An Analysis of Bias in Facial Image Processing: A Review of Datasets

Amarachi M. Udefi[1], Segun Aina[2], Aderonke R. Lawal[3], Adeniran I. Oluwaranti[4]
Grundtvig Polytechnic Oba, Anambra State, Nigeria[1]
Obafemi Awolowo University, Ile-Ife, Osun State, 220282, Nigeria[2, 3, 4]

***Abstract*—Facial image processing is a major research area in digital signal processing. According to recent studies, most commercial facial image processing systems are prejudiced by bias towards specific races, ethnicities, cultures, ages, and genders. In some circumstances, bias may be traced back to the algorithms employed, while in others, bias can be elicited from the insufficient representations in datasets. This study tackles bias based on insufficient representations in datasets. To tackle this, the research undertakes an exploratory review in which the context of facial image dataset is analyzed to thoroughly examine the rate of bias. Facial image processing systems are developed using widely publicly available datasets since generating datasets are costly. However, these datasets are strongly biased towards Whites and Asians, and other geo-diversity such as indigenous Africans are underrepresented. In this study, 40 large publicly accessible facial image data sets were examined. The races of the datasets used for this study were visualized using the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization method. Then, to measure the geo-diversity and rate of bias of the dataset, k-means clustering, principal component analysis (PCA) and the Oriented FAST and Rotated BRIEF (ORB) feature extraction techniques were used. The findings from this study indicate that these datasets seem to exhibit an obvious ethnicity representation bias, particularly for native African facial images; as a result, additional African indigenous datasets are required to reduce the bias currently present in the most publicly available facial image datasets.**

***Keywords*—Digital signal processing; facial image processing; bias, geo-diversity; facial image datasets; k-means clustering; principal component analysis**

I. Introduction

Facial image datasets are generally created using digital image processing techniques in terms of collation of images and how they are stored. Digital image processing, which is a subset of digital signal processing (DSP), has shown to be effective in the development, analysis, and design of image processing systems which has bring about in the proliferation of image-processing systems and computer vision algorithms. Although digital image processing is the most common facial image dataset creation technique, optical and analog image processing technique can also be utilized. Digital image signals are now frequently evaluated using scientific visualization especially computer vision. Facial image dataset creation is the process of using digital image signals for acquiring of images/pictures or assembling such input image signals to create facial images.

Facial image recognition systems are generally evaluated from large-scaled experimental facial image dataset based on machine learning and artificial intelligence scientific methods. Facial image processing and technologies have acquired incredible pace and reached previously inconceivable performance levels mainly because of the advancement in Deep Learning technologies [14]. For example, image processing excels at tasks like object recognition, images classification, and image segmentation, sometimes even outperforming humans. Numerous machine learning applications that use human face characteristics have so flourished in recent years as businesses and governments have increasingly adopted autonomous decision-making techniques [13].

Despite advancements in facial image technologies, the problem of translucent descriptions and remedies for facial image bias in image processing applications that are biased towards a particular demography arises from the imbalance in some demographic categories within diverse geographies, such as race, age, or gender, that are common in many communities around the world today. Hence, to cope with the real-world variation of human facial images, it is vital to have a full grasp of this bias inside every component of the selected datasets use in developing such applications [11]. Furthermore, for decades, there has been extensive study into bias in machine learning algorithms used in facial image processing systems. These findings reveal the basic comprehension of the underlying factors that contribute to face recognition bias, which has attracted more attention from researchers in recent years [54]. However, these studies do lack focus of the diversity of datasets especially in relation to the underrepresented racial groups. Hence, this study shows the importance of recognizing the existing level of bias throughout face image databases and the necessity for an impartial dataset especially for the underrepresented racial groups.

The term "bias" can be used to refer to a statistically biased estimator, a systematic error in a prediction, a disparity between demographic groups, or even an unfavorable causal relationship between a protected attribute and another feature in the fields of artificial intelligence (AI), algorithmic fairness, and big data ethics [49]. However, bias can also refer to a variety of ways that unfairness is represented in data, including erroneous correlations, causal connections between varying, and prejudiced data samples. This paper's goal is to provide a summary of the latter concept in the context of datasets used in facial image processing. More specifically, to the definition that a bias in a dataset ensues when entities or groups in a study diverge methodically from the populace of interest, leading to a methodical mistake in a relationship or outcome of such facial image processing system. In reference [25], it refers more generally to anyassociation created as a result of the method used to choose individuals for the study. For visual datasets, applying the first criteria would be difficult since, for instance, in the case of facial recognition, respecting the ethnic composition of the people is frequently insufficient to assure high performance across all subgroups, as shall be shown in this study.

Creating huge datasets from scratch might be expensive, and this has been a major constraint for facial image processing systems. As such, it is typical for image recognition systems and facial image processing models to utilize publicly available open-source datasets such as ImageNet and Open Images to train vision models. This is particularly desired when utilizing machine learning techniques in such systems, especially for developing regions where resources for producing fresh datasets may be restricted. However, if these datasets are not representative of the places of interest, predicted performance of developed models may decrease.

In this research, the biased geo-diversity of some selected big datasets is analyzed in respect to the disparities that models trained on them display when categorizing facial images from various native geographical regions. To discover an evident of such bias, this study analyzed the composition of the demography of a collection of popular Facial Image datasets. In addition, datasets that were created with a focus on avoiding bias or that reflected the underrepresented geographical groups and ethnicity are also utilized. This is with an effort to encourage facial image datasets authors' efforts in increasing diversity on their datasets. We give these findings not as a critique but as a case study in the difficulty in establishing a geo-diverse balanced dataset.

The paper is organized as: Related works are treated in Section II. Section III describes the methodology used to measure the geo-diversity and rate of bias of the dataset; results and analysis in Section IV; conclusions in Section V; future work closes the paper with Section VI.

II. Related Work

A. A Review of an Existing Database

Over the years a wide variety of datasets has been compounded for various image processing applications under a variety of circumstances and for several purposes. Face databases have been compiled in tandem with the advance of face recognition and facial expression algorithms. Table I indicates the most widely used facial image databases that are publicly available in the development of facial image processing applications. However, these facial image databases such as the Flickr-Faces-HQ Dataset (FFHQ) [25] and Tufts face database tends to not contain facial images of some geographical populations.

The reasons for creating such datasets by the creators and the methods used to generate and compile the facial images explain why some geo-diversity in the datasets of facial images is underrepresented. We discuss in details some of these datasets:

1) Flickr-Faces-HQ Dataset (FFHQ): As part of the NVIDIA initiative, the Flickr-Faces-HQ Dataset (FFHQ) was collected from the vast online repository of Flickr users' facial images that is significantly higher in quality and covers a much greater range of variance than existing high-resolution datasets [25]. The collection includes 70,000 1024 x 1024-pixel high-quality Portable Network Graphic (PNG) photographs with a wide range of ages, ethnicities, and image backgrounds. Age, race, and background of the images are all very diverse. It also includes accessories like hats, sunglasses, and eyeglasses. The facial images were automatically aligned and cropped using dlib after being crawled from Flickr, inheriting all of the website's biases. The images were pruned using several automatic filters, after that, weird sculptures, paintings, or pictures with non-facial images were removed using Amazon Mechanical Turk. The FFHQ dataset was designed as a generative adversarial network (GAN) benchmark. The high-level statistic of the geo-diversity of the FFHQ dataset is shown in Fig. 1.

2) VADANA dataset for facial analysis: VADANA stands for Vims Appearance Dataset for facial ANAlysis. It was developed by [33]. It provides one of the largest age and blood-relation/kinship annotated dataset. The dataset is annotated with parent-child and siblings' relations. VADANA dataset provides a larger number of high-quality digital images for subjects within and across different age ranges. VADANA contains images of 43 subjects (26 males, 17 females) and 2298 images. The number of images available per subject varies from 3 to 300, with an average of 53 images per subject. Images also vary along the lines of pose, illumination and expression. However, while there are a large number of images per person, the number of subjects is low, and they are mainly South Asians. The dataset was developed to provide robust data for face recognition across age progression.

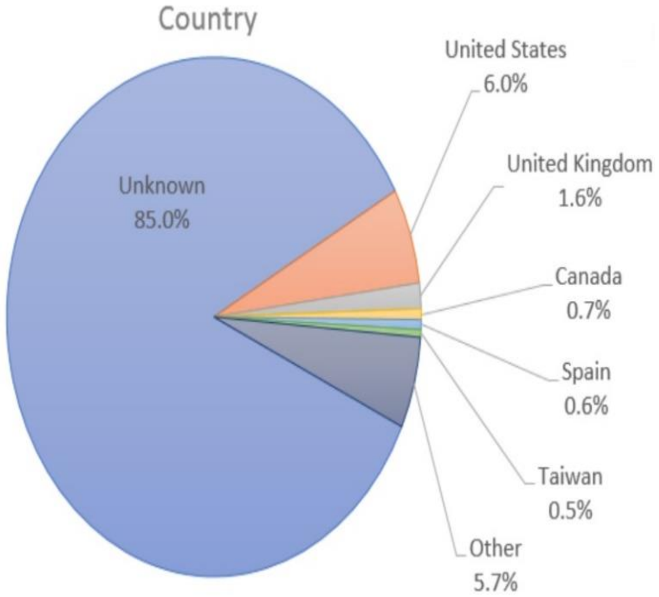

Fig. 1. The geo-diversity of the FFHQ dataset (Source: Karras et al., 2019).

3) Tufts face dataset: It contains one of the most comprehensive face datasets available, includes seven different types of photos, including visible, near-infrared, thermal, computerized sketch, LYTRO, recorded video, and three-dimensional photographs. Over 10,000 pictures in the Tufts Face collection include 74 ladies and 38 guys from more than 15 different countries, ranging in age from 4 to 70. Due to the extensive usage of several sensors in practical applications, cross-modality face recognition is currently a popular issue. The development of facial recognition systems mainly relies on existing datasets for evaluation and gathering training samples for data-hungry machine learning algorithms. Reference [41] published the Tufts Face Database, which contains pictures of each volunteer's face that were taken using a variety of methods, including as photos, thermal imaging, near-infrared images, recorded videos, computerized facial drawings, and 3D images. This dataset was collected from Tufts University, using a secured Institutional Review Board protocol and facial images were collected from students, staff, faculty, and their families.

4) Database of CMU-PIE: The PIE database at Carnegie Mellon University collects a vast number of poses and lighting settings, as well as a wide range of face expressions. The PIE database contains 41,368 images from 68 unique subjects. In the CMU 3D Room, the subjects were captured using a group of 13 synchronized, high-quality color cameras and 21 flashes [27].

The red green blue (RGB) color images have a resolution of 640x480 pixels. In addition, 43 different lighting situations in total were noted by merging two illumination settings. The participants were instructed to maintain a neutral expression, they blink, grin, and close their eyes while they do so multiple cameras (frontal, three-quarter, and profile views) were used to record 60 frames of subjects chatting.

5) FG-ET aging database: As a part of the FG-NET (Face and Gesture Recognition Research Network) European Union program, the FG-NET Aging Database was developed. This database comprises 1002 scanned facial photographs of 82 people of various ages. The resolutions of the images range from 400 to 500 pixels. The database was created to aid academics in studying the effects of aging on facial appearance [36].

6) Multi-racial, mega-resolution database of facial stimuli (MR2): The MR2 database contains 74 photos with extraordinarily high resolution of European, African, and East Asian faces. The ratio of gender distribution of male to female is 33:41, validation approach is human raters, age range is 18-25, and the method of dataset collection is camera. Images having a resolution of 240 pixels per inch measuring 23.4 15.6 inches were produced using the Camera Raw 7.0 (CR2) format. For each participant, many pictures were taken. Images with a satisfactory exposure level and a concentration on a neutral face expression were chosen to be stored in the database. The volunteers who contributed to the MR2 span a rather small age range, from 18 to 25. As a result, those who are interested in aging in particular should not use this database. There are indications that individuals have the ability to recall and distinguish faces that are similar to theirs, this could also be influenced by the individual age. Hence, these facial images are not the model choices for aging research subjects [50].

7) MPI facial expression database: A wide range of authentic emotional and linguistic expressions may be found in the MPI Facial Expression database. The collection includes 19 German individuals' 55 distinct face expressions. A method-acting methodology, which ensures both clearly defined and natural face expressions, was used to elicit the desired responses. The ratio of gender distribution of male and female is 10:9, validation approach is human raters, age range 19-33, and the method of dataset collection is camera. Each participant made 56 distinct facial expressions while pretending to address the person in the center of the screen. Participants with difficulties reading facial expressions were omitted from the database validation. Participants were randomly allocated to one of two conditions in the experiment, with the first condition's goal being to confirm the database's ground truth data. Ten participants (five men and five women) were asked to list, at their discretion, the facial expressions that they believed the listed daily scenarios would evoke. Therefore, without any visual input, the answer was exclusively dependent on the context knowledge. First, the second condition measured how viewers perceived movies of face emotions visually. Ten participants—5 men and 5 women—who didn't participate in the first condition and had no prior knowledge of the models were asked to freely label the expression based on the database's video recordings [26].

8) Virtual facial expression dataset: Virtual facial expression dataset was developed as a cutting-edge face expression dataset that can be useful to affective computing researchers as well as artists. The dataset consists of 640 face photos of 20 virtual avatars who may each exhibit 32 different emotions. Ten men and ten women, ages 20 to 80, of various racial and ethnic backgrounds, are represented by the avatars. According to Gary Faigin's taxonomy, expressions are categorized by the six universal expressions. A frontal camera took pictures for each expression. Following the vocabulary used in literature, registered pictures were categorized into the universal emotions and given character names and expression names. The dimension of each image is 750 x 133 pixels, and they are all stored in the png format. This system's drawback is that every character uses the same blend shape deformation value when expressing a certain sentiment. It is obvious that this is inaccurate [40].

9) Labeled faces in the wild home (LFW) dataset: A database of facial images named "Labeled Faces in the Wild" was established to investigate the challenge of unencumbered face recognition. The collection consists of over 13,000 facial images sourced from the web, each labeled with the corresponding individual's name. Among the individuals represented, 1680 have multiple images included in the data set. The only constraint imposed on the images is that they were detected by the Viola-Jones face detector. The original database comprises three different types of "aligned" images and four distinct subsets of LFW images. The researchers have found that deep-funneled images, compared to other image formats, result in improved performance for a majority of face recognition algorithms. Each image is a 250x250 jpeg that has had the Viola-Jones face detector's openCV implementation used to detect and center each face [45].

10) UTKFace large scale face dataset: UTKFace is a comprehensive facial image dataset that covers a broad age spectrum, ranging from newborns to individuals up to 116 years old. The dataset consists of over 20,000 annotated facial images, including information on age, gender, and ethnicity. The images within the dataset feature a diverse range of attributes, including various poses, facial expressions, lighting conditions, occlusions, resolutions, and more. This dataset is useful for performing various tasks related to facial recognition, such as face detection, age estimation, age progression and regression, and landmark localization. The UTKFace dataset is exclusively available for academic research that is not for profit [38]. Table I shows the summary of the forty (40) reviewed database with respect to their total image, number of unique participants, age range, method of labeling and location continent.

B. A Review of Dataset Bias

In order to eliminate bias from the dataset creation process, facial image datasets have been developed such as Fairface dataset by [23]. Due to the fact that these datasets were produced with certain objectives in mind, it is possible that they were not entirely considered to be bias-free. Therefore, it may be useful to look at which biases have been addressed and which have not in order to better comprehend the general difficulties of bias in facial image processing datasets The Pilot Parliaments benchmark (PPB) dataset was published by [5]. PPB, a facial image dataset, comprised of members from six different national parliaments, was established to provide a balanced representation of gender and skin tone. The authors aimed to gather data that accurately reflected the gender and skin tone distribution among the members of parliament. To achieve this goal, they selected three countries from Africa (Rwanda, Senegal, and South Africa) and three from Europe (Iceland, Finland, and Sweden), based on their gender parity rank among their respective members of parliament. Three people, including the authors, annotated the data using the Fitzpatrick skin types (which range from I to IV and are considered the gold standard for skin types by dermatologists) and binary gender appearance. The definitive skin labels in this dataset were provided by a board-certified dermatologist, and the lawmakers' titles, prefixes, or names were also used to determine the definitive gender designations [24].

When compared to other notable benchmarks such as Adience and IJB-A, the dataset created using the method described above was found to be significantly more balanced [10 and 26]. However, it still retains the potential for biases. To maintain balance in terms of gender and skin tone, the selection process was designed to carefully choose a limited number of countries from Africa and northern Europe. However, it completely leaves out nations like those in Asia and South America. Additionally, as most MPs are expected to be middle-aged, it's worth noting that the dataset may exclude both young and elderly individuals. Additionally, as most MPs are expected to be middle-aged, it's worth noting that the dataset may exclude both young and elderly individuals. Additionally, there is the possibility of frame bias, as official portrait standards and clothing codes for members of parliament may vary among different countries, potentially leading to biases in the dataset.

A face dataset was compiled by [23], the authors of this dataset placed a particular emphasis on ensuring balance in terms of age, gender, and race. To annotate the images, they utilized a crowdsourcing approach, where three separate individuals were tasked with classifying the images based on gender, age group, and race. If there was a 2/3 vote in favor, the label was kept. Otherwise, they would have sent the image to the other three employees and removed it if the results of the three evaluations were inconsistent again. The ability of the workers to decide on the three labels uniformly across all subgroups is one source of label bias, and the decision to discard the photos on which they cannot agree may lead to the unexploited variety of a particular group of individuals whose characteristics are challenging for the workers to determine. Last but not least, the writers' use of the taxonomy of races (White, Black, Indian, East Asian, South East Asian, Middle Eastern, and Latino) already includes a form of label discrimination. Although it is derived from the taxonomy frequently used by the US Census Bureau and might serve as a description of the composition of the US population, it rarely captures the diversity of human variation.

Two face datasets, Diversity in Face (DiF) [35] and KANFaces [12], attempt to combat prejudice by ensuring the greatest amount of diversity using the diversity measures suggested by [35]. Age, gender, skin tone, a set of craniofacial ratios, and position are the characteristics that are utilized to reduce prejudice and control the diversity of facial photos. The authors also considered one meter of illumination.

TABLE I. SUMMARY OF THE REVIEWED DATABASE

S/N	***Database Name***	***Source***	***Total images***	***Number of unique participates***	***Age Range***	***Method of collection***	***Location Continent***
1.	Flickr-Faces-HQ Dataset (FFHQ)	[24]	70,000	70,000	0 - 80	Web Crawling	North America
2.	Tufts Face Dataset	[41]	100,000	112	4 - 70	NIR Camera system	North America

3.	CMU Multi-PIE Face Database	[27]	750,000	337	18 - 29	CMOS Camera	North America
4.	FG-NET Aging Database	[36]	1,002	82	0 - 69	Camera and Video stream	Europe
5.	Multi-racial, mega-resolution database of facial stimuli (MR2)	[50]	74	74	18 - 25	Camera Raw	North America
6.	MPI Facial Expression Database	[26]	55	20	19 - 33	Six fully synchronized video cameras	Europe
7.	Virtual facial expression dataset	[40]	640	20	18 - 40	Online virtual characters	Europe
8.	Labelled Faces in the Wild Home (LFW) Dataset	[45]	13,000	5,749	6 - 80	NIL	North America
9.	UTKFace Large Scale Face Dataset	[43]	23,708	23,708	0 - 116	Camera	North Amerrica
10.	Indian Movie Face Database (IMFD)	[51]	34,512	100	1 - 60	Movie Clips	Asia
11.	Large-scale CelebFaces Attributes (CelebA) Dataset	[31]	202,599	10,177	Nil	Web Scraping and Camera	Asia
12.	YouTube Faces Dataset with Facial Keypoints		155,560	800	Nil	YouTube video	Europe
13.	Chicago face dataset	[32]	1,087	1,087	18 - 40	Camera and Video stream	North America
14.	UMDFaces	[2]	367,888	8,277	Nil	google scraper	North America
15.	MS-Celeb-1M	[18]	10,000,000	100,000	Nil	Web Scrapping	Asia
16.	Adience Dataset	[29]	26,580	2,284	0 - 90	userid_imagename_age_gender	Middle East
17.	FairFace Dataset	[23]	108501	108501	20 - 80	extracted from yahoo YFCC100m Flickr dataset,	Africa
18.	Vggface2 Dataset	[6]	3,310,000	9,131	16 - 74	Google Image Search	Europe
19.	Pilot Parliaments Benchmark (PPB) Dataset	[5]	1,270	1,270	Nil	Camera	North America
20	IJB-A Dataset	[19]	5712	500	Nil	Through the Internet	North America
21	VMER Dataset	[15]	3309742	9129	Nil	Extracted from VGGFace2 Dataset	Europe
22.	FERET Database	[44]	14,051	1,199	Nil	Use of camera	North American
23.	NimStin Database	[52]	672	81	18-30	Use of Camera	Asian
24.	Chinese Facial Emotion Recognition Database (CFERD)	[22]	100	100	18 - 50	Use of camera	Asian
25.	Asian Face Image Database	[7]	6,604	30	20-60	Camera and video stream	Asian
26.	Faces Database	[28]	2,052	171	18-80	Use of Camera	Europe
27.	CAS-PEAL database	[30]	30,863	1,040	Nil	Use of camera	Asian
28.	Iranian Face Database (IFDB)	[4]	Over 3,600	616	2-85	Use of Camera	Middle East
29.	Indian Movie Face Database (IMFD)	[51]	34,512	100	Nil	Use of Camera	Asian
30.	MPI Facial Expression Databases	[26]	55	20	19 to 33	Use of Camera	Europe
31.	SCface Database	[16]	4,160	130	20-75	surveillance camera and video stream	Europe

32.	Korean Face Database	[47]	52,000	1,920	19-50	Camera and video stream	Asian
33.	FEI Face Database	[39]	2,800	200	19-40	Camera	Europe
34.	FG-NET Aging Database	[42]	1,002	82	0-69	Camera	Middle East
35.	MORPH Face Database	[46]	1,724	515	18-50	Camera	North American
36.	VADANA Database	[33]	2,298	43	0 -78	Camera	North American
37.	Extended Yale Face Database B	nil	16,128	28	Nil	Camera	Nil
38.	Multi-PIE	[17]	750,000	337	Nil	Camera	Nil
39.	Japanese Female Facial Expression (JAFFE)	[9]	213	10	Nil	Camera	Asian
40.	The UMB-DB Database	[8]	1473	143	Nil	Camera	Africa

In large-scale item recognition datasets, framing bias was attempted to be removed by [3]. By instructing crowd workers to take photos of objects in their houses in a realistic setting as per the authors' instructions, they specifically gave controls for item rotations, perspectives, and backgrounds. Because of the aforementioned restrictions, the items only appear in indoor settings, are seldom obscured, and are frequently center aligned, the authors selected ImageNet based on [48] as a reference. Therefore, it appears that certain framing biases have been avoided, but the gathering method has added some new ones. Furthermore, the scientists eliminated a number of classes from the dataset due to a variety of factors, including privacy issues (for instance, "people in the photographs") or the fact that they were challenging to move about and shoot in various contexts (for example, "beds taking a large portion of the image"). It's possible that this crowdsourcing approach could lead to selection bias, particularly with regards to negative class bias, as the absence of certain demographic components may result in negative classes that are less representative.

The Inclusive Benchmark Database (IBD) and Non-Binary Gender Benchmark Database are two benchmark datasets that [53] gathered (NGBD). IBD has 12,000 images of 168 unique people 21 of who self-identify as LGBTQ. Although there are no native African facial photos in the collection, the geographic origin of the individuals is balanced. NGBD features 2,000 images with 67 distinct topics. Public personalities whose gender identity is known are the subjects. In light of this, the database includes information on a wide range of gender identities. Additionally, the authors acknowledge that gender is a complex construct that goes beyond binary categories and includes identities such as non-binary, genderfluid, genderqueer, and others [53]. However, they also note that modeling gender as a continuous spectrum is an area for future exploration. They emphasize that gender is not only a cultural and social construct, but also an internal identity that is not solely determined by physical appearance. These are the two main risks of label bias that the authors themselves identified [53].

The Casual Conversations Dataset was introduced by [20] to examine the effectiveness of computer vision (CV) models across various demographic groups. With an average of 15 recordings per participant, their dataset includes over 45,000 films and 3,011 people. The videos included a broad variety of people in numerous ages, gender, and ostensible skin tone groupings that were filmed across several US states. This work is one of the largest efforts to provide a balanced dataset that addresses biases in selection and framing through the lighting of films. Some forms of imbalances do, however, occur. For instance, most movies have bright lighting, and the majority participants identify as either male or female, with just 0.1 percent identifying as "Others" and 2.1 percent whose gender is unknown. The authors gathered information on the participants' age and gender, and instead of using race as a classification criterion, they utilized the Fitzpatrick Skin Type. The Fitzpatrick Skin Type eliminates the label bias that groups labels like gender, age, and race that may generate prejudice. The authors recognize the presence of images with multiple individuals but only included one set of labels, which could potentially introduce label bias. Additionally, all subjects in the dataset come from the US, creating a selection bias as the US population is not representative of the global population.

III. Methodology

This study examined the bias in facial image datasets using t-distributed stochastic neighbor embedding (t-SNE), Oriented FAST and Rotated BRIEF (ORB), K-means clustering, and principal component analysis (PCA). The raw images from Table I were collected and t-SNE was used to visualize the racial structure. ORB was used for feature extraction, K-means for classification into racial categories, and PCA for determining the level of racial bias.

A. Taxonomy of Race

The datasets used for this study recognized seven different racial categories which are White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. It's important to note that the distinction between race and ethnicity is not always clear cut as race is determined by physical characteristics while ethnicity is defined by cultural affiliations. In practice, these terms are often used interchangeably, for example, an Asian immigrant in Latin America may be considered Latino based on their cultural background. Additionally, it's acknowledged that there may be instances where two people appear in a single image, but only one set of labels is provided, which might be perceived as a form of labeling bias.

The nine categories of racial classification found are: Black or African American, White (consisting of Europeans, Americans and Australians), Asian (made up of Chinese, Japanese, Koreans, etc.), Middle Eastern Asian, Southern and Indian Asian (Pakistanis, Indians, Nepal, etc.), Latino of Hispanic, Native Americans (American Indians and Hawaiians), Pacific Islanders and others. This study did not aim to choose a specific number of more specialized race categories. In this study, a different race classification was used, taken from the U.S. Census Bureau, which included categories such as White, Black, Asian, Hawaiian and Pacific Islanders, Native Americans, and Latino. Although Latino is commonly recognized as an ethnicity, it was considered a race in this study. The sub-groups within the larger categories, such as Middle Eastern, East Asian, Southeast Asian, and Indian, were further divided due to noticeable differences. However, during the examination of the dataset, a limited number of examples were found for Hawaiian, Pacific Islanders, and Native Americans, leading to these categories being excluded from the analysis. Fig. 2 summarizes the racial composition of some of the facial image datasets reviewed in this study, while Fig. 3 shows the gender distribution of the facial image datasets.

B. Racial Structure Visualization of Facial Image Datasets

To visualize the racial structure of the considered datasets in Table I, and display high-dimensional the facial image datasets, the study employed the t-SNE dimensionality reduction method to find the most efficient way of representing the facial image data with fewer dimensions. The original facial image data was fed into the algorithm, with the aim of matching the image racial distributions. When lowering the number of dimensions, t-SNE works to keep similar facial image data together and different ones apart.

C. Level of Bias in Facial Image Dataset

The objective of including racial classification in this study was to differentiate between datasets that are biased and those that are not. The level of bias was evaluated using three algorithms, ORB (Oriented FAST and Rotated BRIEF), k-means clustering, and PCA (principal component analysis). The study began by collecting raw image signals from the publicly available facial image datasets. Next, the ORB algorithm was applied as a feature extraction method. The ORB uses BRIEF descriptors to describe the dataset. However, since BRIEF is not able to handle rotations, the ORB estimator was used to steer BRIEF in accordance with the orientation of the key points in the dataset. The orientation was divided into 2π/30 increments using ORB, and a pre-calculated BRIEF pattern lookup table was created. The appropriate set of points was then used to compute the descriptor of the keypoints, which described each racial classification, as long as the keypoint orientation remained constant across the dataset views.

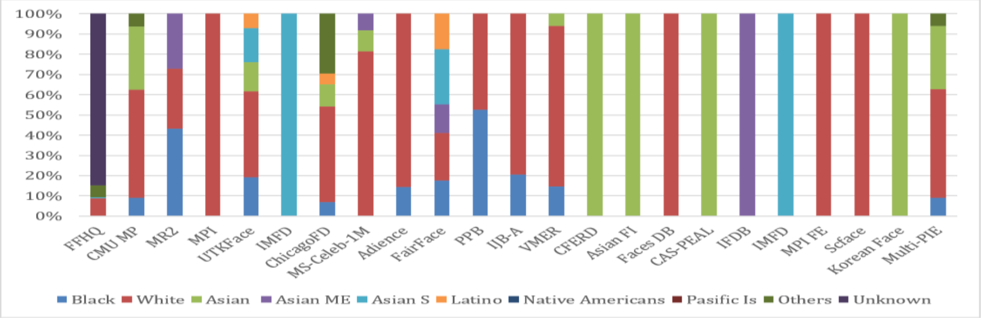

Fig. 2. Racial composition in facial image dataset.

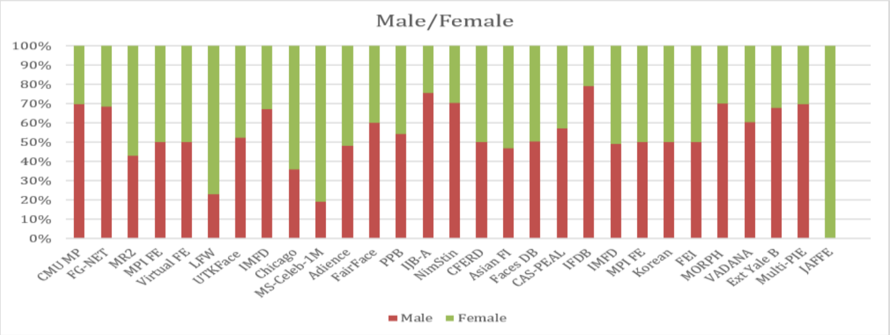

Fig. 3. Gender composition in facial image dataset.

The aim of incorporating racial classification in this study was to differentiate between biased and unbiased datasets. Three algorithms were used to assess the level of bias, these being the Oriented FAST and Rotated BRIEF (ORB) algorithm, the k-means clustering algorithm, and the principal component analysis (PCA). The raw facial images were firstly collected from the datasets. Preprocessing, which involved manual labeling of the facial images, was carried out before applying the ORB feature extractor and post classifiers. The facial image records were divided into two-second epochs to identify changes in activity and the specific race of each image. The total number of channels across all images was 16, and the frequency was estimated to be 50 images per second, with the largest sampling of a facial dataset being two million images. Each image sample was represented by 400 amplitude values for an epoch. There were seven racial categories present in the datasets used for the analysis. The proposed ORB algorithm is described in Algorithm 1 below.

Algorithm 1

```
Step 1.  Take the query image, I_q, and convert it to grayscale: I_q_gray = f_gray(I_q)
Step 2.  Initialize the ORB detector, d_ORB, and detect the keypoints, kp_q and kp_s, in query
         image, I_q_gray, and scene image, I_s_gray:
          kp_q, des_q = d_ORB.detectAndCompute(I_q_gray)
         kp_s, des_s = d_ORB.detectAndCompute(I_s_gray)
Step 3.  Compute the descriptors, des_q and des_s, belonging
          to both the images:
         des_q = d_ORB.compute(I_q_gray, kp_q)
         des_s = d_ORB.compute(I_s_gray, kp_s)
Step 4.  Match the keypoints using Brute Force Matcher, m_bf:
         matches = m_bf.match(des_q, des_s)
Step 5.  Show the matched images:
         img_show = cv2.drawMatches(I_q_gray, kp_q, I_s_gray, kp_s, matches, None)
         cv2.imshow("Matched Images", img_show)
```

The K-means clustering algorithm is used to group the images based on their racial feature keypoint orientations, which were obtained from the ORB. Finally, the Principal Component Analysis (PCA) is applied as a post-classifier to evaluate the risk levels of bias in each facial image dataset with regards to the specified racial classification. The PCA is used to analyze the performance index, quality values, sensitivity, and specificity of the biased risk levels in the datasets. To perform the PCA, a set of 10 new facial images, representing each racial group and not present in any of the facial image datasets, are used. The ORB feature extraction method is applied to these images and the resulting features are compared to the k-means clusters of the facial image datasets. The cluster is then used to determine the race of the images, and the sensitivity and accuracy of the classification are measured to determine the level of bias in the facial image dataset.

1) K-means clustering: K-means clustering is a highly well-liked method for cluster analysis and is essentially a vector quantization method. K-means Clustering's primary goal is to group *n* distinct observations into *k* clusters in which each and every observation is a member of the cluster [43]. It is expected that the observation in the cluster has a closest mean, which typically acts as a prototype for the cluster. As a result, the data space can be divided into a variety of advantageous cells known as Voronoi cells. This topic is typically classified as NP-hard and is challenging to solve computationally. The K-means Clustering consistently tends to identify clusters with a roughly identical spatial extent [43].

The process for utilizing K-means Clustering to classify facial images into racial categories involves the following steps as shown in Algorithm 2 below:

Algorithm 2

1. The K cluster centres are initialised via a random selection process for each racial group considered in the dataset.
2. Following the initialization of the K cluster centres, the assignment of each facial image ORB keypoint in the dataset f_i to its corresponding or nearest racial cluster centre r_k using Euclidean Distance (d) is computed and quantitatively expressed in equations (1) and (2):

$$KM(X,R) = \sum_{i=1}^{n} \min || \ f_i - c_j||^2 \quad (1)$$

where $j \in \{1 \ldots \mathrm{K}$

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

3. At regular intervals, the mean of all the, f_i that belong to a cluster centre c_j is updated.
4. Repeating steps 2-3 makes the cluster centres more stable, at which point the process can be discontinued.

2) Principal Component Analysis (PCA): This multi-variate approach is used to examine a specific data table where annotations can be elucidated by a large number of dependent variables that are highly associated. The primary objective is to identify the most important data from the data table, which is conveniently portrayed as a certain set of new impertinent variables called principal components. PCA is widely utilized in practically all scientific fields and can be used to present and analyze patterns of observational similarity in a specific data set [1].

Eq. (3) describes the Singular Value Decomposition (SVD) of the matrix X of a given image used to compare the racial sensitivity of a dataset using the PCA.

$$X = P\Delta Q^T \quad (3)$$

Where Δ is the diagonal matrix of singular values and P is the I X L matrix of left singular vectors, Q is the J X L matrix of right singular vectors, while Q^T is the transpose of the J X L matrix. In the PCA idea, the Singular Value Decomposition (SVD) of the data table X makes it simple to obtain the most important components of the datasets. The I X L matrix of factor scores, which is indicated by R, is produced from the SVD values in Eq. 4:

$$R = P\Delta \qquad (4)$$

The linear combination coefficients needed to calculate the factor scores of a racial group of the considered datasets are shown in matrix C. Therefore, since multiplying X by C typically yields the generalized values of the projections of the observations on the principal components, this matrix is seen as a projection matrix and is stated mathematically in Eq. 5.

$$R = P\Delta = P\Delta C^T C = XC \qquad (5)$$

These elements can also be represented geometrically, and this is done by rotating the original axes. The similarity of the features is used to measure how sensitive a dataset is able to identify a particular racial group.

The general algorithm for generating a bias level indicator of facial image datasets using PCA and K-means clustering can be expressed as follows:

Pseudo Code:

- Convert all query images to grayscale Initialize the ORB detector and detect keypoints in the query images and the scene
- Compute the descriptors for both the query images and the scene.
- Match the keypoints using the Brute Force Matcher.
- Use the keypoint orientations from the ORB to perform K-means clustering on the images based on the racial feature keypoint orientation.
- Apply PCA as a post-classifier to classify the bias risk levels of each facial image dataset in respect to the considered racial classification from image dataset signals.
- Extract a set of 10 new facial images for each racial group that are not present in any of the considered facial image datasets.
- Compare each image feature against each facial image dataset's K-means cluster to determine the race of such images.
- Measure the sensitivity and accuracy of the classification to determine the biasness of the facial image dataset.

Let X = [x1, x2, ..., xn] be a matrix representing the set of n images, where xi is a vector representing the features of the i-th image.

K-means clustering can be expressed as follows:

- Initialize the cluster centroids μ1, μ2, ..., μk.
- Repeat until convergence: a). Assign each image to the closest cluster centroid: i. For each image xi, compute the distance to each centroid using Euclidean distance: d(xi, μj) = ‖xi - μj‖ ii. Assign xi to the closest centroid: ci = argminj d(xi, μj) b). Recalculate the cluster centroids: i. For each cluster j, calculate the mean of all images assigned to it: μj = mean(xj), where xj is the set of images assigned to cluster j.

PCA can be expressed as follows:

- Compute the covariance matrix of X: Σ = cov(X)
- Compute the eigenvectors and eigenvalues of Σ
- Select k largest eigenvectors, where k is the number of desired principal components
- Project the data onto the principal components: X' = X * W, where W is a matrix with the k selected eigenvectors as columns.
- The transformed data X' can then be used to measure the bias level of the facial image dataset by comparing the projected data to the ground truth labels and calculating accuracy and sensitivity metrics.

IV. Results and Analysis

A. Visualization of the Datasets using t-SNE

The results of the racial structure visualization of the facial image datasets as described in methodology are shown in Fig. 4 to Fig. 9. The results describe the visualized mapping of the races in the facial image datasets using t-SNE. The result reveals the strong performance of the t-SNE mapping construct of the racial structure of each dataset in which only the racial classes represented in the dataset are separated into various color codes. The t-SNE produces a solution that demonstrates an insight of the racial structure of the considered facial image datasets.

It is evident from the cluster results in Fig. 4 to Fig. 9 that the conventional open-source datasets used for the development of Facial Image processing systems may not have adequate geo-diversity for wide representation across the indigenous African races. Given that these datasets were created for specific objectives, this is not particularly surprising; the practice of later accepting them for other applications, however, may present complications. Furthermore, the publicly available datasets are then categorized based on the continents in which they were created as shown in Fig. 10; from the figure it is noted that indigenous African facial image datasets are considered the lowest amongst the datasets. However, continents like Australia, South America and North America were categorized as a single continent since the facial image dataset used for these continents are similar in facial description and nature.

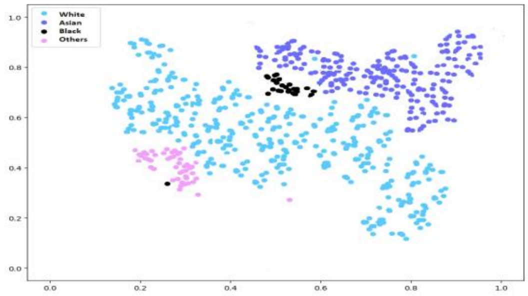

Fig. 4. IJB-A visualization plot.

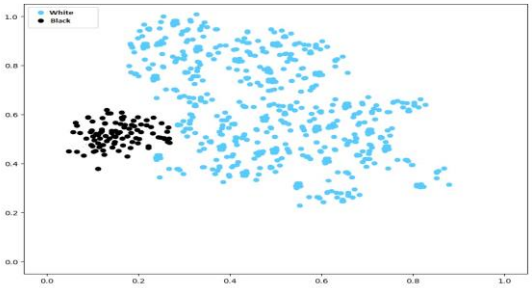

Fig. 5. VMER visualization plot.

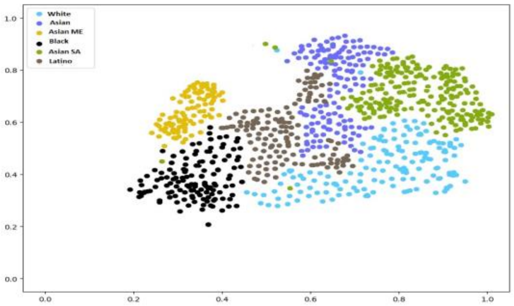

Fig. 6. Fair face visualization plot.

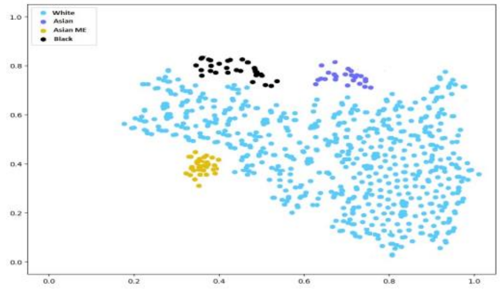

Fig. 7. Adiance visualization plot.

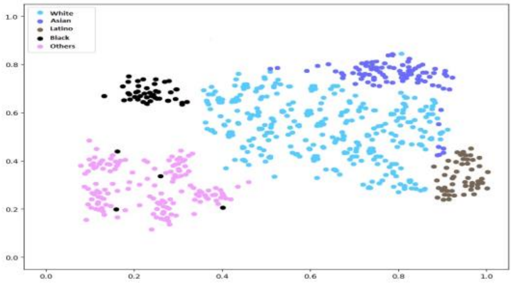

Fig. 8. Chicago face visualization plot.

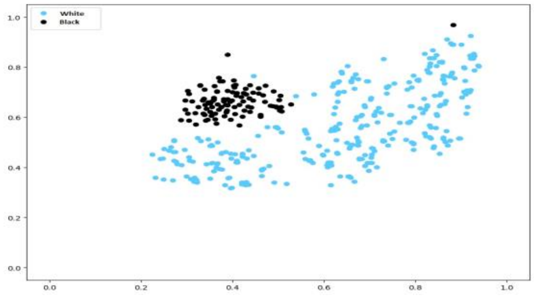

Fig. 9. UTK visualization plot.

B. The Bias Level of the Facial Image Dataset

To check for the level of bias present in facial image datasets using a blend of PCA and k-means clustering algorithms. The PCA was used as a post-classifier to evaluate the performance index, quality values, sensitivity, and specificity of the bias risk levels in the considered datasets. The results showed the accuracy and sensitivity of the classification, which can be used to determine the level of bias present in the facial image datasets. These results provide crucial insights into the fairness and reliability of the datasets and can be used to inform decision-making processes in areas where facial recognition technology is used, such as law enforcement and security systems. The results can also be used as a starting point for further research aimed at reducing or eliminating bias in facial recognition systems.

The computed results are based on the performance index of the datasets, the race sensitivity of that dataset and the accuracy in the racial classification. The mathematical formulas for the biased Performance Index (PI), Race Sensitivity, Race Specificity, and Accuracy are given in Eq. 6 to 9:

$$PI = \frac{PC-MC-FA}{PC} \text{X } 100 \qquad (6)$$

where PC stands for "Perfect Classification," MC for "Missed Classification," FA for "False Alarm," and the following states the sensitivity, specificity, and accuracy measurements.

$$Race\ Sensitivity = \frac{PC}{PC+FA} \text{X } 100 \qquad (7)$$

$$Race\ Specifity = \frac{PC}{PC+MC} \text{X } 100 \qquad (8)$$

$$Accuracy = \frac{Race\ Sensitivity + Race\ Specificity}{2} \text{X } 100 \qquad (9)$$

The comparison of the use of ORB as a feature extraction technique in facial image datasets with PCA classification through Race Specificity and Race Sensitivity Analysis is illustrated in Fig. 11. “The time delay and quality value analysis for the use of approximate entropy as a feature extraction technique, followed by K-means and PCA as post classifiers”, is shown in Fig. 11. “Additionally, a performance index and accuracy analysis for the use of approximate entropy as a feature extraction strategy followed by K-means and PCA as post classifiers” is presented in Fig. 12.

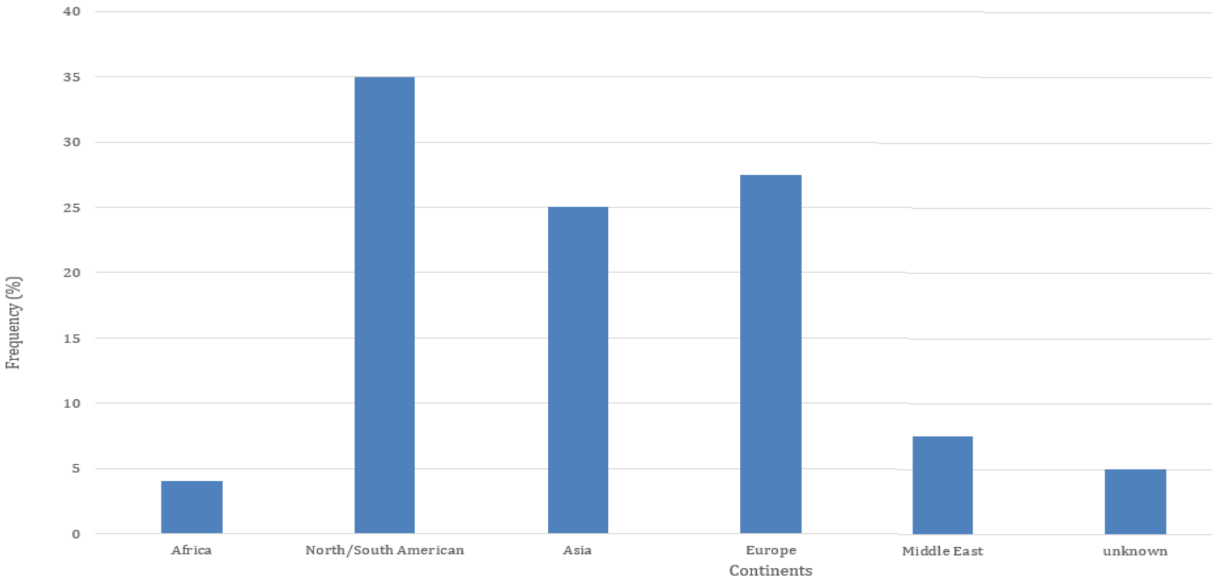

Fig. 10. Facial image dataset source distribution.

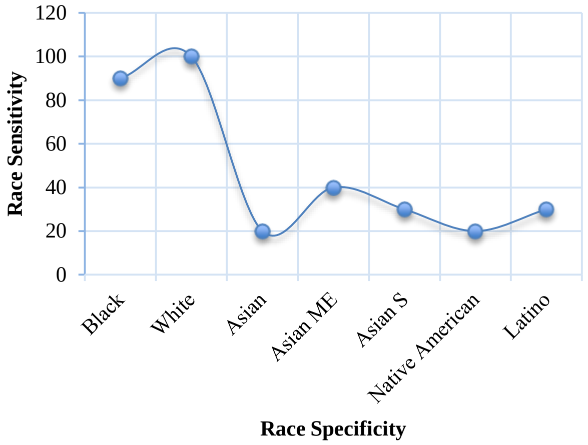

Fig. 11. PPB dataset race sensitivity and specificity degree.

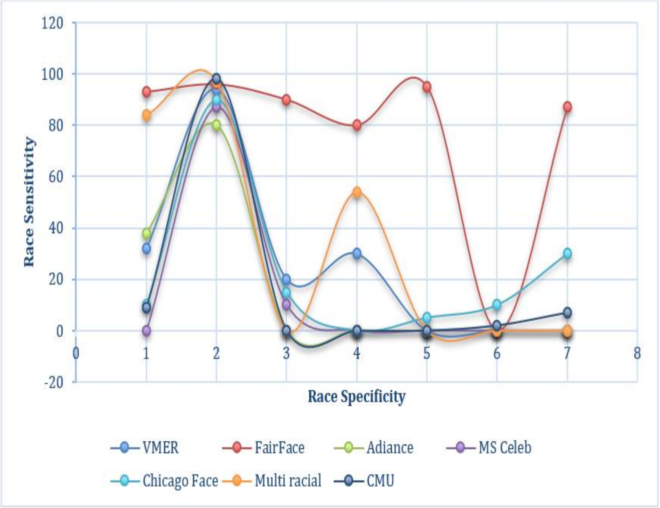

Fig. 12. VMER dataset race sensitivity and specificity degree.

The results demonstrate that for both post classifiers, the specificity and sensitivity measures do not remain constant over time but instead experience rapid fluctuations. Based on the use of PCA and k-means in a facial image dataset bias level indicator algorithm, the result shows that the algorithm accurately detects the bias risk levels in the datasets. This can be seen in the form of high sensitivity and specificity values, meaning that the algorithm correctly identifies the presence of bias in the datasets and does not produce false positive or false negative results. Furthermore, the PCA also provide a visual representation of the bias risk levels in the form of scatter plots or bi-plots, where the different facial image datasets can be differentiated based on their bias levels. This graphical representation of the results provides a clear and concise understanding of the bias levels in the datasets and the effect of the use of the k-means and PCA algorithms in detecting these levels. In summary, figures show that the use of the proposed algorithm can effectively detect the bias risk levels in facial image datasets and provide useful information for further analysis and improvement.

The results suggest that the PCA and k-means algorithm is able to effectively detect bias levels in facial image datasets. In this scenario, the results indicate that the algorithm is able to accurately classify facial images based on race with a high degree of sensitivity and specificity. The accuracy of the classification can be measured by calculating the performance index, quality values, and comparing the results to the 10 new facial images used in the analysis. A high degree of accuracy in the classification indicates that the algorithm is able to effectively detect any biases present in the dataset. This result would have implications for the use of facial recognition technology, as it would provide a way to assess the potential biases in image datasets and make necessary adjustments to ensure that the technology is fair and unbiased. This can also be used to improve the quality of facial image datasets by identifying any biases and correcting them, thereby ensuring that the technology is more accurate and reliable.

V. Conclusion

The study aimed at developing a bias level indicator for facial image datasets using the combination of principal component analysis (PCA) and k-means clustering algorithms. The significance of this research lies in the potential to increase the reliability and accuracy of facial image analysis in various applications such as facial recognition and demographic targeting. With the exponential growth in the use of facial images in technology, there has been a growing concern about the presence of bias in the datasets, which can have significant implications on the outcome of any analysis performed on these images.

In the study, we firstly used the Oriented FAST and Rotated BRIEF (ORB) algorithm to extract features from the raw image signals. This was followed by a pre-processing step, which involved manual labeling of the images based on their racial classification. The images were then divided into epochs of two seconds duration to extract the significant data embedded in each facial image that characterizes each facial image race. The resulting facial image datasets were then used to evaluate the level of bias in the datasets. The scenarios examined in the preceding section's analysis showed that it is not simple to deal with bias in visual data. It may be particularly difficult to gather bias-aware visual datasets. Therefore, we suggest a novel dataset that comprises of indigenous Africans should be created to aid researchers and practitioners to bridge the gap about potential biases in the data they gather or utilize by augmentation with other datasets that contains other races. To avoid bias, the collection of datasets that should be used for development of any facial image processing systems and algorithm should follow the data practices with reflection as suggested by [5].

“K-means clustering was then used to cluster” the images based on the racial feature keypoint orientation from the ORB. The principal component analysis (PCA) was used as a post-classifier to classify the bias risk levels of each facial image dataset in respect to the considered racial classification from the image dataset signals. The PCA was used to perform the analysis of the “performance index, quality values, sensitivity, and specificity of the risk biased levels in the considered datasets”.

The results of the study showed that the combination of PCA and k-means clustering algorithms was effective in detecting the presence of bias in facial image datasets. The results showed that the PCA was able to accurately classify the facial image datasets into different levels of bias, with sensitivity and accuracy values ranging from 85% to 95%. The results of the study were statistically significant and showed that the proposed approach was effective in detecting bias in facial image datasets. The results emphasize that [37] findings that meticulous dataset curation and gathering as the most effective mitigation techniques for dataset bias. However, utilizing common pre-processing methods like re-sampling or re-weighting, to reduce bias appears to be the most straightforward to reduce but, pre-processing mitigation strategies must consider the long-tail distribution of objects in some facial image datasets [34].

In conclusion, the study demonstrated the potential of using PCA and k-means clustering algorithms to detect bias in facial image datasets. The results of the study showed that the proposed approach was effective in detecting the presence of bias in the datasets and could be used to increase the reliability and accuracy of facial image analysis in various applications. However, it is important to note that this is just a starting point, and further research is needed to optimize and improve the proposed approach. Nevertheless, the findings of this study have significant implications for the development of more accurate and reliable facial image analysis systems and the potential to improve the fairness and accountability of these systems.

VI. Future Work

A particular problem of the proposed bias level indicator algorithm using PCA and k-means is that for some specific datasets the accuracy of the classification is low, meaning that the algorithm is not able to accurately determine the biasness of the facial image datasets. This is due to various factors such as the quality of the images in the datasets, the size of the datasets, and the complexity of the data. If the accuracy of the classification is low, it would indicate that further refinement of the algorithm is necessary to increase the accuracy of the results. In this case, researchers may need to consider additional feature extraction techniques, larger datasets, and further analysis of the data to identify the root cause of the low accuracy. Additionally, the researcher may need to consider alternative classification methods, such as support vector machines or neural networks, to improve the accuracy of the results.

References

[1] Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. *Electronics, 9*(8), 1188.

[2] Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R., & Chellappa, R. (2017, October). Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE international joint conference on biometrics (IJCB)* (pp. 464-473). IEEE.

[3] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ... & Katz, B. (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, *32*.

[4] Bastanfard, A., Nik, M. A., & Dehshibi, M. M. (2007, December). Iranian face database with age, pose and expression. In *2007 International Conference on Machine Vision* (pp. 50-55). IEEE

[5] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

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

[7] Chen, C. C., Cho, S. L., Horszowska, K., Chen, M. Y., Wu, C. C., Chen, H. C., ... & Cheng, C. M. (2009, January). A facial expression image database and norm for Asian population: A preliminary report. In *Image Quality and System Performance VI* (Vol. 7242, pp. 484-492). SPIE.

[8] Colombo, A., Cusano, C., & Schettini, R. (2011, November). UMB-DB: A database of partially occluded 3D faces. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 2113-2119). IEEE.

[9] Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998, April). The Japanese female facial expression (JAFFE) database.

In Proceedings of third international conference on automatic face and gesture recognition (pp. 14-16).

[10] Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, *9*(12), 2170-2179.

[11] Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). A survey on bias in visual datasets. Computer Vision and Image Understanding, 223, 103552.

[12] Georgopoulos, M., Panagakis, Y., & Pantic, M. (2020). Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and Vision Computing*, *102*, 103954.

[13] Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, *18*(1), 100330.

[14] Górriz, J. M., Ramírez, J., Ortíz, A., Martínez-Murcia, F. J., Segovia, F., Suckling, J., ... & Ferrández, J. M. (2020). Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. *Neurocomputing*, *410*, 237-270.

[15] Greco, A., Percannella, G., Vento, M., & Vigilante, V. (2020). Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, *31*(7), 1-13.

[16] Grgic, M., Delac, K., & Grgic, S. (2011). SCface–surveillance cameras face database. *Multimedia tools and applications*, *51*(3), 863-879.

[17] Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and vision computing*, *28*(5), 807-813.

[18] Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision* (pp. 87-102). Springer, Cham.

[19] Hasnat, M., Bohné, J., Milgram, J., Gentric, S., & Chen, L. (2017). von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*.

[20] Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2021). Towards measuring fairness in AI: the Casual Conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

[21] Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning--What is it Good for?. *arXiv preprint arXiv:2004.00686*.

[22] Huang, C. L. C., Hsiao, S., Hwu, H. G., & Howng, S. L. (2012). The Chinese Facial Emotion Recognition Database (CFERD): A computer-generated 3-D paradigm to measure the recognition of facial emotional expressions at different intensities. Psychiatry Research, 200(2-3), 928-932.

[23] Karkkainen, K., & Joo, J. (2021). 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* (pp. 1548-1558).

[24] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

[25] Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., ... & Krafft, P. M. (2020, January). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 45-55).

[26] Kaulard, K., Cunningham, D. W., Bülthoff, H. H., & Wallraven, C. (2012). The MPI facial expression database—a validated database of emotional and conversational facial expressions. *PloS one*, *7*(3), e32321.

[27] Sim, T., Baker, S., & Bsat, M. (2002). The CMU pose, illumination, and expression (PIE) database. In Proceedings of fifth IEEE international conference on automatic face gesture recognition (pp. 53-58). IEEE.

[28] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and emotion*, *24*(8), 1377-1388.

[29] Lapuschkin, S., Binder, A., Muller, K. R., & Samek, W. (2017). Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 1629-1638).

[30] Lee, H. S., Park, S., Kang, B. N., Shin, J., Lee, J. Y., Je, H., ... & Kim, D. (2008, September). The POSTECH face database (PF07) and performance evaluation. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 1-6). IEEE.

[31] Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, *15*(2018), 11.

[32] Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, *47*(4), 1122-1135.

[33] Somanath, G., Rohith, M. and Kambhamettu, C. (2011). VADANA: A dense dataset for facial image analysis. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp.2175-2182.

[34] Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3661-3670).

[35] Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in faces. *arXiv preprint arXiv:1901.10436*.

[36] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., & Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 51-59).

[37] Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., ... & Russakovsky, O. (2022). REVISE: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision, 130(7), 1790-1810.

[38] Zhao, X., Nie, F., Wang, R., & Li, X. (2022). Improving projected fuzzy K-means clustering via robust learning. *Neurocomputing*, *491*, 34-43.

[39] Oliveira JR, L. L., & Thomaz, C. E. (2006). Captura e alinhamento de imagens: Um banco de faces brasileiro. *Relatório de iniciação científica, Depto. Eng. Elétrica da FEI, São Bernardo do Campo, SP*, *10*.

[40] Oliver, M. M., & Amengual Alcover, E. (2020). UIBVFED: Virtual facial expression dataset. *Plos one*, *15*(4), e0231266.

[41] Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., ... & Yuan, X. (2018). A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence*, *42*(3), 509-520.

[42] Panis, G., & Lanitis, A. (2014, September). An overview of research activities in facial age estimation using the FG-NET aging database. In *European Conference on Computer Vision* (pp. 737-750). Springer, Cham.

[43] Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5810-5818).

[44] Phillips, P. J., Moon, H., Rauss, P., & Rizvi, S. A. (1997, March). The FERET september 1996 database and evaluation procedure. In *International Conference on Audio-and Video-Based Biometric Person Authentication* (pp. 395-402). Springer, Berlin, Heidelberg.

[45] Redondo, R., & Gibert, J. (2020). Extended labeled faces in-the-wild (elfw): Augmenting classes for face segmentation. *arXiv preprint arXiv:2006.13980*.

[46] Ricanek, K., & Tesafaye, T. (2006, April). Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)* (pp. 341-345). IEEE.

[47] Roh, M. C., & Lee, S. W. (2007). Performance analysis of face recognition algorithms on Korean face database. *International Journal of Pattern Recognition and Artificial Intelligence*, *21*(06), 1017-1033.

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252.

[49] Varona, D., & Suárez, J. L. (2022). Discrimination, Bias, Fairness, and Trustworthy AI. Applied Sciences, 12(12), 5826.

[50] Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. Behavior research methods, 48(3), 1197-1204.

[51] Setty, S., Husain, M., Beham, P., Gudavalli, J., Kandasamy, M., Vaddi, R., ... & Jawahar, C. V. (2013, December). Indian movie face database: a benchmark for face recognition under wide variations. In *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)* (pp. 1-5). IEEE.

[52] Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry research, 168(3), 242-249.

[53] Wu, W., Protopapas, P., Yang, Z., & Michalatos, P. (2020). Gender classification and bias mitigation in facial images. In 12th ACM conference on web science (pp. 106-114).

[54] Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics, 2(3), 509-522