–183.

[40] K. Karkkainen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.

[41] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[42] C. Conde, A. Serrano, and E. Cabello, "Multimodal 2D, 2.5D & 3D face verification," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 2061–2064.

[43] FEI, Centro Universitario da FEI. (Mar. 14, 2012). *Fei Face Database*. [Online]. Available: https://fei.edu.br/~cet/facedatabase.html

[44] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images Detection Alignment Recognit.*, Oct. 2008, pp. 1-15.

[45] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 256–263.

[46] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 341–345.

[47] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 707–711, May 2002.

[48] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[49] M. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nov. 2001, pp. 202–207.

[50] S. Tamura, H. Kawai, and H. Mitsumoto, "Male/female identification from 8 × 6 very low resolution face images by neural network," *Pattern Recognit.*, vol. 29, no. 2, pp. 331–335, Feb. 1996.

[51] T. Jabid, Md. H. Kabir, and O. Chae, "Gender classification using local directional pattern (LDP)," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2162–2165.

[52] S. Baluja and H. A. Rowley, "Boosting sex identification performance," *Int. J. Comput. Vis.*, vol. 71, no. 1, pp. 111–119, Jan. 2007, doi: 10.1007/s11263-006-8910-9.

[53] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," in *Proc. Class Project Stanford CS231N, Convolutional Neural Netw. Vis. Recognit., Winter Semester*, vol. 2014, no. 5, 2014, p. 2.

[54] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[55] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation," 2019, *arXiv:1904.11685*.

[56] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2000–2012, Aug. 2019.

[57] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," 2018, *arXiv:1811.08180*.

[58] X. Di, V. A. Sindagi, and V. M. Patel, "GP-GAN: Gender preserving GAN for synthesizing faces from landmarks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1079–1084.

[59] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proc. ICML*, 2018, pp. 1–39.

[60] S. Ramachandran and A. Rattani, "Deep generative views to mitigate gender classification bias across gender-race groups," 2022, *arXiv:2208.08382*.

[61] A. Dhillon and G. K. Verma, "Convolutional neural network: A review of models, methodologies and applications to object detection," *Prog. Artif. Intell.*, vol. 9, no. 2, pp. 85–112, Dec. 2019.

[62] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2016, pp. 650–657.

[63] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[64] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021.

[65] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[66] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining Knowl. Manage. Process*, vol. 5, no. 2, pp. 1–11, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:61877559

[67] R. Giot, M. El-Abed, and C. Rosenberger, "Fast computation of the performance evaluation of biometric systems: Application to multibiometrics," *Future Gener. Comput. Syst.*, vol. 29, no. 3, pp. 788–799, Mar. 2013.

[68] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2020, pp. 1–10.

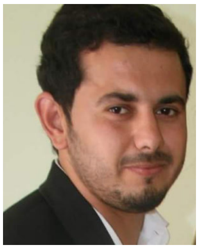
[69] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 1–11.

[70] C. Thomaz, G. Giraldi, J. Costa, and D. Gillies, "A priori-driven PCA," in *Computer Vision—ACCV* (Lecture Notes in Computer Science), J.-I. Park and J. Kim, Eds. Berlin, Germany: Springer, 2013, pp. 236–247.

[71] J. E. Tapia and C. A. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of LBP, intensity, and shape," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 488–499, Mar. 2013.

[72] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, 2008, pp. 340–353.

[73] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1637–1644.



**MOHAMED OULAD-KADDOUR** received the engineering and magister degrees from Ecole Nationale Supèrieure d'Informatique (ESI), Algiers, in 2011 and 2015, respectively, where he is currently pursuing the Ph.D. degree. He is an Assistant Professor with Ecole Nationale Supèrieure des Travaux Publics (ENSTP), Algiers. He is writing the Ph.D. in collaboration with the Face Recognition and Artificial Vision (FRAV) Research Group, Universidad Rey Juan Carlos, Madrid. His research interests include image classification, biometric categorization, machine learning, and image processing.

**HAMID HADDADOU** is currently a Professor with Ecole Nationale Supèrieure d'Informatique (ESI), Algiers. He is also the Head of the Applied Mathematics Team, Computer Systems Communication Laboratory (LCSI). His research interests include biometrics, image processing, optimization, and multiscale mathematics modeling.

**CRISTINA CONDE VILDA** received the B.S. degree in physics (electronics) from the Complutense University of Madrid, in 1999, and the Ph.D. degree from Universidad Rey Juan Carlos, Madrid, in 2006. She has worked in the private sector for several years. In 2001, she joined Universidad Rey Juan Carlos, as an Assistant Professor. For seven years, she was the Vice Dean of Studies with the Computer Science School. She is currently a Full Professor. She has coordinated several national and European projects. Her research interests include image and video analysis, pattern recognition, and machine learning in both classical and biologically inspired computation.

**DANIEL PALACIOS-ALONSO** was born in Madrid, Spain. He received the B.S. and M.S. degrees in computer science and the Ph.D. degree in advanced computation from Universidad Politécnica de Madrid (UPM), in 2009 and 2017, respectively. He was a Team Leader at a technological consulting firm for five years. Since 2013, he has been a member of the Neuromrphic Speech Processing Laboratory, Center for Biomedical Technology. He is currently an Associate Professor with Universidad Rey Juan Carlos (URJC). He is also the Head of the Bioinspired Systems and Applications Group (SA-BIO). His research interests include stress and emotional states, neurodegenerative diseases, such as Parkinson's, ALS, and Alzheimer's, among others, artificial vision, pattern recognition, and biomedical signal processing. He was a recipient of several best paper awards, including ICPRS 2016, BIOSIGNALS 2019, and JID 2020, and the Doctoral Consortium Award from the Spanish Association of Artificial Intelligence, in 2013. He is a reviewer of national and international journal articles.

**KARIMA BENATCHBA** has been a member of the Higher School of Computer Science of Algiers for more than 20 years. She is currently a Full Professor with Ecole Nationale Supèrieure d'Informatique (ESI), Algiers. She is also the Head of Laboratory des Méthodes de Conception de Systémes (LMCI), ESI, in which she coordinates the optimization research group. Her research interests include optimization methods, artificial intelligence, image segmentation, machine learning, and data mining.

**ENRIQUE CABELLO** (Member, IEEE) received the B.S. degree in physics (electronics) from the University of Salamanca and the Ph.D. degree from the Polytechnic University of Madrid. In 1990, he joined the Computer Science Department, University of Salamanca. He joined Universidad Rey Juan Carlos, in 1998, where he has been the Head of the Face Recognition and Artificial Vision Group, since 2001. He is currently a Full Professor. His research interests include image and video analysis, pattern recognition, and machine learning using classic and bioinspired approaches.