Comparative Analysis of Artificial Intelligence Models for Facial Image Modification Detection

Diana Levshun, Dmitry Levshun
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
{diana.levshun, levshun}@comsec.spb.ru

***Abstract*—Challenges associated with detecting facial image modifications have become even more significant due to advances in photo editing software and image generation techniques. This paper presents a comprehensive comparative analysis of artificial intelligence models for detecting facial image modifications, including retouching, makeup, and digital manipulation. The research examines a broad spectrum of approaches, including traditional hand-crafted facial image features alongside deep features that are extracted through convolutional neural networks. It also explores different detection techniques, ranging from classical machine learning algorithms to advanced deep neural network architectures. Additionally, the study considers the use of principal component analysis for feature dimensionality reduction and its impact on detection. The experimental results demonstrate the advantages of ensemble methods and deep features in improving the accuracy of facial modification detection. The findings have practical implications for developing image authentication and fake detection systems in the field of digital forensics and visual content quality control.**

I. Introduction

Photo editing and the use of filters have become widespread, particularly in social media posts. Retouching and beauty filters enable users to smooth skin texture, eliminate blemishes, and much more. Platforms like Instagram, TikTok, Snapchat, and others dedicate substantial resources to developing features that assist users in creating visually attractive content. Additionally, users utilize augmented reality (AR) filters and artificial intelligence (AI) technologies for photo enhancement.

We can note that over 50-70% of young people report using AR beauty filters and applications to enhance their photos [1], [2]. While many individuals enjoy the creative expression that filters provide, there is a growing concern about the long-term effects on mental health and self-esteem [3], [4]. Continuous exposure to beautified images can create a skewed perception of beauty, making unfiltered appearances seem less acceptable.

Concern on the part of governments has led to a number of countries now having laws requiring digitally retouched photographs to be labeled as “edited photographs” or “retouched image”. For example, in Israel, Norway, and France, any commercial image that has been digitally altered must carry a similar warning [5]. The advertising industry is also introducing a number of bans on the use of filters to alter faces. For example, in 2020, the UK Advertising Standards Authority (ASA) introduced a rule requiring public figures to disclose the filters or retouching used when advertising cosmetic products [6]. In a number of other countries, there are currently ongoing discussions on defining responsible behavior related to the digital modification of images on media.

Also, any user on social networks can appear as a completely different person by changing not only their age but also their gender in a photo. This becomes fertile ground for online fraud. Concerns about filters are also relevant for users of online dating services. These scams often involve creating deceptive profiles that misrepresent individuals’ true appearances and identities, making it easier for fraudsters to manipulate victims emotionally and financially [7].

Facial modifications can also compromise user privacy when used to trick AI models and facial recognition tools [8]. For example, distorting facial landmarks makes deepfakes more difficult to detect, making it easier for attackers to use these technologies for unethical purposes. Filters can significantly alter biometric features, complicating tasks like face recognition and classification. Faking images creates significant security risks for biometric authentication systems [9], [10].

Thus, while facial modification technologies offer creative and privacy-preserving opportunities, they also present significant risks that need to be addressed. The development of robust detection methods and ethical guidelines is crucial to mitigate these dangers and ensure responsible use.

Detecting facial retouching poses distinct challenges compared to other types of image manipulation because beautification is typically subtle and intended to look natural. Moreover, the wide variety of retouching techniques and tools makes detection more complex. AI-based detection methods have become the leading approach for identifying these manipulations, as they can learn the statistical patterns and artifacts introduced during the retouching process [11]–[25].

We note that we are only considering facial modifications in the digital realm, and do not take into account the application of physical makeup or plastic surgery.

In this paper, we propose to compare different artificial intelligence models for detecting facial modification in photos. To do this, we test both individual machine learning and deep learning models and their ensembles on the MakeupWild [26] and B-LFW [27] datasets. Comparison of different types of features and classifiers allows us to determine their performance and identify the most effective combinations that are good at detecting modifications in data from different sources. Classifier ensembles, in turn, combine the strengths of individual models and reduce the likelihood of misclassification.

Our contributions are as follows:

- **Comprehensive comparison**: The study provides a large-scale comparative analysis of many facial modification detection models, incorporating a variety of feature extraction and classification methods. Unlike approaches that focus on either a single feature type or a single classifier, our study systematically considers and tests multiple approaches in a single experimental environment.
- **Impact of feature space dimensionality**: We conduct an experiment by applying principal component analysis (PCA) to different feature sets to study how dimensionality reduction affects facial modification detection.
- **Incorporation of ensemble methods**: We use model ensembles to analyze how they affect the accuracy of detection compared to individual classifiers.
- **Advances in digital forensics**: Our study expands the knowledge base in image authentication and manipulation detection by providing a comprehensive analysis of state-of-the-art methods and datasets, which contributes to improved defense against deepfakes.

The paper is organized as follows. Section II reviews related work and datasets. Section III proposes a methodology for analyzing AI models for facial image modification detection. Section IV presents experiments to evaluate the performance of the models. Section V includes the discussion. Section VI concludes the study.

II. Related Works

A. Modification Detection

In the field of artificial intelligence, facial image manipulation detection is typically considered a binary classification problem. Trained models are used to classify between original images and manipulated ones. Feature extraction involves transforming raw face images into a reduced feature space that preserves the most discriminatory information.

Early efforts to detect facial retouching and beautification relied on conventional (traditional) machine learning techniques, which extracted hand-crafted features from images before classification. One of the most popular methods in this area is the *support vector machine* (SVM). Bharati et al. [11] use SVMs in combination with restricted Boltzmann machine (RBM) to classify images as original or retouched. Rathgeb et al. present an SVM-based retouch detection system that analyzes spatial and spectral features extracted using Photoresponse Numerical Non-Uniformity Analysis (PRNU) [17] and texture descriptors, facial landmarks, and deep face representations [18]. Local binary patterns (LBP) are extracted as texture features. While these traditional machine learning methods provided foundational approaches to retouching detection, they have gradually been superseded by deep learning techniques that can automatically learn relevant features from data. Rasti et al. [14] extract makeup color features such as average skin tone (AST) and texture using histogram of directional gradient (HOG). The authors also use principal component analysis before classification using SVM.

Convolutional neural networks (CNN) are the most commonly adopted for extracting deep image features. CNNs offer significant advantages in detecting facial retouching compared to earlier methods, primarily due to their ability to learn complex features and patterns from images. Akhtar et al. [19] provide a comparative analysis of widely used CNN architectures such as VGG16, SqueezNet, DenseNet, ResNet, and GoogleNet for facial manipulation detection. The authors also analyze which manipulations are the least and most difficult to detect: changes in hairstyle, tattoos, and glasses are easy to spot, while age manipulations and facial feature corrections are more difficult to detect.

Together with an SVM classifier, CNNs models are used in Jain et al. [13], Kotwal et al. [15], Rathgeb et al. [18], Hedman et al. [21] and Sharma et al. [23]. The authors also use well-known convolutional network architectures for face recognition, such as LigthCNN, ResNet50 and others. Also, Hedman et al. [21] use the XGBoost classifier. Typically, such deep neural networks are pre-trained for face recognition, for example, on ImageNet [28] or VGGFace2 [29] datasets.

Also, a fully connected layer (FCL) of a convolutional neural network can be used as a classifier. Thus, Wang et al. [16] present an approach to detect facial deformation manipulations performed with Adobe Photoshop using a Dilated Residual Network variant (DRN-C-26). Majumdar et al. [22] discuss the use of deep learning models, specifically ResNet50 and XceptionNet, for detecting retouched and altered facial images. Sheth and Vora [24], [25] apply ResNet50 and VGG16 models with transfer learning to detect facial retouching.

CNNs can also be the basis for two neural network architectures, the *autoencoder* (AE) and the *generative adversarial network* (GAN). The first is typically used for image reconstruction and data compression, while the second is used for generating new images. For example, Bharati et al. [12] use a sparse autoencoder for feature extraction in conjunction with an SVM classifier for demography-based retouch detection. Alzahrani et al. [20] use a convolutional autoencoder to detect makeup.

B. Image Datasets

Sources of modified image data may also include studies aimed at the task of recognizing faces using retouching and makeup [27], [30] and the problem of makeup transfer [26], [31] or transfer of face attributes [13], [32]. Their authors also use various programs and generative models to create images.

As input for modification, researchers often use publicly available face image datasets, such as the Collection B of Notre Dame database (ND-Collection B) [33], CASIA Near infrared vs. Visible light (NIR-VIS) 2.0 face database [34], cross-spectral cross-resolution video dataset (CSCRV) [35], CelebFaces Attributes (CelebA) dataset [36], Open Images dataset [37], Face Recognition Technology (FERET) database [38], Face Recognition Grand Challenge (FRGCv2) dataset [39], Milborrow/University of Cape Town (MUCT) face database [40], FairFace dataset [41], Labeled Faces in the Wild (LFW) database [42] and others.

TABLE I. FACIAL IMAGE MODIFICATION DATASETS

Dataset	Image Count	Modifictaion Type: Makeup	Beauty	Attributes	AR-filers	Obtaining	Tool	Initial data	Access
ND-IIITD Retouched Face Database [11]	4875	✓	✓			Manual	PortraitPro Studio Max	ND-Collection B ND-Collection B,	On request
Multi-Demographic Retouched Faces (MDRF) [12]	3600	✓	✓			Manual	BeautyPlus, PortraitPro Studio Max	CASIA-NIR-vsVIS v2, CSCRV	On request
Dataset by Jain et al. [13]	18 000			✓		Auto	StarGAN	CelebA	Private
Dataset by Wang et al. [16]	1110000		✓	✓		Auto	Adobe Photoshop	Open Images	Open
Datasets by Rathgeb et al. [17], [18]	147712	✓				Manual	AirBrush, FotoRus, InstaBeauty,	FERET	Private
	144032	✓				Manual	Polarr, YouCam Perfect	FRGCv2	Private
Deepfake MUCT dataset [19]	9608			✓		Manual	FaceApp	MUCT	Private
Makeup Wild [26]	3834	✓				Auto	PSGAN	Own	Open
FairBeauty [27]	108501	✓			✓	Auto	OpenFilter	FairFace	Open
B-LFW [27]	13000	✓			✓	Auto	OpenFilter	LFW	Open
LFW-Beautified [21]	34592	✓			✓	Auto	Instagram API	LFW	Open
CelebAMask-HQ [32]	30000			✓		Auto	MaskGAN	CelebA	On request

TABLE II. AI MODELS FOR DETECTING FACIAL IMAGE MODIFICATIONS

Ref.	Year	Feature Extraction	Manipulation Detection	Pre-training Data	Experimental Data	Accuracy
[11]	2016	RBM	SVM		ND-IIITD	87.10%
[12]	2017	AE	SVM		MDRF	95.00%
[13]	2018	CNN	SVM		ND-IIITD	99.65%
					CelebA based	99.73%
[14]	2018	AST, HOG, PCA	SVM		YMU	97.50%
[15]	2019	LigthCNN	SVM		AIM	93.35%
					YMU	93.88%
					MIW	96.10%
					MIFS	93.27%
[16]	2019	DRN-C-26	FCL	ImageNet	Open Images based	99.80%
[17]	2020	PRNU	SVM		FRGCv2 based	86.30%
[18]	2020	LBP, ResNet50	SVM		FRGCv2 based	88.29%
					FERET based	91.84%
[19]	2020	CNN	FCL		MUCT	96.25%
		VGG16	FCL	ImageNet		94.98%
		SqueezeNet	FCL			97.33%
		DenseNet	FCL			99.42%
		GoogLeNet	FCL			92.17%
		ResNet183	FCL			99.33%
[20]	2021	VGG17	AE	ImageNet	YMU	88.33%
[21]	2022	ResNet34, ResNet50, SqueezeNet	SVM, XGBoost	CelebA, VGGFace2	LFW-Beautified	95.30%
[22]	2022	ResNet50	FCL	ImageNet	ND-IIITD	50.21%
		XceptionNet	FCL			56.22%
[23]	2023	CNN	SVM		ND-IIITD	99.84%
					YMU	83.70%
[24]	2024	ResNet50	FCL	ImageNet	ND-IIITD	98.52%
[25]	2024	VGG16	FCL	ImageNet	ND-IIITD	98.08%

Separately, we can highlight such datasets as Age Induced Makeup (AIM), YouTube Makeup dataset (YMU), Makeup in the Wild (MIW) and Makeup Induced Face Spoofing (MIFS) [15]. These datasets were obtained by extracting images from videos with manual makeup application and from makeup images from the Internet.

Table I contains the description of the most popular datasets of modified facial images. We define the size of the dataset, the type of modification (makeup, beautification, attribute manipulation, or AR-filters), the method and tool of obtaining, the initial data, and the access to the dataset.

Table II describes AI models from relevant research. We present face feature extraction methods, AI models for modification detection, a dataset for pretraining these models, a dataset for testing the models, and the detection accuracy.

In this study, we propose the broader analysis of facial image modification detection models, including ensemble models. Unlike prior works that focus on a single type of feature (e.g., only deep features or hand-crafted features), this study systematically combines diverse feature representations: LBP, HOG, facial landmark extraction, and deep features. We also consider the role of PCA in modification detection. As classifiers, in addition to SVM, we also propose to use random forest (RF) and gradient boosting (GB). For deep features, we use a fully connected layer of the AI model.

III. METHODOLOGY

Fig. 1 shows the proposed scheme of AI model analysis for facial image modification detection. Explanations for the symbols in the scheme are given below in the text. The analysis involves several key steps: (1) image preprocessing; (2) feature extraction; (3) principal component analysis; (4) AI model training; (5) image modification detection.

A. Image Preprocessing

Let us denote the input array of facial images as $X_n = \{x_1, x_2, ..., x_n\}$, where n is the number of images, x is the individual image.

Each image has a width w and a height h. By preprocessing, the input image is a numerical array: $x_{w \times h}$. Image preprocessing typically includes color transformations, normalization, resizing, cropping, noise removal, and others. These transformations can be combined depending on the task and the type of model used.

B. Feature Extraction

There are many methods for extracting features from a face image [43], [44]. Among the main groups of feature extraction techniques, we can distinguish:

1) *geometric-based* (encode the shape and spatial relationships between facial components such as the eyes, nose, and mouth by measuring distances and angles);
2) *appearance-based* (focus on the overall visual properties of the face) like LBP and HOG;
3) *deep learning-based* (automatically learn the most discriminative features directly from the data);
4) *hybrid* (combine the strengths of different feature extraction methodologies).

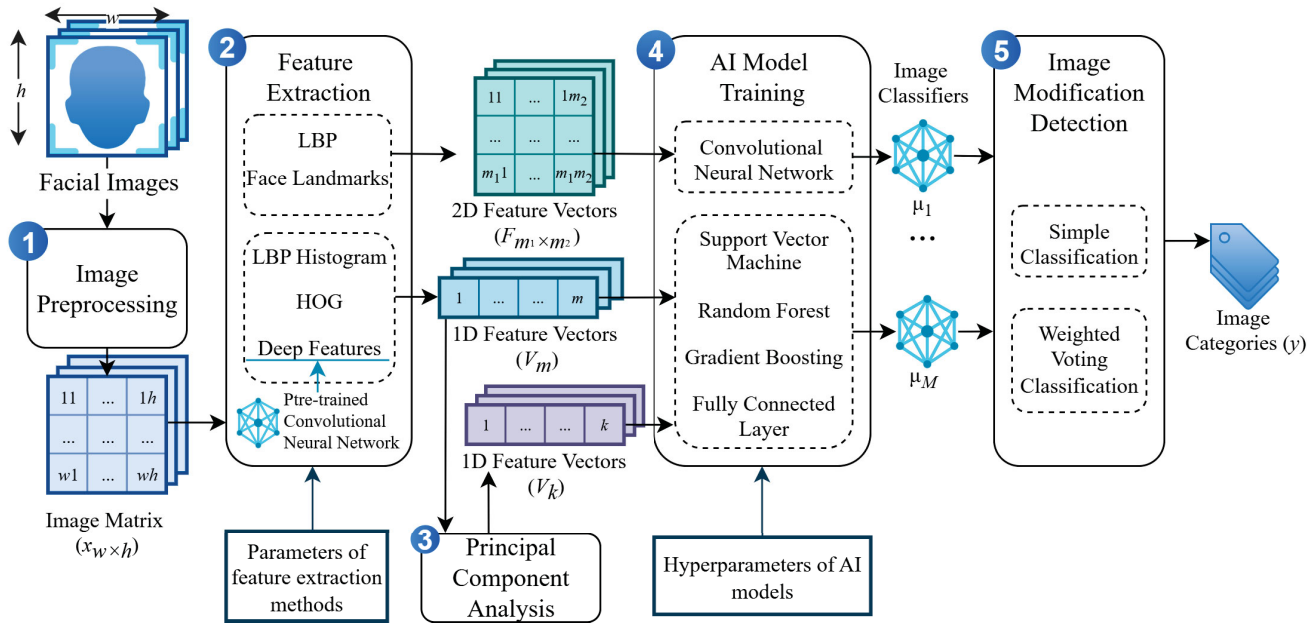


Fig. 1. AI model analysis for facial image modification detection

Using the feature extraction method ϕ allows us to transform the image matrix into a feature vector V_m (m is the length of the vector) or a feature matrix $F_{m_1 \times m_2}$ (m_1 is the number of rows of the matrix, m_2 is the number of columns).

Then the resulting array of facial images can be transformed into a two-dimensional array:

$$\phi : X_n \to V_{n \times m},$$

or a three-dimensional array:

$$\phi : X_n \to F_{n \times m_1 \times m_2}.$$

Based on the conducted research review, we propose the methods described below for extracting image features.

To obtain two-dimensional (2D) feature vectors, we use:

- *local binary patterns* (LBP) to determine the texture:

$$\phi^{LBP}(x_{w \times h}, p, r) = F^{LBP}_{w \times h}, \quad (1)$$

where p is the number of circularly symmetric neighboring given points, r is the circle radius, $m_1 = w$, $m_2 = h$;

- 3D *face landmarks* (FL):

$$\phi^{FL}(x_{w \times h}, L) = F^{LMarks}_{l \times 3}, \quad (2)$$

where L are parameters of the facial landmark detection detector, l is the number of landmarks, $m_1 = l$, $m_2 = 3$.

To obtain one-dimensional (1D) feature vectors, we use:

- *local binary patterns histogram* (LBPH) to determine the texture distribution of a face image:

$$\phi^{LBPH}(F^{LBP}_{w \times h}, d) = V^{LBPH}_m, \quad (3)$$

where d is the number of histogram cells, $m = d$;

- *histogram of oriented gradients* (HOG) descriptor to describe the distribution of intensity gradients:

$$\phi^{HOG}(x_{w \times h}, b, c, cc) = V^{HOG}_m, \quad (4)$$

where b is the number of orientation cells, c is the cell size, cc is the number of cells in each block, $m = w \cdot h$;

- *deep features* of face images:

$$\phi^{CNN}(x_{w \times h}, HP) = V^{CNN}_m, \quad (5)$$

where HP are the CNN parameters, m is the size of the last convolutional layer (the number of output features).

C. Principal Component Analysis

One-dimensional image features can be furthered preprocessed using principal component analysis (PCA):

$$pca(V_{n \times m}, k) = V_{n \times k}, \quad (6)$$

where k is the number of components to keep.

Principal components capture the maximum variance in the data, which often reduces noise and redundancy while retaining most of the important information. Using PCA can help us potentially improve the performance of the model by removing correlated or less informative features.

D. AI Model Training

Next, we pass the resulting array of features to the input of a classifier based on machine learning. The purpose of classification is to predict the label y:

$$\mu(V_m, HP) = y, \quad (7)$$

$$\mu(F_{m_1 \times m_2}, HP) = y, \quad (8)$$

where HP are the classifier parameters.

Classifiers can be either traditional machine learning models such as support vector machine (SVM), random forest (RF), gradient boosting (GB) or deep models such as multilayer perceptron (MLP). In the case of CNN, the classifier can be a fully connected layer of a neural network.

E. Image Modification Detection

In case of binary classification, the label determines whether the face in the image is modified or not. In case of multiclassification, the specific type of modification (retouching, makeup, etc.) is also determined. Let us denote the set S of known categories of face images when modified, including originals: $S = \{s_1, ..., s_z\},\ y \in S$, where z is the number of categories.

In ensemble classification, the results for each individual model are fed into a decision module that determines the final category of the image. We propose using weighted voting as a decision-making method. The weight of a particular model is determined during the training and testing phases.

The weighted vote count for class $s_i \in S$ is:

$$W(s_i) = \sum_{j=1}^{M} w_j \cdot \mathbf{1}[\mu_j = s_i], \tag{9}$$

where M is the number of classification models, w_j is the weight of the j-th classification model, $\mathbf{1}[]$ is the indicator function (1 if the condition is true, 0 otherwise).

The final predicted image class is the one with the highest weighted vote:

$$y = \arg\max_{s \in S} V_c(x). \tag{10}$$

The evaluation of models in the ensemble is based on such quality metrics of face modification detection as accuracy (*ACC*), precision (*P*), recall (*R*) and F-measure (*F*1).

To identify significant differences in the obtained metrics, the Friedman test is used. For a model, we rank the results across different data sets, then calculate the Friedman statistic (Q), which is compared with the critical value of the Chi-square distribution (χ^2): if Q is higher, then the models are not equal to each other. We also calculate the p-value as $\mathbb{P}(\chi^2 \leq Q)$, which is compared with the critical significance level α equal to 0.05.

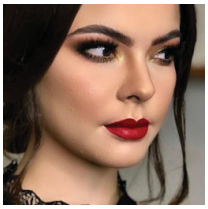
IV. Experiments

A. Datasets

As experimental facial image datasets, we use MakeupWild [26] and B-LFW [27].

MakeupWild dataset is created using the PSGAN makeup transfer generative model and contains 384 images of faces with makeup and 334 images without makeup. The image size is 256×256. Fig. 5 shows an example of makeup and no makeup images from this dataset.

B-LFW is a beautified version of the LFW dataset, designed to study and evaluate facial recognition systems. Modifications were made using OpenFilter, a framework for applying AR filters available on social media platforms. We use both original LFW images and modified B-LFW images in the ratio of 52:48, respectively. We refer to this entire dataset as B-LFW in the text below. The number of images is 25233, the size is 112×112. Fig. 5 shows an original and AR filter-changed images from this dataset.



(a) Makeup image

(b) No makeup image

Fig. 2. Example images from the MakeupWild dataset

(a) Orginal image

(b) AR filter image

Fig. 3. Example images from the B-LFW dataset

B. AI Models

To implement the proposed model analysis, we create a software prototype in Python 3.11 using keras, mediapipe, scikit-image, scikit-learn, and other libraries.

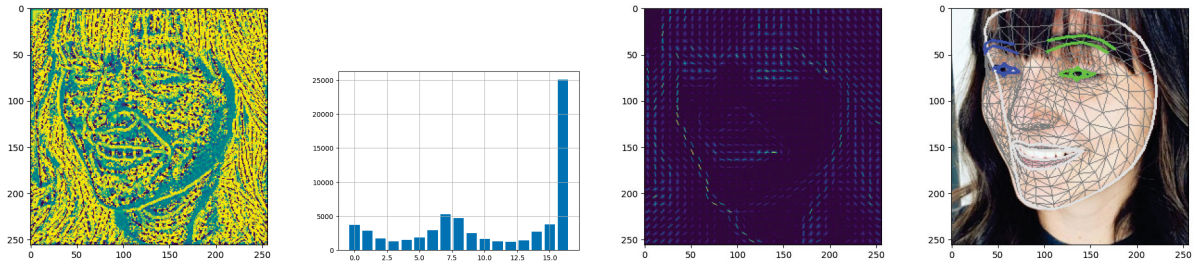
We need to note that the hyperparameters for the models listed below are selected based on theoretical analysis of relevant works, including publications related to the models. In this paper, we do not consider the hyperparameter optimization as it's beyond the scope of the current study, and will be a continuation based on the current experiments.

For LBP and HOG extraction, we use the skimage library. The methods are run with the following parameters – LBP: p = 15, r = 3; LBPH: p = 15, r = 3, d = 17; HOG: b = 9, c = 8, cc = 3. Thus, for the LBP histogram, PCA is applied with the parameter number of components $k = 15$, and for the HOG vector $k = 100$.

For facial landmarks, we use the MediaPipe library. Face landmarker uses the model FaceMesh [45], which outputs an estimate of 478 3D facial landmarks. The architecture of this model includes a MobileNetV2 convolutional neural network with custom blocks for real-time processing.

Fig. 4 shows examples of feature extraction from a facial image in the MakeupWild dataset.

Deep features are extracted using convolutional neural network models pretrained on the ImageNet dataset and provided in the keras applications library. For comparison, we selected the Xception [46], EfficientNetB3 (EffNetB3) [47], EfficientNetV2B3 (EffNetV2B3) and EfficientNetV2S (EffNetV2S) [48] models.



(a) LBP (b) LBPH (c) HOG image (d) Face landmarks

Fig. 4. Feature extraction examples

TABLE III. AI MODELS FOR EXPERIMENTS

№	Name	Feature Extraction								PCA	Classification					
		LBPH	LBP	HOG	FL	Xception	EffNetB3	EffNetV2B3	EffNetV2S		SVM	RF	GB	MLP	CNN	FCL
1	LBPH-SVM	✓									✓					
2	LBPH-PCA-SVM	✓								✓	✓					
3	LBPH-RF	✓										✓				
4	LBPH-PCA-RF	✓								✓		✓				
5	LBPH-GB	✓											✓			
6	LBPH-PCA-GB	✓								✓			✓			
7	LBPH-MLP	✓												✓		
8	LBPH-PCA-MLP	✓								✓				✓		
9	LBP-CNN		✓												✓	
10	HOG-SVM			✓							✓					
11	HOG-PCA-SVM			✓						✓	✓					
12	HOG-RF			✓								✓				
13	HOG-PCA-RF			✓						✓		✓	✓			
14	HOG-GB			✓									✓			
15	HOG-PCA-GB			✓						✓				✓		
16	HOG-MLP			✓										✓		
17	HOG-PCA-MLP			✓						✓						
18	FaceMesh-CNN				✓										✓	
19	Xception					✓										✓
20	EffNetB3						✓									✓
21	EffNetV2B3							✓								✓
22	EffNetV2S								✓							✓
23	ens-EffNet						✓	✓	✓							✓
24	ens-Model		✓		✓				✓						✓	✓

For classification models based on traditional machine learning, we use the following hyperparameters:

- SVM: regularization parameter C = 1, *kernel* = *'rbf'*, *degree* = 3, kernel coefficient *gamma* = *'scale'*;
- RF: the number of trees *n_estimators* = 100, *min_samples_split* = 2, split *criterion* = *'gini'*;
- GB: *learning_rate* = 0.1, the number of boosting stages to perform *n_estimators* = 100, *max_depth* = 3.

Hyperparameters of the MLP model for 1D features:

- number of layers: *n_layers* = 2;
- number of units on layers: *units* = [512, 256];
- hidden activation function: ReLu;
- output activation function: sigmoid.

Hyperparameters of the CNN model for 2D features:

- number of convolutional blocks: *n_blocks* = 3;
- number of units on layers: *conv_units* = [256, 128, 64];
- number of units on fully-connected layers: *fcl_units* = 64;
- size of a convolutional filter: *kernel_size* = 3;
- hidden activation function: ReLu;
- output activation function: sigmoid.

The fully connected layer for pretrained models includes *units* = 512. The number of training epochs for the MLP model is 100, the batch size is 128. The number of training epochs for all CNN models is 100, and the batch size is 32.

C. Results

Table III contains the description of 24 obtained AI models. Model ensembles are denoted as "ens-". Model 23 combines pre-trained EfficientNets models. Model 24 merges LBP-CNN, FaceMesh-CNN and modEfficientNetV2S models.

Table IV contains the results of evaluating the effectiveness of AI models for detecting image modifications for the MakeupWild dataset, and Table V contains the results for the B-LFW dataset. The best results are highlighted in bold.

Fig. 5 shows the accuracy of modification detection on the MakeupWild dataset, and Fig. 6 – on the B-LFW dataset. In these figures, the accuracy values for the models are ranked from highest to lowest to visualize better results.

Table VI shows the results of the Friedman test calculations for the obtained metrics. The number of degrees of freedom for χ^2 is 23 (one less than the number of models).

TABLE IV. PERFORMANCE OF IMAGE MODIFICATION DETECTION ON MAKEUPWILD

№	Name	*ACC*	*P*	*R*	*F1*
1	LBPH-SVM	0.792	0.842	0.780	0.810
2	LBPH-PCA-SVM	0.806	0.827	0.744	0.783
3	LBPH-RF	0.833	0.914	0.780	0.842
4	LBPH-PCA-RF	0.819	0.914	0.744	0.820
5	LBPH-GB	0.812	0.899	0.756	0.821
6	LBPH-PCA-GB	0.833	0.903	0.793	0.844
7	LBPH-MLP	0.688	0.66	0.886	0.757
8	LBPH-PCA-MLP	0.882	0.853	0.892	0.872
9	**LBP-CNN**	**0.999**	**0.999**	**0.999**	**0.999**
10	HOG-SVM	0.833	0.857	0.835	0.846
11	HOG-PCA-SVM	0.868	0.969	0.785	0.867
12	HOG-RF	0.833	0.857	0.835	0.846
13	HOG-PCA-RF	0.806	0.823	0.823	0.823
14	HOG-GB	0.806	0.793	0.873	0.831
15	HOG-PCA-GB	0.819	0.835	0.835	0.835
16	HOG-MLP	0.917	0.873	0.954	0.912
17	HOG-PCA-MLP	0.743	0.818	0.684	0.745
18	FaceMesh-CNN	0.804	0.768	0.875	0.818
19	Xception	0.833	0.756	0.971	0.850
20	EffNetB3	0.931	0.941	0.914	0.928
21	EffNetV2B3	0.938	0.969	0.900	0.933
22	EffNetV2S	0.931	0.929	0.929	0.929
23	**ens-EffNet**	**0.999**	**0.999**	**0.999**	**0.999**
24	**ens-Model**	**0.978**	**0.975**	**0.983**	**0.979**

TABLE V. PERFORMANCE OF IMAGE MODIFICATION DETECTION ON B-LFW

№	Name	*ACC*	*P*	*R*	*F1*
1	LBPH-SVM	0.767	0.763	0.745	0.754
2	LBPH-PCA-SVM	0.789	0.788	0.765	0.777
3	LBPH-RF	0.782	0.778	0.761	0.769
4	LBPH-PCA-RF	0.791	0.793	0.762	0.777
5	LBPH-GB	0.779	0.776	0.757	0.766
6	LBPH-PCA-GB	0.788	0.792	0.754	0.773
7	LBPH-MLP	0.700	0.624	0.950	0.753
8	LBPH-PCA-MLP	0.864	0.855	0.858	0.859
9	LBP-CNN	0.985	0.996	0.971	0.984
10	HOG-SVM	0.968	0.967	0.966	0.967
11	HOG-PCA-SVM	0.933	0.933	0.927	0.930
12	HOG-RF	0.908	0.920	0.888	0.903
13	HOG-PCA-RF	0.850	0.869	0.812	0.839
14	HOG-GB	0.932	0.940	0.918	0.929
15	HOG-PCA-GB	0.841	0.857	0.806	0.831
16	HOG-MLP	0.966	0.964	0.962	0.963
17	HOG-PCA-MLP	0.938	0.920	0.949	0.934
18	**FaceMesh-CNN**	**0.986**	**0.971**	**0.999**	**0.985**
19	Xception	0.833	0.756	0.971	0.850
20	EffNetB3	0.852	0.853	0.828	0.840
21	EffNetV2B3	0.974	0.973	0.973	0.973
22	EffNetV2S	0.980	0.978	0.981	0.979
23	**ens-EffNet**	**0.990**	**0.989**	**0.990**	**0.990**
24	**ens-Model**	**0.989**	**0.976**	**0.982**	**0.978**

TABLE VI. FRIEDMAN TEST RESULT FOR CLASSIFICATION MODELS

Metric	χ^2	Q	p-value	α
ACC	35.172	36.560	0.036	< 0.05
P	35.172	35.238	0.049	< 0.05
R	35.172	39.844	0.016	< 0.05
F1	35.172	37.329	0.030	< 0.05

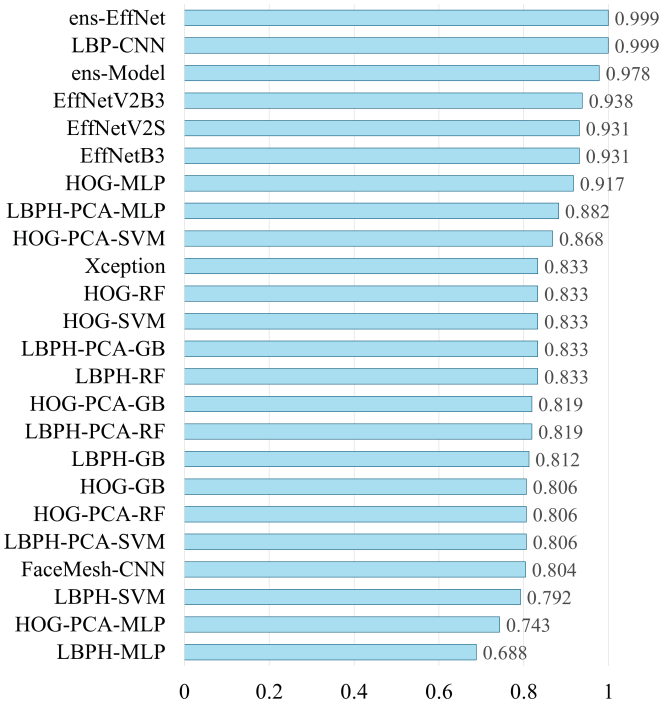


Fig. 5. Accuracy of modification detection on MakeupWild dataset

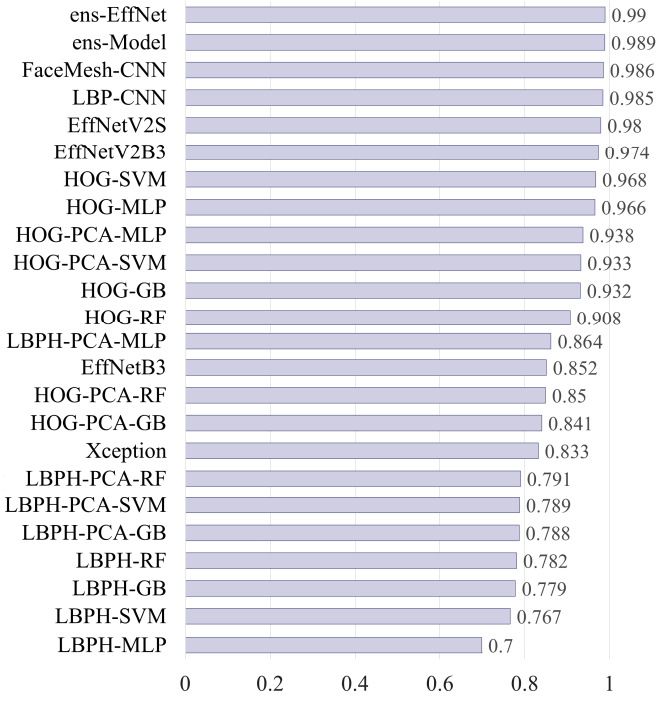


Fig. 6. Accuracy of modification detection on B-LFW dataset

V. Discussion

The obtained results demonstrate high performance of the considered models for detecting facial image modifications, including both generated makeup effects and augmented reality filters applied to images. On the MakeupWild dataset, F-measure ranges from 74.5% to 99.9%, and on the B-LFW dataset, from 75.3% to 99%. The remaining metrics, such as accuracy, precision and recall, have a value of at least 74.5%. The results of the Friedman test demonstrate that the obtained differences in the values of these metrics are statistically significant. In particular, we note the differences in the models in terms of classification recall with a smaller p-value = 1.6%.

We can note that ensembles of models show improved performance. For example, the ensemble of pre-trained models modEfficientNetB3, modEfficientNetV2B3 and modEfficientNetV2S (model 23) shows a better result than these models individually. This suggests that combining different neural network architectures enables the ensemble to leverage specific strengths and capture diverse feature representations.

Also, the ensemble of models that uses different features (model 24) has a higher efficiency than models based on individual features. This model combines the selection of geometric-based (the three-dimensional shape and structure of the face), appearance-based (skin texture and color) and deep (hidden patterns and regularities of facial image) features. This highlights the importance of multi-faceted feature representation in improving the detection of subtle facial modifications.

The extraction of deep features using pretrained models is most preferable (models 20, 21 and 22). EfficientNet models are highly effective in image classification tasks due to the optimal balance between various neural network parameters. Among the hand-crafted features tested, LBP combined with a CNN classifier (model 9) yield the best performance. HOG-based models (models 10-17) also achieve strong results and generally outperform LBP-based (models 1-8). These features better capture texture and edge information that can indicate image alterations.

The use of PCA improves the detection performance of facial modifications in a number of cases (models 1 and 2, 5 and 6, 7 and 8, 10 and 11, 14 and 15 for MakeupWild dataset; models 1 and 2, 3 and 4, 5 and 6, 7 and 8 for B-LFW dataset). PCA is particularly effective in extracting meaningful patterns from LBP histograms of facial images by reducing noise and emphasizing discriminative components, thereby enhancing the classifier's ability to detect modifications.

In summary, these results highlight several *findings*:

- deep learning-based approaches utilizing pretrained models provide superior feature representations for detecting facial image manipulations;
- combining multiple model architectures and feature types into ensembles can significantly boost performance of facial image modification detection;
- careful feature engineering techniques such as LBP and HOG combined with dimensionality reduction methods like PCA remain valuable tools in this domain.

We should also describe some *limitations* of the AI models:

- limited generalizability when trained on certain types of modifications and image forgeries;
- dependence on the quality and quantity of training data;
- vulnerability to adversarial attacks;
- uninterpretability of modification detector decisions.

Our future research areas include studying the interpretability and explainability of facial image manipulation detection, including previously unknown modifications. To explaination we can use approaches such as feature importance analysis, visualization of importance maps, local explanations, and others. We also plan to consider the possibility of generalizing the models to advanced deepfakes that go beyond beautifications.

VI. Conclusion

In this paper, we compare different AI models for facial modification detection that employ both traditional and deep feature extraction. We find that while hand-crafted features can be effective for detecting makeup, retouching, and AR filters, they often lack the flexibility and generalization capabilities provided by convolutional neural networks.

The research evaluates various classifiers, including SVM, random forests, gradient boosting, and deep neural networks. Ensembles, which combine multiple classifiers or features, are particularly effective, leveraging their mutual strengths. We also show that implementing PCA reduces computational complexity without significantly reducing performance.

We can note that an approach combining deep feature extraction, dimensionality reduction, and ensemble learning offers a promising path to developing robust facial modification detection systems. This is of high importance for combating disinformation and ensuring the integrity of digital content. In future work, we plan to also consider the possibility of detecting different types of facial modifications, including those with explainability. We also plan to explore the possibility of detecting unknown types of manipulations.

Acknowledgment

The study was funded by the budget project FFZF-2025-0016 and FASIE agreement No. 50GYCodeAIS13-D7/94529.

References

[1] I. Šiđanin, B. Milić, K. Mitrović, and J. Spajić, "Use of beauty applications and ar beauty filters among young people: Trends and challenges," in *2023 19th International Scientific Conference on Industrial Systems IoT Technologies (IS23)*, 2023, pp. 299–303.

[2] E. Hafeez and F. Zulfiqar, "How false social media beauty standards lead to body dysmorphia," *Pakistan Journal of Humanities and Social Sciences*, vol. 11, no. 3, pp. 3408–3425, 2023.

[3] S. D. McCrackin, F. Mayrand, C. Wei, and J. Ristic, "Filtered realities: navigating the social consequences of edited photographs," *Current Psychology*, vol. 44, no. 8, pp. 6989–7000, 2025.

[4] A. F. Ismail, "Excessive editing of selfies on social media: The illusion of sustainability in mental health among female adolescents," *Journal of Ecohumanism*, vol. 3, no. 7, pp. 3487–3494, 2024.

[5] R. F. Rodgers and K. Laveway, "Retouchée au féminin: The gendered nature of the french law mandating labeling of digitally modified images," *Laws*, vol. 10, no. 3, p. 62, 2021.

[6] L. Mishra and B. D. Kurmi, "Cosmetics regulations and standardization guidelines," *Pharmaspire*, vol. 15, pp. 137–150, 2023.

[7] S. Bharne and P. Bhaladhare, "Investigating online dating fraud: An extensive review and analysis," in *2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*. IEEE, 2022, pp. 141–147.

[8] N. Jagadeesha, "Facial privacy preservation using fgsm and universal perturbation attacks," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1. IEEE, 2022, pp. 46–52.

[9] T. Hassan, A. Asaad, D. Ali, and S. Jassim, "Artificial image tampering distorts spatial distribution of texture landmarks and quality characteristics," *arXiv preprint arXiv:2208.02710*, 2022.

[10] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay, "Facial biometrics in the social media era: An in-depth analysis of the challenge posed by beautification filters," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[11] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903–1913, 2016.

[12] A. Bharati, M. Vatsa, R. Singh, K. W. Bowyer, and X. Tong, "Demography-based facial retouching detection using subclass supervised sparse autoencoder," in *2017 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 474–482.

[13] A. Jain, R. Singh, and M. Vatsa, "On detecting gans and retouching based synthetic alterations," in *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2018, pp. 1–7.

[14] S. Rasti, M. Yazdi, and M. A. Masnadi-Shirazi, "Biologically inspired makeup detection system with application in face recognition," *IET Biometrics*, vol. 7, no. 6, pp. 530–535, 2018.

[15] K. Kotwal, Z. Mostaani, and S. Marcel, "Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 15–25, 2019.

[16] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting photoshopped faces by scripting photoshop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 072–10 081.

[17] C. Rathgeb, A. Botaljov, F. Stockhardt, S. Isadskiy, L. Debiasi, A. Uhl, and C. Busch, "Prnu-based detection of facial retouching," *IET Biometrics*, vol. 9, no. 4, pp. 154–164, 2020.

[18] C. Rathgeb, C.-I. Satnoianu, N. E. Haryanto, K. Bernardo, and C. Busch, "Differential detection of facial retouching: A multi-biometric approach," *IEEE Access*, vol. 8, pp. 106 373–106 385, 2020.

[19] Z. Akhtar, M. R. Mouree, and D. Dasgupta, "Utility of deep learning features for facial attributes manipulation detection," in *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)*. IEEE, 2020, pp. 55–60.

[20] T. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "Deep learning models for automatic makeup detection," *AI*, vol. 2, no. 4, pp. 497–511, 2021.

[21] P. Hedman, V. Skepetzis, K. Hernandez-Diaz, J. Bigun, and F. Alonso-Fernandez, "On the effect of selfie beautification filters on face detection and recognition," *Pattern Recognition Letters*, vol. 163, pp. 104–111, 2022.

[22] P. Majumdar, A. Agarwal, M. Vatsa, and R. Singh, "Facial retouching and alteration detection," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing Cham, 2022, pp. 367–387.

[23] K. Sharma, G. Singh, and P. Goyal, "Ipdcn2: Improvised patch-based deep cnn for facial retouching detection," *Expert Systems with Applications*, vol. 211, p. 118612, 2023.

[24] K. R. Sheth and V. Vishal S, "Transfer learning based fine-tuned novel approach for detecting facial retouching," 2024.

[25] K. R. Sheth and V. S. Vora, "Preserving authenticity: transfer learning methods for detecting and verifying facial image manipulation," 2024.

[26] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan, "Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5194–5202.

[27] P. Riccio, B. Psomas, F. Galati, F. Escolano, T. Hofmann, and N. Oliver, "Openfilter: A framework to democratize research access to social media ar filters," 2022.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[29] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[30] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay, "Impact of digital face beautification in biometrics," in *2022 10th European workshop on visual information processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[31] Y. Sun, L. Yu, H. Xie, J. Li, and Y. Zhang, "Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 584–24 594.

[32] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5549–5558.

[33] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4*. Springer, 2003, pp. 44–51.

[34] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 348–353.

[35] M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa, "Cross-spectral cross-resolution video database for face recognition," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2016, pp. 1–7.

[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[37] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[38] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[39] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 947–954.

[40] S. Milborrow, J. Morkel, and F. Nicolls, "The muct landmarked face database," *Pattern recognition association of South Africa*, vol. 201, no. 0, p. 535, 2010.

[41] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[42] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[43] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, "Feature extraction methods: a review," in *Journal of Physics: Conference Series*, vol. 1591, no. 1. IOP Publishing, 2020, p. 012028.

[44] G. Singh, S. Kaur, and S. Devi, "Facial feature extraction review for contemporary applications," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.

[45] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *arXiv preprint arXiv:1907.06724*, 2019.

[46] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[47] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[48] ——, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.