http://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Series A2-A3
Volume 64, Number 2, Pages 129-151 (2022)
DOI: 10.33769/aupse.1184112
ISSN 1303-6009 E-ISSN 2618-6462

Research Article; Received: October 04, 2022; Accepted: November 02, 2022

A METHOD FOR ANALYZING SUSPECT-FILLER SIMILARITY USING CONVOLUTIONAL NEURAL NETWORKS

Dervis Emre AYDIN[1] and Yilmaz AR[2]

[1]Lawyer, Ankara Bar, Ankara, TURKEY
[2]Department of Computer Engineering, Ankara University, Ankara, TURKEY

ABSTRACT. Eyewitness misidentifications are one of the leading factors in wrongful convictions. This study focuses on the structure of the lineups, which is one of the factors that cause misidentification, and the use of artificial intelligence (AI) technologies in the selection of fillers to be included in the lineups. In the study, AI-based face recognition systems are used to determine the level of similarity of fillers to the suspect. Using two different face recognition models with a Convolutional Neural Network (CNN) structure, similarity threshold values close to human performance were calculated (VGG Face and Cosine similarity = 0.383, FaceNet and Euclidean l2 = 1.16). In the second part of the study, the problems that are likely to be caused by facial recognition systems used in the selection of fillers are examined. The results of the study reveal that models responsible for facial recognition may not suffice alone in the selection of fillers and, an advanced structure using CNN models trained to recognize other attributes (race, gender, age, etc.) associated with similarity along with face recognition models would produce more accurate results. In the last part of the study, a Line-up application that can analyze attributes such as facial similarity, race, gender, age, and facial expression, is introduced.

1. INTRODUCTION

Extensive research on wrongful convictions reveals that misidentifications of eyewitnesses are leading factors that cause errors [1-4]. Like legal experts, scientists are aware of this problem. For decades, scientists have been working to develop more appropriate identification procedures to determine the factors that cause

Keywords. Eyewitness identification, line-up, suspect-filler similarity, stopping point, convolutional neural network.
dervis.emre.aydin@gmail.com–Corresponding author; 0000-0001-6128-6514
ar@ankara.edu.tr; 0000-0003-2370-357X.

misidentification and to minimize the risks that may arise when collecting eyewitness evidence [5].

Eyewitness experts examine factors affecting eyewitness identifications under two main groups: System variables and estimator variables [6]. Crime and eyewitness-related factors such as stress, violent content of the crime, exposure time, and perpetrator characteristics are called estimator variables [7-12]. The factors under the control of the justice system and related to the accuracy of the identification are covered under the heading of system variables [13]. Identification methods [14], police interrogation tactics [15], feedback [16], and post-event information [17] are the main areas of study on system variables.

Identification procedures are also one of the system variables affecting eyewitness identifications [18]. There are many factors related to the accuracy of the identification, such as the way the lineup is shown (live or photographed; sequential or simultaneous), the number of fillers, witness instructions, and eyewitness confidence [13, 19, 20]. One of these factors is the criteria for the selection of fillers to be included in the lineups [21].

In 1998, as a result of The Executive Committee of the American Psychology-Law Society of the APA, recommendations based on scientific findings on how to collect eyewitness evidence were published under four headings [22]. A new study has recently been published, including five new recommendations in addition to the four recommendations from the previous study [13]. This recent study explored selecting fillers for lineups under the fourth recommendation [13].

Wells et al. (2020) suggest that "there should be only one suspect per lineup and the lineup should contain at least five appropriate fillers who do not make the suspect stand out in the lineup based upon physical appearances or other contextual factors such as clothing or background". Two main approaches are discussed on the similarity of fillers in the last study [13].

1.1. **Filler Search Approaches**

1.1.1. **Match-to-description.** The first approach, called match-to-description, takes into account only the characteristics of the eyewitness in the description when selecting fillers [23]. It is assumed that this approach prevents possible misidentification due to the match of the fillers in the lineup with the description of the culprit, and facilitates the recognition of the perpetrator thanks to the variety in the facial shapes of the members of the lineup [21].

1.1.2. **Match-to-appearance.** In the other approach called match-to-appearance or resemble-suspect, the fillers are selected among those that physically resemble the suspect while creating a lineup [23]. There is a lot of research showing that innocent

suspects are more likely to be misidentified in lineups created with fillers with a low likeness to suspects [24]. At the same time, a study examining eyewitness identification procedures around the world suggests that the match-to-appearance approach is much more widely used [25]. However, there is also some criticism of this approach.

The most important criticism of the resemble-suspect approach is that it has no criterion or “stopping point” for determining how similar the fillers should be [21, 13]. When the lineup is created with fillers very similar to the suspect, it will be very difficult to identify the perpetrator from within the lineup [13]. The opinion of experts who advocate the match-to-description approach is that this approach does not include such a risk due to its natural stopping point (the witness’ description of the culprit) [21].

Although large-scale studies comparing the two approaches have highlighted findings supporting the match-to-description approach [24, 26] experts say the match-to-description approach should only be used when a complete description is obtained [13]. In cases where no exact description is taken or when the description of the eyewitness and the possible suspect are inconsistent with some physical characteristics, the selection of fillers that match the current appearance of the suspect stands out as a generally accepted hybrid approach [13].

The current recommendation from the U.S. Department of Justice on the similarity criterion includes a hybrid approach:

"3.2. Filler should generally fit the witness's description of the perpetrator, including such characteristics as gender, race, skin color, facial hair, age, and distinctive physical features. They should be sufficiently similar so that a suspect's photograph does not stand out, but not so similar that a person who knew the suspect would find it difficult to distinguish him or her. When viewed as a whole, the array should not point to or suggest the suspect to the witness [27]".

According to Wells et al., most research on suspect-filler similarity is based on subjective similarity criteria. Therefore, they noted that the appropriate level of suspect-filler similarity remains a question that science has not yet answered [13]. Based on this problem, we aimed to use AI-based facial recognition technologies to detect the appropriate level of similarity in suspect-filler similarity. First, we reviewed the computer-based systems already used when preparing the lineups.

1.2. **Computer-Based Systems Used in the Selection of Fillers**

1.2.1. **Match-to-description based systems.** Several studies show that computer systems are used by many police departments in the U.S. to access photo databases (driver's license and criminal photos) to create lineups [28-30] Officers have access to filler photos with desired characteristics by entering suspect’s physical

characteristics (race, gender, hair color, etc.) [29]. The VIPER (Video Identification Parade Electronic Recording) system used in the United Kingdom is also a filler database based on the match-to-description approach [31].

1.2.2. **Match-to-appearance / resemble-suspect based systems.**

Tredoux's Euclidean Approach. We see that Tredoux (2002) was the first researcher to use a statistically based suspect-filler similarity approach in the selection of fillers, and he used the euclidean distance between two faces as a similarity criterion based on Valentine's (1991) multidimensional 'face space' system [23, 32, 33]. In Valentine's (1991) multidimensional face space, the similarity between human faces is rated as multidimensional features such as race, age, gender, face shape, eye size, etc. This system, called multidimensional scaling, calculates the similarity between faces using the ratings of multiple features [34].

We have not been able to find any information on whether this approach used by Tredoux (2002) in the selection of fillers is utilized in real-life and field research. But there has been research in recent years showing that face recognition software has been used to select fillers from large photo databases [28].

2. Face Recognition Systems

In recent years, we have seen facial recognition systems widely used in areas such as video surveillance, identification, and autonomous vehicles. There are various techniques in face recognition that focus on the features of certain elements of the human face or use the entire face or use these approaches in a hybrid way [35]. It is possible to classify these approaches used in facial recognition as feature-based and image-based approaches [36].

Feature-based Approaches: The feature-based approach is based on data on the facial features such as eyes, eyebrows, and noses on the face and the intensity of these features [37]. It is thought that the most distinctive elements in the face area are the forehead and eye area [38].

Image-based Approaches: In image-based approaches used in facial recognition systems, artificial neural networks are widely used [36]. Artificial neural networks are one of the subfields in machine learning, a subset of artificial intelligence [39]. In the image-based approach based on artificial neural networks, techniques from statistical analysis and machine learning are utilized to find facial characteristics [36]. The face recognition models in the Deepface framework that we used in our study are also Convolutional Neural Networks-based face recognition models, a subtype of artificial neural network structure [40].

Convolutional Neural Networks: Convolutional Neural Networks (CNN) is a kind of deep learning model inspired by the working structure of the animal visual

cortex [41]. It is also preferred for tasks such as face recognition and gender classification due to its common use in computer vision tasks such as image classification and object recognition [42]. In a CNN structure, the input image is processed on layers with different tasks, and a vector output is obtained as a result [43].

The most important layer of a basic CNN structure is the Convolutional Layer, which consists of convolutional filters. The input image is filtered with convolutional filters of different sizes and feature maps are created. The pooling layer is used to create a subsampling of the feature maps. The activation function (non-linearity) used to match the input is a basic function in neural network structures and decides whether or not a neuron is fired by referring to a specific entry. The Fully-Connected Layer, which is usually found at the end of CNN structures, is responsible for the classification [39].

3. The Current Study

Although it is thought that face recognition software may be more objective than face recognition by humans due to their mathematical algorithm-based ratings [34], research by Bergold and Heaton (2018) shows that there are some risks. The results of the study using a face recognition software named Betaface (Demo) by Bergold and Heaton (2018) where they studied the effect of the facial image database size on the accuracy of the identification showed that identification decreased when the fillers were selected from the large face database [28]. Moreover, there is concern that this situation may pose a serious problem as the use of these systems becomes more common [13]. In our opinion, it seems possible to detect an objective "stopping point" by using facial recognition software in order to overcome this problem of over-similarity.

In our research, we first focused on detecting an objective similarity rate that would be used in suspect-filler similarity using an AI-based face recognition model. For this purpose, we used the VGG Face and FaceNet facial recognition models included in the Deepface framework. We determined a stopping threshold value where the performance of the models is closest to the human performance. Then we created a list of potential fillers using Betaface (demo) software. With this sample list, we've explored possible problems with facial recognition systems. Finally, we have developed an open-source AI-based application that analyzes the suspect filler similarity.

3.1. **Method and Material**

3.1.1. **Deepface framework and facial recognition models.**

***Deepface Framework*.** Deepface is an open-source facial recognition package released with an MIT license developed for the Python programming language [40]. We used the Deepface package in our study as it is able to perform the four common steps (face detection, alignment, vector representation, and verification) in modern facial recognition systems in a practical way. The framework includes Convolutional Neural Networks (CNN) based facial recognition models such as VGG-Face, Google FaceNet, OpenFace, and Facebook DeepFace [44, 40].

VGG Face Facial Recognition Model. One of the facial recognition models we used in our study was the VGG Face facial recognition model, which is the default model in the Deepface framework. Developed by the Visual Geometry Group at the University of Oxford, VGG-Face is a CNN model with 22 layers and 37 deep units. The image size of the model's input layer is 224x224x3, producing a feature vector of 2622 at the end of the process. The accuracy performance of the model in the LFW data set [45] was 98.78% [46]. Human performance in the LFW data set was found to be 97.5% [47, 43].

***FaceNet Facial Recognition Model*.** The second facial recognition model used in our study is FaceNet face recognition model, which is also included in the Deepface package. FaceNet, developed by Google, is a CNN model of 140 million parameters. FaceNet produces a 128-dimensional feature vector from the face image it receives as inputs of 160 x 160 x 3 dimensions [48]. The accuracy performance of the model in the LFW data set [45] was 99.63% [46].

3.1.2. **Metrics.** CNN-based face recognition models are responsible for producing feature vectors that best express the face images they receive as inputs. The similarity between the two faces is measured by the distance between the vectors. Different metrics such as Cosine Similarity, Euclidean Distance and Euclidean L2 are used when calculating the distance between vectors [40]. As the similarity between the two face images increases, the calculated distance value decreases, while the distance value increases as the similarity decrease. A threshold value must be determined to decide whether the images belong to the same person [40].

In this way, it can be concluded that the images belong to the same person in cases where the distance between the vectors is below the threshold value. The threshold value to be used for this purpose is not a fixed value and should be determined separately according to different metrics for each facial recognition model.

The threshold value of the VGG Face model in the cosine metric by Serengil was calculated as 0.3751 through the statistical approach [49]. It is also seen that the threshold value in the cosine metric of the VGG Face model is calculated as 0.3147 through the decision tree-based C.4.5 algorithm, and the threshold value of the

FaceNet model in the Euclidean l2 metric is calculated as 0.90 through the decision tree-based C.4.5 algorithm [40].

The equations for the evaluation metrics precision, recall, F1-score and accuracy are given as follows:

Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
Accuracy = (TN + TP) / (TN + TP + FN + FP)
(TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative)

3.1.3. **Our Datasets.** Some research on facial recognition systems reveals that the systems are more successful in males than females; they also produce more successful results in light-skinned people than in dark-skinned people [42]. Since we don't want this issue to cause a bias, which is important when determining the threshold value, we have created two separate datasets with face images. The first dataset, comprising a total of 48 images, includes four photographs of six female (2 Black, 2 White, and 2 Asian) and six male (2 Black, 2 White, and 2 Asian) celebrities. In the second dataset, there are four photographs of twelve different Black female celebrities. Since factors such as the quality of face images and the ratio of light affect the results [50], datasets were compiled from images with similar quality and light intensity.

4. Results

All the calculations in our study were performed using the Jupyter Notebook development environment through Python 3.9 programming language on a computer with Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz 2.49 GHz processor and 8.00 GB Ram hardware power. VGG Face and FaceNet facial recognition models included in the Deepface framework were used to calculate the distance values between face images.

4.1. Calculation of the Threshold Value We Use in Our Method

4.1.1. **VGG face and cosine similarity.** In order to determine the first threshold value we used in our model, we first pre-labeled the facial images in the data sets we prepared according to whether they belonged to the same person. Then, using the *verify()* function in the Deepface package, we calculated the distance between the vector values of the facial images calculated by the VGG Face model according to the cosine metric. In this way, we have determined the distance values between faces

belonging to different people (False Positive) and distance values between faces belonging to the same people (True Positive) (Table 1).

TABLE 1. Datasets outcomes (VGG Face, Cosine Similarity).

	Mixed Dataset		**Black Women Dataset**	
Decision Context	*False Positive*	*True Positive*	*False Positive*	*True Positive*
N	1056	72	1056	72
Distance Mean	0.6576	0.1863	0.4833	0.1754
Distance SD	0.1284	0.61	0.0975	0.0649

Note. The results were calculated with Deepface Framework on Jupyter Notebook. Calculations were repeated 3 times until the constant value was received. Minor changes in values (0.015) can be observed based on the system's hardware performance.

Since we will use distance values between faces belonging to the same people when calculating the threshold value, we studied whether there is a statistically significant difference between the two groups in terms of this parameter. After we found that the groups confirm the normality assumption of True Positive distance values [51], we used the Independent Sample t-Test to analyze whether there was a significant difference between the groups (Table 2).

TABLE 2. Result of analysis black women dataset with mixed dataset (VGG Face, Cosine Similarity).

	Mixed Dataset		**Black Women Dataset**		**t(142)**	***p***	**Cohen's *d***
Parameters	*M*	*SD*	*M*	*SD*			
True Positive Dist.*	0.186	0.061	0.175	0.064	1.041	0.30	0.063

Note. *Distance values between faces belonging to the same person.

As a result of the analysis, we found that there was no statistically significant difference between the datasets (M=0.186, SD=0.061), which had a balanced distribution in terms of gender and race, and the true positive distance values of the data set (M=0.175, SD=0.064) consisting solely of Black female celebrities (t(142)=1.041, p=0.30). Therefore, we continued our study with the findings of the Mixed dataset. In the next stage, we calculated the potential threshold values based on the statistical approach used by Serengil to calculate the threshold value (Threshold Value = True Positive Mean + Sigma * True Positive Standard Deviation) (Table 3). When we set the threshold value of 0.3693 (3σ), we found that

the model performed with 98.31 % accuracy. When we used 0.199 and 0.383 as the threshold value, we found that the model had the closest accuracy performance to human performance (97.5%) in the LFW data set (Table 3).

TABLE 3. Mixed dataset threshold values (VGG Face, Cosine Similarity).

Parameters	**Mixed Dataset Outcomes**				
		σ	*2σ*	*3σ*	
Distance	0.199*	0 .2473	0.3083	0.3693	0.383*
Precision	100.0%	100.0 %	100.0 %	79.12 %	72.0 %
Recall	61.11 %	83.33 %	97.22 %	100.0 %	100.0 %
F1 score	75.86 %	90.90 %	98.59 %	88.34 %	83.72 %
Accuracy	97.5% (Human)	98.93 %	99.82 %	98.31 %	97.5% (Human)

Note. The results were calculated with Deepface Framework on Jupyter Notebook. 1 sigma corresponds to 68.27% confidence, 2 sigma corresponds to 95.45% confidence, and 3 sigma corresponds to 99.73% confidence. * When we set the threshold at 0.199 and 0.383, the facial recognition system performs closest to human performance (97.5%) in the LFW dataset.

TABLE 4. Black women dataset threshold values (VGG Face, Cosine Similarity).

Parameters	**Black Women Dataset Outcomes**				
		σ	*2σ*	*3σ*	
Distance	0.199	0.2403	0.3083	0.3693	0.383
Precision	98.07 %	95.08 %	67.64 %	33.02 %	28.57 %
Recall	70.83 %	80.55 %	95.83 %	100.0 %	100.0 %
F1 score	82.25 %	87.21 %	79.31 %	49.65 %	44.44 %
Accuracy	98.04 %	98.49 %	96.80 %	87.05%	84.04 %

Note. The results were calculated with Deepface Framework on Jupyter Notebook.

4.1.2. **FaceNet and Euclidean L2.** We used the same method to determine the appropriate threshold value in the FaceNet model and the Euclidean L2 metric. Using the *verify()* function in the Deepface framework, we calculated the distance between the vector values of face images calculated by the FaceNet model according to the Euclidean metric (Table 5).

TABLE 5. Datasets outcomes (FaceNet, Euclidean L2).

	Mixed Dataset		**Black Women Dataset**	
Decision Context	*False Positive*	*True Positive*	*False Positive*	*True Positive*
N	1056	72	1056	72
Distance Mean	0.3722	0.6354	1.221	0.6087
Distance SD	0.0948	0.1015	0.1335	0.1262

Note. The results were calculated with Deepface Framework on Jupyter Notebook.

After finding that the true positive distance values of the images in mixed and Black women datasets confirm the assumption of normality according to the Kolmogorov-Smirnov Test, we used the Independent Sample t-Test to analyze whether there was a statistically significant difference between the groups (Table 6).

TABLE 6. Result of analysis black women data set with mixed data set (FaceNet, Euclidean L2).

	Mixed Dataset		**Black Women Dataset**		**t(142)**	***p***	**Cohen's *d***
Parameters	*M*	*SD*	*M*	*SD*			
True Positive Dist.*	0.635	0.101	0.609	0.127	1.398	0.164	0.114

Note. *Distance values between faces belonging to the same person (TPD). M = It is the mean of the TPD values within the group. SD= It is the standard deviation of the TPD values within the group.

As a result of the analysis, we found that there was no statistically significant difference between the datasets (M=0.635, SD=0.101) with a balanced distribution in terms of gender and race and the true positive distance values of the data set (M=0.609, SD=0.127) consisting solely of Black female celebrities (t(142)=1.398, p=0.164). Therefore, we continued our study with the findings of the Mixed dataset. When we set the threshold value of 0.9399 (3σ) in the conditions in which we used the FaceNet facial recognition model and the Euclidean l2 metric, we found that the

model performed with 99.73% accuracy. When we used 0.68 and 1.16 as the threshold value, we found that the model had the closest accuracy performance to human performance (Table 7).

TABLE 7. Mixed dataset threshold values (FaceNet, Euclidean L2).

Parameters	**Mixed Data Set Outcomes**				
		σ	*2σ*	*3σ*	
Distance	0.68*	0.7369	0.8384	0.9399	1.16*
Precision	100.0 %	100.0 %	100.0 %	96.0 %	72.0 %
Recall	63.88 %	86.11 %	100.0 %	100.0 %	100.0 %
F1 score	77.96 %	92.53 %	100.0 %	97.95 %	84.21 %
Accuracy	97.69 %	99.11 %	100.0 %	99.73 %	97.60 %

Note**.** The results were calculated with Deepface Framework on Jupyter Notebook. 1 sigma corresponds to 68.27% confidence, 2 sigma corresponds to 95.45% confidence, and 3 sigma corresponds to 99.73% confidence. *When we set the threshold at 0.68 and 1.16, the facial recognition system performs closest to human performance in the LFW dataset.

TABLE 8. Black female dataset threshold values (FaceNet, Euclidean L2).

Parameters	**Black Women Data Set**					
		σ	*2σ*		*3σ*	
Distance	0.68	0.7349	0.8611	0.955*	0.9873	1.16
Precision	100.0%	100.0%	94.52 %	72.72%	58.53 %	17.69%
Recall	72.22 %	83.33 %	95.83 %	100.0%	100.0 %	100.0 %
F1 score	83.87 %	90.90 %	95.17 %	84.21%	73.84 %	30.06 %
Accuracy	98.22%	98.93 %	99.37 %	97.60%	95.47 %	70.30%

Note. The results were calculated with Deepface Framework on Jupyter Notebook. * When we set the threshold at 0.955, the facial recognition system performs closest to human performance in the LFW dataset.

4.2. **Similarity Results of Potential Fillers Listed with Betaface (demo).** In the second part of our study, we used facial recognition software called Betaface (demo) to create a sample fillers list. Betaface (demo) is a feature-based facial recognition system [34] that calculates over 22 and 101 points on the face [52].

Bergold and Heaton (2018) selected the top 12 people who most resembled the suspect from the database they created with facial recognition software called Betaface (demo). Then, with an experienced homicide detective selecting five people from 12 photographs, lineups were created to be used in the research. We also used software called Betaface (demo) to identify the top 10 people who most resembled the corresponding author of this article from within the celebrities' database (40000+ faces of famous people). We then calculated the vector distances of these 10 people according to the Cosine metric of the VGG Face model and the Euclidean L2 metric of the FaceNet model (Tables 9 and 10).

TABLE 9. Betaface (Demo) test (1-5).

Similarity Parametres	**BetaFace (Demo) Images**				
	1st	2nd	3rd	4th	5th
Betaface Similarity	83.3%	79.9%	79.7%	79.6%	78.7%
VGG-Face (Cosine)*	0.3094	0.3310	0.2386	0.3520	0.2422
FaceNet (EuclideanL2)*	0.9104	1.0300	1.0217	1.0686	0.9236

Note. *The results were calculated with Deepface Framework on Jupyter Notebook.

TABLE 10. Betaface (Demo) test (6-10).

Similarity Parametres	**BetaFace (Demo) Images**				
	6th	7th	8th	9th	10th
Betaface Similarity	78.3%	78.3%	77.6%	77.6%	77.1%
VGG-Face (Cosine)*	0.2171	0.2644	0.3717	0.4461	0.4448
FaceNet (EuclideanL2)*	1.0231	0.9693	1.1876	1.1519	1.1658

Note. *The results were calculated with Deepface Framework on Jupyter Notebook.

5. Discussion

5.1. **Appropriate Threshold Value.** In our research, we first focused on detecting an optimal level of similarity of fillers to the suspect that could be used as a stopping point. One of the problems that we encountered at this stage was that the facial recognition models we used had a higher accuracy performance than human performance (In Mixed data set (3σ): VGG Face = 98.31%, FaceNet = 99.73% (Table 3.7), in LFW data set: VGG Face = 98.95% FaceNet = 99.63% [46]). There is a risk that a system with a higher accuracy rate than humans is listing fillers who are difficult to distinguish by humans. Therefore, we preferred a system that has a similar accuracy performance to that of humans (97.5%).

5.1.1. Appropriate Threshold Value in VGG Face and Cosine Metric We found that the VGG Face facial recognition model performed 98.31% accuracy under a threshold of 0.3693 (3σ). When we set the threshold at 0.199 and 0.383, we found that the system was approaching the human performance level of 97.5% accuracy (Table 3). We decided to use 0.383 (sigma > 3) because the value 0.199 (sigma < 1) is not statistically secure enough. In this way, we tried to minimize the risk of people being included in the lineup, who cannot be distinguished by eyewitnesses because they are so similar. As a result, we decided that the optimal level of similarity (as a stopping point) that should be used in terms of suspect-filler similarity should be 0.383 under the condition that the VGG Face facial recognition model and cosine similarity metric are used.

5.1.2. **Appropriate threshold value in faceNet and Euclidean L2 metric.** We found that the FaceNet facial recognition model performed with 99.73% accuracy under a threshold of 0.3693 (3σ). When we set the threshold at 0.68 and 1.16, we found that the system was approaching the human performance level of 97.5% accuracy (Table 7). We decided to use 1.16 (sigma > 3) because the value of 0.68 (sigma < 1) is not statistically secure enough. In this way, we tried to minimize the risk of people being included in the lineup, who cannot be distinguished by eyewitnesses because they are so similar. As a result, we decided that the optimal level of similarity (as a stopping point) that should be used in terms of suspect-filler similarity should be 0.383 under the condition that the FaceNet face recognition model and Euclidean l2 metric are used.

5.2. **Risks in Selecting Fillers with Facial Recognition Software**

5.2.1. **Too much facial similarity.** In the second part of our study, we used Betaface (demo) software to list the top 10 results that resemble the corresponding author of this article from the celebrities' database. The similarity results calculated with

Betaface (demo), VGG Face, and FaceNet can be seen in Tables 9 and 10. We used the threshold values we identified at the previous stage to assess whether the 10 listed people were eligible to be included in the lineup in terms of their similarity to the suspect. When we used the threshold value of 0.383, which we set for VGG Face and cosine metric, we concluded that 8 out of 10 people listed by Betaface (demo) software were too similar to be used in the lineup. In the case where we used the threshold value of 1.16, which we set for FaceNet and Euclidean l2 metric, we concluded that 8 out of 10 people were too similar to be used in the lineup (Table 9-10). The results at this stage also supported the concern expressed by Wells et al. (2020) that facial recognition software would choose more similar fillers to the suspect if there was no stopping point.

5.2.2. **Other similarity factors.** Apart from facial geometry-based similarity, there are other factors to consider in the selection of fillers. Factors such as race, skin color, age, gender, eye color, hair color, hairstyle, facial hair, distinctive physical features, and photo background are some of the other factors to consider when creating a lineup [27,13]. In this second part of our study, we noticed that some of the people listed by Betaface (demo) software had discrepancies in skin color, and one had a long beard. However, we found that the result of facial similarity of these problematic samples remained below the threshold values that we set as a stopping point in both face recognition models. This is not surprising, especially given the verification-oriented working structure of CNN-based face recognition models, but it is also an indication that they alone will not be sufficient to create a lineup in terms of other factors associated with similarity. However, we also know that models using CNN structure can be trained in the classification of such features [40]. Therefore, we have concluded that an advanced structure created with other feature-based trained models together with the facial recognition model will allow for the use of both match-to-description and match-to-appearance approaches based on common and objective criteria.

5.3. **The Line-up Application.** The results of our study in the second part showed that models responsible only for facial recognition may not be sufficient in the selection of fillers. Therefore, we concluded that an advanced structure using CNN models trained to recognize other characteristics (race, gender, age, etc.) associated with similarity along with facial recognition models would produce more accurate results. Based on this result, we developed open-source software that can be analyzed based on factors such as facial similarity, race, gender, age, and facial expression (Figure 1, Note. The photos were created by https://generated.photos/ and do not belong to real people).

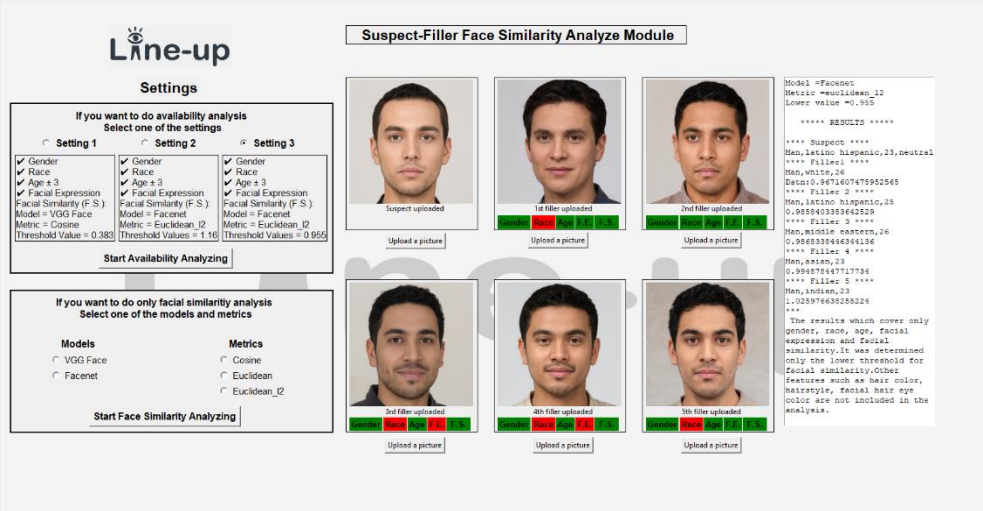

FIGURE 1. Line-up app.

We used the Python programming language and fully open-source software libraries, especially the Deepface framework when developing the Line-up application. The app allows users to perform suspect-filler similarity analysis by selecting one of the VGG Face and FaceNet facial recognition models and one of the cosines, Euclidean, and Euclidean l2 metrics according to their request. In cases where one of the default settings is selected or the appropriate threshold value range is determined, the face similarity eligibility results of the fillers uploaded to the system are listed. At the same time, conformity analysis of other similarity-related characteristics such as race, age, gender, and facial expression can be performed. With future developments, we aim to analyze other factors that will affect eyewitness identification, such as the eye color, hair color, hairstyle, facial hair, and photo background of the fillers.

6. LIMITATIONS

6.1. **Performance Differences of Face Recognition Models**

6.1.1. **Demographic Challenges.** There are many reservations about the use of artificial intelligence-based facial recognition systems in public areas. One of the most commonly cited ethical issues is the risk of increasing social bias due to the change in system performance according to gender and skin color [42]. Research by Buolamwini (2017) shows that darker-skinned females are 32 times more likely to be misclassified than men with lighter skin. It has been concluded that this resulting bias is largely related to the data sets used to train artificial intelligence systems [53].

As a result of research on biases caused by facial recognition models, systems have evolved and interracial accuracy performances have come close to each other [54]. Research shows the importance of training models with balanced data sets in these improvements [55]. Although we use trained CNN models, these standards are also important for determining appropriate threshold values. Possible performance differences resulting from the use of a fixed threshold value can also lead to bias and negative consequences for disadvantaged groups.

Since we did not want the face recognition models, we used in our study to cause such bias, when calculating the threshold values, we used two separate data sets, one with a balanced diversity of race and gender and the other consisting solely of Black female celebrities. We found that the VGG Face facial recognition model performed over 96% accuracy in both data sets (Mixed Accuracy = 99.82%, BF Accuracy = 96.80%) (Sigma = 2) (Table 3, 4). We found that the FaceNet facial recognition model performed over 99% accuracy in both data sets (Mixed Accuracy = 100.0%, BF Accuracy = 99.37%) (Sigma = 2) (Table 7, 8). In our analysis, we used distance values of the same individuals' images with each other as dependent variables and found that there was no statistically significant difference between the groups (Table 2, 6). However, when we used the appropriate threshold values that we identified (VGG Face and cosine similarity = 0.383, FaceNet and Euclidean l2 = 1.16) we found that both models performed poorly in the Black women data set (Table 4,8).
When adjusting the threshold values of models, we used human accuracy performance, measured in the LWF dataset, as a criterion. Since the LWF dataset includes gender and racial diversity, we were able to compare the accuracy performance with the mixed dataset unlike the Black women dataset.

However, at this stage, we think that the threshold value equivalent to 97.5% accuracy performance can be used in the Black women dataset as the threshold value in disadvantaged groups to avoid a bias that can cause by the performance of facial recognition models. As a result of our calculations for this purpose, we found that the VGG Face model works with 97.5% accuracy performance at a threshold of 0.278 in the Black women data set, but is statistically low in reliability (sigma < 2). We found that the FaceNet model works with 97.6% accuracy performance at a threshold of 0.955 on the Black woman data set and has high statistical reliability (sigma > 2). As a result, we have concluded that the threshold value of 0.955, which we calculated with the FaceNet model and Euclidean l2 metric, can be used in disadvantaged groups to avoid a bias caused by the performance of facial recognition models. However, we still think that setting a different threshold range for each group will minimize this potential risk to avoid a possible bias caused by the performance differences of face recognition models.

6.1.2. **Image quality-based challenges.** We know that facial recognition models need to be trained with balanced data sets in terms of gender and race so as not to cause bias. At the same time, factors such as the quality of facial images, and shadow and light ratio directly affect the results produced by face recognition models [50]. Although we use trained CNN models, these standards are also important for determining appropriate threshold values.

When preparing the dataset we used in our study, we made sure that the images had similar quality and light intensity. Nevertheless, we should say that the threshold values obtained in our study are largely valid in the conditions in the data sets prepared by us. However, when we set the threshold values at a high confidence level (sigma = 3), we found that both models produced results close to the original accuracy performance achieved in the LFW dataset, so we think the results can be generalized (For an argument that the LWF dataset does not represent a balanced sample, see [56]).

In order to achieve more reliable results, the data set should be created using photos that have standard qualities. The face recognition model and metric to be used afterward should be determined and a sample set of images should be prepared from the images in this data set and the appropriate threshold value should be determined with the method we described. Similarity calculations should be made with computers with equivalent system performance. Finally, we think that these statistically determined threshold values should be supported by experimental studies in order to be used in real life.

6.2. **Upper Threshold Value.** The fillers selected with the resemble-suspect approach must be within the acceptable similarity range [21]. Therefore, in addition to the lower threshold value that we have set as a stopping point, an upper threshold value should be determined. The upper threshold value is important to prevent the suspect from standing out in the lineup. However, since we can't determine how different the results above the threshold we have determined as a stopping point differ by statistical method, we think that the upper threshold value should be determined via experimental studies. In this respect, we think it is more accurate to select the people closest to the lower threshold (stopping point) as fillers when searching a database until a scientific threshold value is determined.

7. Conclusions

AI-based systems can be used to control the system variables under the control of the justice system and ensure equal and objective distribution of justice. At the same time, using these technologies in areas that directly affect convictions, such as the legality and reliability of evidence, seems important for the justice system to free

itself from prejudices and improve. Of course, this is only possible provided that the artificial intelligence systems to be used comply with ethical requirements such as the "European Ethical Charter on the use of artificial intelligence (AI) in judicial systems [57]" and legal regulations such as "the Artificial Intelligence Act [58]".

A great deal of research examining wrongful conviction cases reveals that the problem was largely due to misidentification [1-3]. In our study, we focused on the lineup structure, which is one of the factors that cause misidentification, and the approaches used in the selection of fillers to be included in the lineup. The results of the research conducted by Bergold and Heaton showed that fillers selected from large databases with face recognition software bore a lot of similarity to the suspect, which reduced their identification rates [28]. These results have raised concerns that using the suspect resemblance approach in this way, which has yet to have an objective stopping point, will become an increasingly common problem. [13].

In our research, we first focused on detecting an objective stopping point that could be used in the suspect-filler similarity. Using two different facial recognition models with a CNN (Convolutional Neural Network) structure, we calculated optimal threshold values close to human performance (VGG Face and cosine similarity = 0.383, FaceNet and Euclidean l2 = 1.16). To avoid a possible bias caused by the interracial performance of facial recognition models, we have identified a safe threshold value that can be used in disadvantaged groups (FaceNet and Euclidean l2 = 0.955).

However, in order to reach the most reliable threshold values, we have concluded that the face images database to be selected for fillers must be created from images with standard attributes, and that the appropriate threshold value should be determined using the method described in a sample created from face images in this database, and that these statistically determined threshold values should be supported by experimental studies in order to be used in real life.

In the second part of our research, we conducted a study on the problems that are likely to arise from the use of facial recognition software without stopping. With the Betaface (demo) software used by Bergold and Heaton (2018) in their research, we have listed the 10 celebrities who most resembled the corresponding author of this article from the celebrities' database. We found that 8 out of 10 of these people fell below the threshold values we set in both models, which means that these people are too similar to be a filler. The results of the study supported concerns raised by Wells et al. (2020) that facial recognition software would choose fillers that looked too similar to the suspect if they did not have a stopping point.

The results of our study also reveal that models responsible only for facial recognition may not be sufficient in the selection of fillers. Therefore, we concluded that an advanced structure using CNN models trained to recognize other

characteristics (race, gender, age [59]) associated with similarity along with facial recognition models would generate more accurate results.

Author Contribution Statements All authors provided critical feedback and helped shape the research, analysis and manuscript.

Declaration of Competing Interests The authors have no conflict of interest to declare.

REFERENCES

[1] Connor, E., Lundregan, T., Miller, N., McEwan, T., Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence After Trial, Office of Justice Programs, 1996.

[2] Garrett, B., Convicting the Innocent: Where Criminal Prosecutions Go Wrong, Harvard University Press, 2011.

[3] Innocence Project, Eyewitness Identification Reform, 2018. Retrieved February 17, 2022, https://innocenceproject.org/eyewitness-identification-reform/.

[4] Saks, M. J., Koehler, J. J., The coming paradigm shift in forensic identification science, *Science*, 309 (5736) (2005), 892–895, https://doi.org/10.1126/science.1111565.

[5] Berkowitz, S. R., Loftus, E., Misinformation in the courtroom, H. Otgaar, M. L. Howe (Eds.), Finding the Truth in the Courtroom: Dealing with Deception, Lies, and Memories, Oxford University Press, 2018, 11–20.

[6] Wells, G. L., Applied eyewitness-testimony research: system variables and estimator variables, *J. Pers. Soc. Psychol.*, 36 (12) (1978), 1546-1557, https://doi.org/10.1037/0022-3514.36.12.1546.

[7] Meissner, C. A., Sporer, S. L., Schooler, J. W., Person descriptions as eyewitness evidence, In book: Handbook of Eyewitness Psychology: Memory for People, 2007, 3-34.

[8] Reisberg, D., Heuer, F., Remembering emotional events, Memory and Emotion, Oxford University Press, 2004, 3-41.

[9] Semmler, C., Dunn, J., Wixted, J. T., The role of estimator variables in eyewitness identification, *J. Exp. Psychol. Appl.*, 24 (3) (2018), 400-415, https://doi.org/10.1037/xap0000157.

[10] Siegel, J. M., Loftus, E. F., Impact of anxiety and life stress upon eyewitness testimony, *Bull. Psychon. Soc.*, (12) (1978), 479-480.

[11] Yarmey, A. D., Jacob, J., Porter, A., Person recall in field settings, *J. Appl. Soc. Psychol.*, 32 (11) (2002), 2354-2367, https://doi.org/10.1111/j.1559-1816.2002.tb01866.x.

[12] Wells, G. L., Olson, E. A. Eyewitness testimony, *Annu. Rev. Psychol.*, (54) (2003), 277-

295, https://doi.org/10.1146/annurev.psych.54.101601.145028.

[13] Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., Wixted, J. T., Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence, *Law Hum. Behav.*, 44 (1) (2020), 3-36, https://doi.org/10.1037/lhb0000359.

[14] Wells, G. L., Loftus, E., Eyewitness memory for people and events, Handbook of Psychology: Forensic Psychology, Vol. 11, John Wiley and Sons Inc, 2003, 149-160.

[15] Memmon, A., Higham, P. A., A review of the cognitive interview, *Psychol. Crime Law*, 5 (1-2) (1999), 177-196, https://doi.org/10.1080/10683169908415000.

[16] Semmler, C., Brewer, N., Wells, G. L., Effects of postidentification feedback on eyewitness identification and nonidentification confidence, *J. Appl. Psychol.*, 89 (2) (2004), 334-346, https://doi.org/10.1037/0021-9010.89.2.334.

[17] Davis, D., Elizabeth, L. F., Internal and external sources of misinformation in adult witness, The Handbook of Eyewitness Psychology, Vol. 1. Memory for Events, 2007, 195-237.

[18] Wells, G. L., Eyewitness identification: A System Handbook, Carswell Legal Publications, 1988.

[19] Fitzgerald, R., Price, H. L., Valentine, T., Eyewitness identification: Live, photo, and video lineups, *Psychol. Public Policy Law*, 24 (3) (2018), 307-325, http://doi.org/10.1037/law0000164.

[20] Wells, G. L., Steblay, N. K., Dysart, J. E., A Test of the Simultaneous vs. Sequential Lineup Methods an Initial Report of the AJS National Eyewitness Identification Field Studies, 2011.

[21] Luus, C. A. E., Wells, G. L., Eyewitness identification and the selection of distracters for lineups, *Law Hum. Behav.*, 15 (1991), 43-47, https://doi.org/10.1007/BF01044829.

[22] Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., Brimacombe, C.A.E., Eyewitness identification procedures: Recommendations for lineups and photospreads, *Law Hum. Behav.*, 22 (6) (1998), 603-647, http://doi.org/10.1023/A:1025750605807.

[23] Fitzgerald, R. J., Oriet, C., Price, H. L., Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology, *Law Hum. Behav.*, 39 (2015), 62-74, http://doi.org/10.1037/lhb0000095.

[24] Fitzgerald, R. J., Price, H. L., Oriet, C., Charman, S. D., The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis, *Psychol. Public Policy Law*, 19 (2) (2013), 151–164, https://doi.org/10.1037/a0030618.

[25] Fitzgerald, R., Rubinova, E., Juncu, S., Eyewitness identification around the world, Methods, Measures, and Theories in Eyewitness Identification Tasks, 2021, 294-322, http://doi.org/10.4324/9781003138105-16.

[26] Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Carlson, M. A., Wooten, A. R., Lineup fairness: propitious heterogeneity and the diagnostic feature-

detection hypothesis, *Cogn. Res. Princ. Implic.*, 4 (2019), 20–26. http://doi.org/10.1186/s41235-019-0172-5.

[27] Yates, S. Q., Memorandum for Heads of Department Law Enforcement Components All Department Prosecutors, Department of Justice, 2017.

[28] Bergold, N. A., Heaton, P., Does filler database size influence identification accuracy?, *Law Hum. Behav.*, 42 (3) (2018), 227–243, https://doi.org/10.1037/lhb0000289.

[29] National Research Council, Identifying the Culprit: Assessing Eyewitness İdentification, The National Academies Press, 2014.

[30] Police Executive Research Forum (PERF), Library of Congress, 2013. Retrieved December 23, 2021, https://www.loc.gov/item/lcwaN0009235/.

[31] Memon, A., Havard, C., Clifford, B., Gabbert, F., Watt, M., A field evaluation of the VIPER system: A new technique for eliciting eyewitness identification evidence, *Psychol. Crime Law*, 17(8) (2011), 711-729, https://doi.org/10.1080/10683160903524333.

[32] Tredoux, C., A direct measure of facial similarity and its relation to human similarity perceptions, *J. Exp. Psychol. Appl.*, 8 (3) (2002), 180-193, https://doi.org/10.1037/1076-898X.8.3.180.

[33] Valentine, T., A unified account of the effects of distinctiveness, inversion, and race in face recognition, *Q. J. Exp. Psychol. A.*, 43 (1991), 161-204, https://doi.org/10.1080/14640749108400966.

[34] Lee, J., Mansour, J., Penrod, S., Validity of mock-witness measures for assessing lineup fairness, *Psychol. Crime Law*, 28 (3) (2021), 215-245, https://doi.org/10.1080/1068316X.2021.1905811.

[35] Yassin, K., Jridi, M., Falou, A. A., Atri, M., Face recognition systems: a survey, *Sensors*, 20 (2) (2020), 342, https://doi.org/10.3390/s20020342.

[36] Kumar, A., Kaur, A., Kumar, M., Face detection techniques: a review, *Artif. Intell. Rev.*, 52 (2019), 927-948, https://doi.org/10.1007/s10462-018-9650-2.

[37] Manjunath, B. S., Chellappa, R., von der Malsburg, C., A feature based approach to face recognition, *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (1992), 373-378, https://doi.org/10.1109/CVPR.1992.223162.

[38] Kalocsai, P., von der Malsburg, C., Horn, J., Face recognition by statistical analysis of feature detectors. *Image Vis. Comput.*, 18 (4) (2000), 273-278, https://doi.org/10.1016/S0262-8856(99)00051-7.

[39] Alzubaidi, L., Zhang, J., Humaidi, A. J., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, *J. Big Data*, 8 (53) (2021), https://doi.org/10.1186/s40537-021-00444-8.

[40] Serengil, S. İ., Özpınar, A., LightFace: A hybrid deep face recognition framework, *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, (2020), 23-27, https://doi.org/10.1109/ASYU50717.2020.9259802.

[41] Yamashita, R., Nishio, M., Do, R.K.G., Convolutional neural networks: an overview and application in radiology, *Insights. Imaging.*, 9 (2018), 611-629, https://doi.org/10.1007/s13244-018-0639-9.

[42] Buolamwini, J., Gender Shades [Master Thesis, Massachusetts Institute of Technology], MIT Media Lab, 2017, https://www.media.mit.edu/publications/full-gender-shades-thesis-17/.

[43] Taigman, Y., Yang, M., Ranzato, M., Wolf, L., DeepFace: Closing the gap to human-level performance in face verification, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2014), 1701-1708. https://doi.org/10.1109/CVPR.2014.220.

[44] Serengil, S. İ., deepface/README.md at master · serengil/deepface · GitHub, 2020. Retrieved November 13, 2021, https://github.com/serengil/deepface/blob/master/README.md.

[45] Huang, G. B., Ramesh, M., Berg, T., Learned-Miller, E., Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, 2007.

[46] Parkhi, O. M., Vedaldi, A., Zisserman, A. Deep Face Recognition, Oxford Robotics Institute, 2015. Retrieved November 23, 2021, https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf.

[47] Kumar, N., Berg, A., Belhumeur, A., Nayar, S., Attribute and simile classifiers for face verification, *Proceedings of the IEEE International Conference on Computer Vision*, (2009), 365-372, https://doi.org/10.1109/ICCV.2009.5459250.

[48] Schroff, F., Kalenichenko, D., Philbin, J., FaceNet: A unified embedding for face recognition and clustering, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 815-823, https://doi.org/10.1109/CVPR.2015.7298682.

[49] Serengil, S. İ., deepface/Fine-Tuning-Threshold.ipynb at master · serengil/deepface. GitHub, 2020. Retrieved November 13, 2021, https://github.com/serengil/deepface/blob/master/tests/Fine-Tuning-Threshold.ipynb.

[50] Krishnapriya, S., Kushal, V., Michael, K., Vítor, A., Kevin, B., Characterizing the variability in face recognition accuracy relative to rac, (2019), arXiv:1904.07325, https://arxiv.org/abs/1904.07325.

[51] Tabachnick, B. G., Fidell, L. S., Using Multivariate Statistics (6th ed.), Boston, MA: Pearson, 2013.

[52] Betaface, Betaface free online demo - Face recognition, Face search, Face analysis, Betaface API. Retrieved February 7, 2022, https://www.betafaceapi.com/demo.html.

[53] Buolamwini, J., Gebru, T., Gender shades: Intersectional accuracy disparities in commercial gender classification, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research*, 81 (2018), 77-91, https://proceedings.mlr.press/v81/buolamwini18a.html.

[54] Raji, I., Buolamwini, J., Actionable auditing: investigating the impact of publicly

naming biased performance results of commercial AI products, *Conference on Artificial Intelligence, Ethics, and Society*, (2019), https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf.

[55] Tian, J., Hailun, X. A., Hu, S., Liu, J., Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism, *Frontiers*, 2021. Retrieved December 26, 2021, https://www.frontiersin.org/articles/10.3389/fncom.2021.620281/full.

[56] Han, H., Jain, A. K., Age, gender and race estimation from unconstrained face images (Rep.), Michigan State University, 2014.

[57] CEPEJ (European Commission for the Efficiency of Justice), European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment, *Council of Europe*, 2018. Retrieved November 11, 2021, https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

[58] European Commission, EUR-Lex. Retrieved November 11, 2021, https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

[59] Serengil, S. İ., Özpınar, A., HyperExtended LightFace: A facial attribute analysis framework, *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, (2021),1-4, https://doi.org/10.1109/ICEET53442.2021.9659697.