IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2025, accepted 7 April 2025, date of publication 9 April 2025, date of current version 18 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3559310

RESEARCH ARTICLE

Multi-Task Faces (MTF) Data Set: A Legally and Ethically Compliant Collection of Face Images for Various Classification Tasks

RAMI HAFFAR, DAVID SÁNCHEZ, (Senior Member, IEEE), AND JOSEP DOMINGO-FERRER, (Fellow, IEEE)

Department of Computer Engineering and Mathematics, Center for Cybersecurity Research of Catalonia (CYBERCAT), Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain

Corresponding author: Rami Haffar (rami.haffar@urv.cat)

This work was supported in part by European Commission under Project H2020-871042 "SoBigData++," in part by MCIN/AEI/10.13039/501100011033 and "ERDF A Way of Making Europe" under Grant PID2021-123637NB-I00 "CURLING" and Grant PRE2019-089210, and in part by INCIBE and European Union NextGenerationEU/PRTR (Project "HERMES" and INCIBE-URV Cybersecurity Chair). The work of David Sánchez and Josep Domingo-Ferrer was supported by the Government of Catalonia (ICREA).

ABSTRACT Human facial data offers valuable potential for tackling classification problems, including face recognition, age estimation, gender identification, emotion analysis, and race classification. However, recent privacy regulations, particularly the EU General Data Protection Regulation, have restricted the collection and usage of human images in research. As a result, several previously published face data sets have been removed from the internet due to inadequate data collection methods and privacy concerns. While synthetic data sets have been suggested as an alternative, they fall short of accurately representing the real data distribution. Additionally, most existing data sets are labeled for just a single task, which limits their versatility. To address these limitations, we introduce the Multi-Task Face (MTF) data set, designed for various tasks including face recognition and classification by race, gender, and age, as well as for aiding in training generative networks. The MTF data set comes in two versions: a non-curated set containing 132,816 images of 640 individuals, and a manually curated set with 5,246 images of 240 individuals, meticulously selected to maximize their classification quality. Both data sets were ethically sourced, using publicly available celebrity images in full compliance with copyright regulations. Along with providing detailed descriptions of data collection and processing, we evaluated the effectiveness of the MTF data set in training five deep learning models across the aforementioned classification tasks, achieving up to 98.88% accuracy for gender classification, 95.77% for race classification, 97.60% for age classification, and 79.87% for face recognition with the ConvNeXT model. Both MTF data sets can be accessed through the following link. https://github.com/RamiHaf/MTF_data_set

INDEX TERMS Face images, image data set, image classification, deep learning.

I. INTRODUCTION

Artificial intelligence (AI) is highly dependent on the availability of data, which is the cornerstone for training and evaluating AI models and unlocking their full potential [1]. Since the accuracy and the effectiveness of AI models are directly influenced by the quality of the data on which those models are trained, meticulous collection and precise labeling of data are paramount in machine learning.

Data collection procedures vary significantly depending on the type of data gathered. In the case of tabular data, collecting information can be done explicitly through surveys and polls that require people to answer. An alternative procedure is to obtain this type of data from public (e.g., government) administrative records. When it comes to textual data, there are significant limitations due to copyright restrictions. The

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

right to copy, digitize, and collect text is heavily curtailed by copyright holders [2]. Nevertheless, there are valuable sources of text data that are not copyrighted or available as public domain material. For example, Wikipedia provides enormous and curated amounts of text data that can be used for training purposes. The situation is similar for image data, where there are restrictions related to the rights to copy and modify images. Yet, one difference between text and image data is that public-domain image sources are much more limited.

Images are especially complex data. On the one hand, their redistribution may or may not be authorized according to the applicable copyright licenses [3]. On the other hand, they may be subject to special regulations according to their contents. For instance, medical images are essential to train computer-aided diagnosis systems, but their use requires either explicit consent by the patients or the removal of image traits or metadata that might lead to privacy disclosure [4]. Also, images containing human faces are employed in a wide range of applications —including automated identification of people— that raise profound legal and ethical issues. In particular, their use is restricted by regulations concerning the protection of personally identifiable information, such as the General Data Protection Regulation (GDPR) [5] or the Ethics Guidelines for Trustworthy AI [6] of the European Union.

More specifically, the GDPR defines personal data in its Article 4 as ''any information relating to an identified or identifiable natural person ('data subject').'' Based on this definition, we can confidently state that facial images are legally considered personal data because they have the capability to easily identify the individuals they depict. The GDPR also limits the processing of personal data to ensure privacy. In Article 9, it states that ''processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.'' Facial images, due to their re-identifying nature and the information they contain about an individual's ethnic origin, race, and biometric features fall into this category of non-processable data. However, the GDPR indicates exceptions to this prohibition in the same article, specifically in its second paragraph. One such exception is ''(e) processing relates to personal data which are manifestly made public by the data subject.'' Hence, processing of personal data on individuals who have chosen to make them public without any restrictions is allowed. This applies to images that have been published under public-domain licenses or with Creative Commons licenses that grant the right to share and modify the content.

Face images are crucial in many areas of AI research, such as facial recognition [7], emotion detection [8], age estimation [9], gender classification [10], facial biometric analysis [11], and other related tasks, such as face anonymization [12]. However, the availability of such data has significantly decreased in the last years.

In particular, several data sets of face images have recently become access-restricted or have been completely removed from the internet due to privacy concerns or to comply with data protection regulations. Furthermore, many image data sets lack clarity regarding the methods employed for image collection, which makes them susceptible to potential removal in the near future. The withdrawal of previously public data sets has several shortcomings and, in particular, hampers the reproducibility of research results obtained from those data. As an alternative, some data sets are composed only of synthetic images, whose applicability is limited because they tend to deviate from real-world data distributions.

A. CONTRIBUTIONS AND PLAN

To overcome the limited availability of human face data sets, in this paper, we present the Multi-Tasks Faces (MTF) data set, a collection of real-face images aimed at training and evaluating AI models for various tasks, including face recognition (FR), gender classification (GC), age classification (AC), and race classification (RC). The data set is presented in two flavors: a non-curated version that includes 132,816 images of 640 individuals, and a manually curated version with 5,246 images of 240 individuals meticulously selected to maximize their classification quality. The compilation of the images has been done in compliance with current legal regulations. In particular, all original images are publicly available and correspond to well-known celebrities, which avoids identity disclosure issues while keeping data identified. In addition, the compiled images have a copyright license that is either public domain or Creative Commons with permission to modify, share, and commercial use. Both data sets have been compiled under the umbrella of the SoBigData++ project,[1] and they have been approved by the ethical and legal board of the project. SoBigData++ is a project funded by the European Union that aims to design and deploy an integrated infrastructure for social mining and big data analytics, and it adheres to the world's strictest data protection regulations.

In addition to describing the data collection and processing procedures that were carried out to build the data sets, we also report the results of a comprehensive evaluation of the curated and non-curated MTF data sets on several well-known deep learning (DL) models. These results constitute baselines for future research endeavors that employ images for the intended tasks. Furthermore, we highlight the significance of manual data processing by comparing the performance of each DL model trained on the curated MTF data set with the same model trained on the non-curated version. This justifies the necessity and importance of the different steps followed to create the curated MTF data set.

[1] https://plusplus.sobigdata.eu/

The remainder of this paper is organized as follows. Section II discusses previously published data sets of face images and highlights their limitations. Section III details how our data sets have been collected, processed, and labeled. Section IV describes the structure of the data sets in detail. Section V reports the baseline results obtained when using the curated MTF data set with five different DL models, and compares its results with the non-curated version. The final section summarizes the key contributions of this work and outlines several lines for further research.

II. RELATED WORK

CMU Face Images [13] is a small data set consisting of 640 black and white face images corresponding to 20 individuals (32 images per individual). It is labeled according to pose (straight, left, right, up), expression (neutral, happy, sad, angry), eyes (open, sunglasses), and scale (full-resolution, half-resolution, quarter-resolution). Information is not available on the type of consent the 20 individuals gave in regard to releasing their images. On the other hand, this data set allows tasks related to pose/emotion classification, but its small size, its lack of race diversity, the low resolution of some images, and the fact that they are black and white limit its applicability.

The Large-scale CelebFaces Attributes (CelebA) data set [14] is a widely employed data set consisting of 202,599 face images featuring 10,177 celebrities. Each image is labeled according to the corresponding identity, in addition to 40 binary attributes depicting the appearance of the individual (e.g., eyeglasses, bangs, wearing hats, mustaches), among others. However, the CelebA data set was compiled by crawling images from the Internet, irrespective of the copyright ownership held by the owners of the images. Consequently, the availability of the CelebA data set is potentially at risk due to privacy concerns and copyright laws, particularly within the European Union [15].

The recently published DigiFace-1M data set [16] comprises 1.22 million synthetic face images. It is divided into two parts: i) 720,000 images featuring 10,000 virtual identities (72 images per identity), and ii) 500,000 images encompassing 100,000 virtual identities (5 images per identity). Although this data set offers a substantial number of images, it lacks information regarding the distribution ratio between males and females or between different age groups. It is important to note that the sole purpose of this data set is to train FR models. However, it is worth considering that all the images in this data set are synthetically generated. This fact limits the data set's representativeness, as it does not accurately reflect real-world data.

The Labeled Faces in the Wild (LFW) data set [17] is another noteworthy data set of face images. It contains 13,233 images, corresponding to 5,749 unique identities. The publishers of the data set emphasize that all the data was legally and ethically collected from original images under the Creative Commons copyright licenses. However, LFW presents two significant drawbacks. Firstly, the images are not 'cleaned', meaning that the backgrounds of the images contain a substantial amount of information, sometimes including other individuals' faces. Secondly, each identity in the data set has a very limited number of images, with only 1,680 identities having more than one image and none of the identities having more than three images. These drawbacks limit the applicability of this data set to certain unsupervised biometric identification applications and render it unsuitable for training supervised models.

Another accessible data set is Megaface [18], which consists of one million face images that have been ethically collected from images that fall under Creative Commons copyright licenses. However, Megaface is specifically designed to facilitate the training of *face detection* models. The data set exclusively offers bounding boxes that specify the facial regions within the images, while it does not include identity labels for face recognition purposes. This makes Megaface a valuable resource for advancing face detection algorithms, but it may not be suitable for tasks that require identity recognition or other demographic classification tasks.

Similarly to Megaface, other data sets have been released with the specific purpose of facilitating the training of face detection models. Two such data sets are the WIDER FACE data set [19], which comprises 32,203 images, and the Multi-Attribute Labeled Faces (MALF) data set [20], which comprises 5,250 images. Both of these data sets have been carefully collected, ensuring ethical collection practices, and each image has been labeled to indicate the precise location of faces within the images. However, both data sets are exclusively suitable for face detection tasks and cannot be directly employed on any other AI training tasks without additional processing and relabeling of the images.

IMDB-WIKI [21] consists of 523,051 face images, which were obtained from two well-known websites (IMDB and Wikipedia) providing images from the public domain. Nevertheless, this data set is only labeled for age estimation.

The Flickr-Faces-HQ (FFHQ) data set [22] is an extensive collection of 70,000 high-quality images portraying human faces with remarkable diversity in terms of age, ethnicity, and image backgrounds. These images were directly downloaded from Flickr, and they are subject to Creative Common copyright licenses. Additionally, FFHQ lacks any processing or labeling, as its primary purpose is to aid in the development of generative adversarial networks.

On the other hand, several other popular face image data sets have been confronted with accessibility challenges due to the way their images were compiled. VGGFace2 [23] featured 3.3 million face images captured in real world conditions, representing more than 9,000 distinct identities. This data set was specifically labeled for FR, with an average of 362 images per identity. However, the images within the data set were obtained by downloading them from Google image search without taking into account the copyright ownership of the images. As a result, VGGFace2 is no longer publicly available. Although the specific reasons for

its withdrawal from public access have not been explicitly disclosed, they are likely to be related to concerns about privacy and legal considerations.

UMDFaces [24] is another significant data set that contains 367,888 facial images of 8,277 individuals, which is labeled for face recognition, but also provides information on 21 key points that capture essential biometric details. However, the authors have not specified the methodology employed to collect and process the data, thereby raising concerns about the ethical aspects of the creation of this data set. Currently, UMDFaces is not available for download. The data set's website indicates that the authors are working towards making it accessible again, which implies that they are making efforts to address compliance with relevant regulations.

To our knowledge, the aforementioned data sets represent the only existing collections of human faces specifically intended for training AI models. Unfortunately, all of them suffer from a variety of drawbacks that diminish their usefulness as benchmarks for FR models: some of them are no longer available due to legal and privacy concerns, and others risk removal for the same reasons. On the other hand, other data sets are entirely synthetic, lack a sufficient number of images to train robust FR models, or are labeled for tasks other than FR. Table 1 summarizes their characteristics and compares them with the MTF data sets we present in this paper (non-curated and curated).

III. DATA COLLECTION, PROCESSING, AND LABELING

The MTF data set has been carefully collected to take advantage of the above-mentioned exception in Article 9 of the GDPR, which allows for the collection and processing of personal data that have been voluntarily made public by the data subject/owner. Along this line, our data collection process prioritized privacy by exclusively focusing on publicly known individuals. This ensures the legal basis needed for publicly releasing the data set and hence its availability and longevity. Furthermore, we sought and obtained approval from the Board of Operational Ethical and Legality Evaluation (BOEL) of the SoBigData++ H2020 project.

This section provides an overview of the methodologies employed for selecting, downloading, processing, and labeling the images in the non-curated and curated MTF data sets.

A. DATA COLLECTION

The initial phase of our data collection process involved selecting the celebrities to be included in the data set. We utilized the IMDB website to search for celebrity names across different regions. To enhance the diversity and inclusivity of the data sets, we intentionally included four distinct ethnicities, which are consistent with those employed by the United States Census Bureau: Asian (Chinese/Korean), Asian (Indian), Black, and White.

In order to maintain a balanced and realistic distribution, we included an equal number of male and female individuals. This approach ensures fairness and avoids any gender bias within the collected data. Recognizing age as a crucial factor, we also sought to include an equal number of young and old celebrities in the data sets. This consideration adds another dimension to the data and enables research related to age-based analysis.

To distinguish between young and old identities within both MTF data sets, we labeled as young people between ages 18 and 49, and as old people aged 50 or older.

The search we conducted on the IMDB website resulted in the selection of an equal number of celebrities from each of the four ethnicities considered. Our selection process relied on the published lists available on the IMDB website. We conducted searches based on specific criteria and identified celebrities who appeared on these lists. The selection was made according to their presence and ranking on these lists.

The distribution of the selected celebrities from each ethnic group is as follows: 40 old male celebrities, 40 old female celebrities, 40 young male celebrities, and 40 young female celebrities. This totals 160 celebrities selected from each of the four ethnicities, resulting in a grand total of 640 distinct identities. This systematic approach ensures an initial balanced representation within each ethnicity and across different age and gender groups.

To download the images of the chosen celebrities, we used the icrawler[2] library. For this purpose, we chose the Bing search engine as a source of the images due to its advanced filtering capabilities by image copyright. Bing offers six different options: ''Creative Commons'', ''public domain'', ''free to share and use non-commercially'', ''free to share and use commercially'', ''free to modify, share, and use non-commercially'', and ''free to modify, share, and use commercially''. In contrast, other search engines such as Google only provide two options: ''Creative Commons'' without any additional specifications, and ''commercial and others''.

The icrawler library proved to be an excellent resource as it provided us with the ability to select a copyright filter for the downloaded images employing the filter options offered by the search engine. By leveraging this functionality, we were able to narrow down our image selection to those aligned with our desired copyright requirements. Specifically, we focus on images that fall under the public domain or Creative Commons licenses, granting us the rights to modify, share, and use the images commercially.

To compile a comprehensive data set, we chose not to impose a limit on the number of images downloaded per identity due to the limited availability of images with the desired copyright licenses. The image crawling process stopped once no further images were obtainable.

At the end of this phase, the data set consisted of a total of 204, 712 images. However, the distribution of these images among the various identities was not uniform. This

[2]https://pypi.org/project/icrawler/

TABLE 1. Features of the surveyed data sets on human faces. The last two rows show the features of our MTF data sets (non-curated and curated).

Data set	*Images*	*Annotations*	*Accessibility*	*Legal compliance*	*Observations*
CMU Face Images	640	Pose, expression, eyes, size	Available	Contains images of the research team with their consent	Cannot be used to train FR models
CelebA data	202,599	5 landmark locations, 40 binary attributes	Available	Published under the Chinese copyright law	Images collected regardless of their original copyright. Cannot be used to train FR models
DigiFace-1M	1,220,000	Identities	Available	Synthetic images, no copyright laws apply	Can be used only to train FR models. The synthetic nature limits the data's representativeness
LFW	13,233	5,749 identities	Available	Created from images with Creative Commons copyright licenses	The data cannot be used before being processed and cleaned
Megaface	1,000,000	Face location in the image	Available	Created from images with Creative Commons copyright licenses	Cannot be used to train FR models
WIDER FACE	32,203	Face location in the image	Available	Ethical collection practices	Cannot be used to train FR models
MALF	5,250	Face location in the image	Available	Ethical collection practices	Cannot be used to train FR models
IMDB-WIKI	523,051	Age estimation	Available	Collected from IMDB and Wikipedia	Cannot be used to train FR models
FFHQ	70,000	N/A	Available	Downloaded from Flickr, they are subject to the Creative Common copyright licenses	Cannot be used to train FR models. Can only aid in the development of generative adversarial networks
VGGFace2	3,300,00	9,000 identities	Not available	Downloaded from Google image search without taking into account the copyright ownership of the images	Not accessible due to legal and ethical implications
UMDFaces	367,888	8,277 identities, 21 biometric keypoints	Not available	No information is given	Not accessible due to legal and ethical implications
Non-curated MTF	132,816	640 identities, age, gender, race	Available	Ensuring ethical collection practices, created from images with Creative Commons or public-domain copyright licenses	Automatically labeled and processed
Curated MTF	5,246	240 identities, age, gender, race	Available	Ensuring ethical collection practices, created from images with Creative Commons or public-domain copyright licenses	Manually labeled and processed

uneven distribution was primarily influenced by two factors: the availability of images on the internet for each celebrity and the individual choices made by celebrities regarding copyright permissions. As a result, the data set exhibited non-iidness (nonidentical and independent distribution among identities). Nevertheless, this non-iidness accurately reflected the true distribution of data available online for those specific celebrities under the required licenses.

B. AUTOMATIC PROCESSING

To ensure accurate training of AI models for facial image classification, identity recognition, and human face generative models, it is crucial to have facial images that focus on the faces themselves, while minimizing background details. Therefore, a necessary pre-processing step involves cropping the faces from the crawled images.

To this end, each downloaded image was processed through the Haar cascades method [25], a machine learning-based approach that involves training a cascade function with input data to automatically detect facial regions within the images. For this purpose, we leveraged the pre-trained Haar cascades method available in the OpenCV [26] library. This automated process proved efficient in providing the bounding boxes of the facial areas in the images. These boxes varied in size depending on the appearance of each face in the original images. We then utilized these bounding boxes to crop and isolate the facial regions from the acquired images. As a result, we obtained a collection of 132, 816 cropped images, representing 640 distinct identities.

Note that the total number of images was lower than that of the initially crawled images mainly due to the presence of mismatches caused by the search engine. Some of the crawled images were not relevant to the desired facial recognition task and did not even contain any recognizable faces. It is also worth mentioning that the cropped images had varying resolutions according to the size of their respective original images. To make the images more amenable to AI model training, they underwent a final automated process to standardize their sizes with techniques similar to those used in previous data sets like LFW, WIDER FACE, and MALF.

As a result, all images were resized to a uniform resolution of 1024 × 1024 pixels. *The resulting images constitute the non-curated MTF data set*.

C. MANUAL PROCESSING

At this stage, the images were manually processed in order to build a curated version of the MTF data set focusing on maximizing classification quality. Three experienced human evaluators were involved to visually analyze each cropped image. Their primary objective was to confirm the presence of the desired celebrity's face in each image. This thorough examination process encompassed the following steps:

- First, the human experts filtered the data set by eliminating each face image that did not belong to the correct identity. For instance, many of the crawled images featured groups of celebrities or celebrities attending festivals with a large number of fans in the background, leading to multiple faces being cropped from the same image. However, only one of those faces belonged to the desired celebrity, which required the removal of the rest. Additionally, search engine results occasionally included images of celebrity relatives, such as spouses, children, or co-workers, which were also excluded by the experts during this stage.
- Second, to ensure the suitability of the data set for training robust AI models, the experts implemented several further filtering criteria. Images that contained hidden or too dark parts of the celebrity's face, such as sunglasses or hands covering the mouth or eyes, were removed. In addition, images that were hand drawn, artificially altered, or generated by AI algorithms were excluded. The data set was further refined by eliminating images where the celebrity's face appeared unnatural due to artistic effects or intrusive movie-related makeup, typically seen in images taken during movie shoots. Furthermore, images with visual disturbances such as heavy pixelation or blurring, which could potentially mislead the AI model, were also excluded from the data set. These filtering criteria ensured that the data set consists only of high-quality, natural, and unaltered facial images suitable for training AI models.
- Third, to mitigate the risk of data leakage and avoid any potential additional unnecessary computational cost during AI model training, the experts implemented measures to identify and remove duplicate or highly similar images (e.g., burst shoots) from the data set. They employed three techniques for this purpose. The first technique involved an automated code written using the OpenCV library. This code calculated the similarity between images and alerted the experts about potential duplicates. The second technique utilized the Duplicate Photos Fixer[3] application, which systematically scanned the data set and compiled a list of potentially duplicated or very similar images. Lastly, the experts visually inspected the data set to ensure that no very similar or duplicate images were overlooked. This manual examination served as a final verification step to confirm the absence of any remaining duplicates.
- Fourth, to ensure the usefulness of the data set in different tasks, experts eliminated images that did not meet the criteria for all desired tasks. For instance, images depicting outdated representations of currently old individuals that could impact the age classification task, or images of the celebrities during their childhood that could affect both face recognition and age classification, were removed.

Finally, the experts conducted a thorough examination of the remaining images to assess their usefulness. During this stage, the experts counted the number of images for each identity that remained in the data set after undergoing all the aforementioned processes. Individuals with less than five remaining images were excluded. This step was essential because a limited number of images would not provide sufficient data for effective training, validation, and testing on the face recognition task. After this meticulous manual curation, the size of the data set was reduced to 5, 246 images and 240 identities, but each had a sufficient amount of data available. *These images constitute the curated MTF data set*.

Although this manual curation resulted in a significant reduction in the number of images, it is important to note that this filtering process is a clear advantage over other data sets such as Megaface, VGGFace2, and UMDFaces, which include a large number of systematically crawled internet images that can introduce significant noise during training.

D. DATA LABELING

After the data processing phase, the images were labeled based on the predetermined criteria established during the collection phase. These labels were specifically tailored for the four tasks outlined below:

1) *Face recognition*: For the non-curated MTF data set, each image was assigned one of 640 identities they were crawled from. In the case of the curated MTF data set, each image was assigned one of the 240 remaining identities.
2) *Race classification*: Each image was assigned one of the following four labels according to the race of the corresponding individual:
   - Asian (Chinese/Korean)
   - Asian (Indian)
   - Black
   - White
3) *Gender classification*: Each image was assigned one of the two following labels:
   - Male
   - Female
4) *Age classification*: Due to images corresponding to different ages of the corresponding individual, we treated this task as a binary classification. Each of the images was assigned one of the following two labels:

[3]https://www.duplicatephotosfixer.com

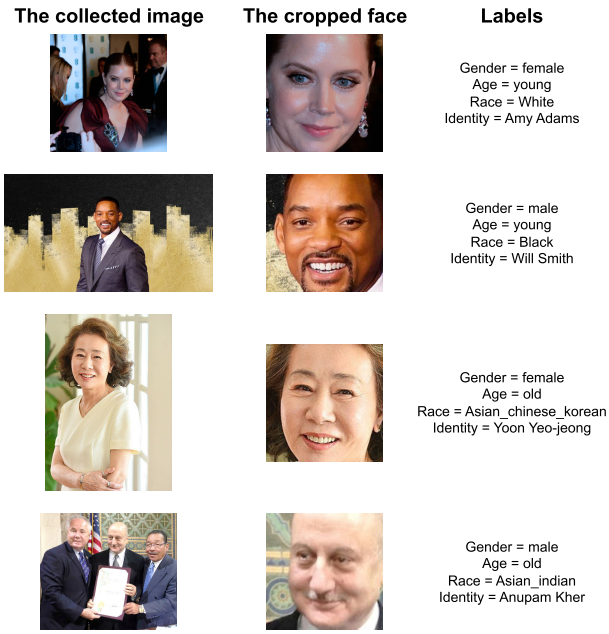


FIGURE 1. Examples of collected, processed, and labeled images from the MTF data sets.

- Young
- Old

Fig. 1 illustrates the procedures explained in this section, by showing the transition from the original collected images to their corresponding cropped versions and the assigned labels.

IV. THE MULTI-TASK FACES (MTF) DATA SETS

After completion of the processing and labeling of all images, the distribution of identities assigned to each label in the four different tasks was shown in Table 2. While our initial efforts aimed to crawl a balanced number of images across all tasks and labels, the actual distribution of the data available online resulted in an imbalance within the curated data set. This imbalance can be attributed to several factors: i) celebrities from different regions of the world publish their images at different rates and under different copyright licenses; ii) young celebrities tend to publish their images more frequently than old celebrities; and iii) old celebrities often have more images from their younger days than from their current age.

The four classification tasks are remarkably diverse.

- The FR task involves a classification problem with many categories, as it requires labeling identities based on their names.
- The second is the RC task, which presents a multi-label classification problem with four labels. The distribution of identities in this task is imbalanced for the curated data set, as it features two majority groups and two minority groups. The majority groups consist of Asians (Chinese/Korean) and Whites, as celebrities from these regions tend to share their images with our target copyright licenses more frequently than celebrities from the other two regions, namely Asian (Indian) and Black celebrities.
- The GC task represents the first binary classification problem. The data are relatively well balanced between the two labels in the curated data set, with slightly more male identities present. However, both male and female identities share their images at a similar rate, although females tend to share a higher number of images overall.
- Finally, the AC task also represents a binary classification problem. In contrast to the GC task, the AC task exhibits an extremely unbalanced data distribution in the curated data set. This disparity is due to the reasons listed above, resulting in the majority of identities and images falling under the "young" category, while the "old" category comprises only 514 images that belong to a mere 50 identities.

TABLE 2. Distribution of the identities and images among the different tasks of both data sets.

Task	Label	Curated MTF		Non-curated MTF	
		Identities	Images	Identities	Images
Face recognition	Identities	240	5,246	640	132,816
Race classification	Asian (Chinese /Korean)	80	1,715	160	32,257
	Asian (Indian)	49	820	160	31,503
	Black	35	478	160	31,612
	White	76	2,133	160	37,444
Gender classification	Males	130	2,490	320	68,702
	Females	110	2,756	320	64,114
Age classification	Young	190	4,682	320	67,462
	Old	50	514	320	65,354

These demographic imbalances, particularly those that involve underrepresented age and racial groups, have important implications for fairness. Models trained on skewed datasets may exhibit biased performance, favoring majority groups while underperforming on minority ones. This can lead to unfair or unreliable outcomes when such models are deployed in real-world applications. Consequently, researchers using the MTF data sets are encouraged to carefully consider these imbalances, especially when evaluating or reporting model performance, and to apply bias mitigation techniques or rebalancing the data where necessary. Note that *if users of the curated MTF data set wish to rebalance the distribution within each task, they can do so by selecting balanced subsets of the data set.*

Moreover, the MTF data sets offer potential for a wide range of tasks that extend beyond the four previously mentioned. Their flexibility allows defining multi-criteria classification tasks involving two or even three attributes, which provides diverse options for AI model development. In Table 3, we present the count of identities and the respective count of images for the cross-tabulation of race and gender, as well as the cross-tabulation of race and age for the curated MTF data set.

TABLE 3. Counts of identities and images when cross-tabulating by race and gender, and by race and age in the curated MTF data set. The first figure in each cell refers to the count of identities, and the second figure to the count of images.

Labels		Gender labels		Age labels	
		Males	Females	Young	Old
Race labels	Asian (Chinese /Korean)	36/1,020	44/695	69/1,612	11/103
	Asian (Indian)	22/300	27/520	43/749	6/71
	Black	18/194	17/384	30/533	5/45
	White	54/976	22/1,157	48/1,788	28/345

Similarly, Table 4 reports the number of identities and their corresponding images for the cross-tabulation of age and gender in the curated data set.

TABLE 4. Counts of identities and images when cross-tabulating by age and gender in the curated MTF data set. The first figure in each cell refers to the count of identities, and the second figure to the count of images.

Labels		Gender label	
		Males	Females
Age labels	Young	92/2,049	98/2,633
	Old	38/441	12/123

A. CURATED DATA SPLITTING FOR TRAINING AND TESTING

As the main purpose of the curated MTF data set is to train and test classification models, we performed two rounds of data splitting. In the first split, we allocated 70% of the curated data (3, 662 images) for training purposes, while the remaining 30% was set aside for validation and testing. Then we took a second split on the 30% portion, which we divided into two subsets: approximately 20% of this portion (332 images) was designated as validation data, to be employed for fine-tuning and optimization of the model. The remaining 80% (1, 252 images) were allocated as test data, to be used as an independent evaluation set to assess the generalization and accuracy of the trained model. To facilitate a consistent evaluation in future experiments conducted on the data set, we maintained the same data splitting across all tasks. We did so following a methodology that ensured that each subset of the data contained at least one image that represents each label within each task. Specifically, we made sure to include at least one image per label in each of the validation and test subsets, as well as a minimum of three images per label in the training subset. This approach was particularly important for the FR task, where the data is more spread across the 240 identity labels, which requires a more evenly distributed data set.

B. DATA RELEASE

Both MTF data sets have been made available under the Creative Commons copyright license, granting permissions for sharing, modifying, and commercial usage. This respects the copyright of the original images. Users are obligated to provide proper attribution to the authors of the data sets and the owners of the original images when utilizing any of the MTF data sets. This includes citing the present paper and specifying the source of the data.

The curated MTF data set has been made accessible through the European SoBigData++ catalogue. It can also be accessed alongside the non-curated data set using the following GitHub webpage that points to the catalogue: https://github.com/RamiHaf/MTF_data_set

The release includes the full labeling of images by organizing them in an appropriate folder structure depicted in Fig. 2. The main folder of the curated data set comprises three subsets,'Train', 'Val', and 'Test' to be used, respectively, for AI model training, AI model validation and hyperparameter tuning, and AI model testing (evaluation of the trained model). Inside each subset, which is also the root folder of the non-curated data set, there are four folders corresponding to race classification denoted as 'Asian_chinese_korean', 'Asian_indian', 'Black', and 'White'. Within each race folder, there are two additional folders named 'Males' and 'Females', representing the two labels used for gender classification. Inside each of the gender folders, there are two more folders labeled 'Young' and 'Old', which indicate the two age categories used for age classification. Within each 'Age' folder, there are folders corresponding to the identities.

With this structure, researchers have the flexibility to rearrange the data within the data sets according to their specific preferences and requirements, e.g., by defining multi-criteria classification tasks involving several attributes as mentioned in Tables 3 and 4.

Additional items made available through the above-mentioned GitHub webpage are as follows:

- To facilitate and support future research on the MTF data, we have released a second curated version that is pre-split and organized specifically for users interested in only one of the four tasks. This version includes four separate folders: 'race_classification', 'gender_classification', 'face_recognition', and 'age_classification'. Each folder contains all the images that are labeled according to the corresponding task.
- We release also the trained models we evaluate in the next section. Since these models provide baseline results for the different tasks supported by the curated data set, by releasing them we aim at facilitating future investigations conducted on this data.
- Finally, we release the Python code we used for evaluating both data sets (non-curated and curated) with the various DL models mentioned in the next section. This code can serve as a baseline for other researchers to compare their work with the results reported here.

V. EXPERIMENTAL RESULTS

To assess the efficacy of the curated MTF data set for training DL models and establish a performance baseline across its supported tasks, we conducted evaluations using five well-known DL classification models. In the following we describe

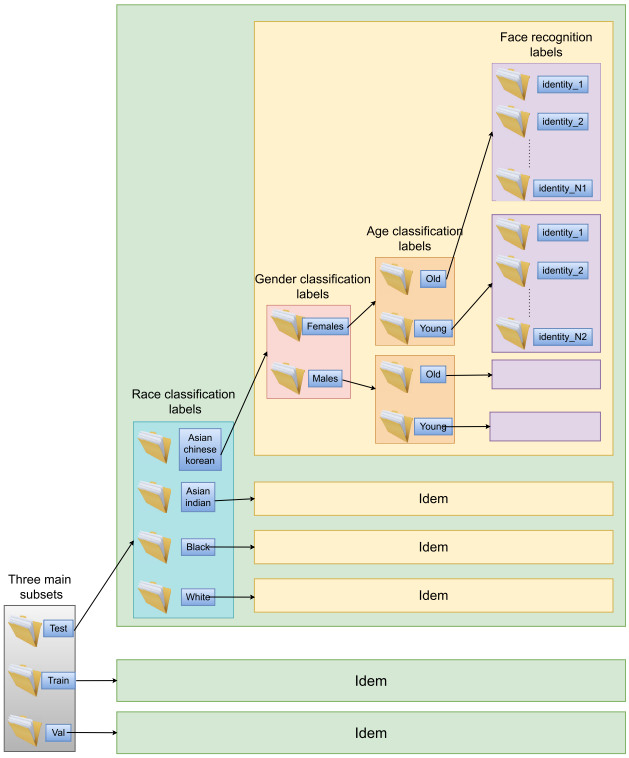


FIGURE 2. Organization of the folders in the released data set.

the evaluation metrics employed to assess the performance achieved, the specific models used in the experiments, and the obtained results.

A. EVALUATION METRICS

We employed the following standard evaluation metrics to measure the performance of DL models trained on the MTF data set:

- *Accuracy*: Ratio of correctly predicted instances to the total number of predicted instances in the data set. It provides an overall measure of the classification correctness:

$$Accuracy = \frac{Number\_of\_correct\_predictions}{Total\_number\_of\_predictions}.$$

- *Precision*: Proportion of true positive predictions to the sum of true positive and false positive predictions. It indicates the accuracy of positive predictions:

$$Precision = \frac{True\_positives}{True\_positive + false\_positive}.$$

- *Recall*: Proportion of true positive predictions to the sum of true positive and false negative predictions. It captures the model's ability to identify positive instances:

$$Recall = \frac{True\_positives}{True\_positive + false\_negative}.$$

- *F1 Score*: Harmonic mean of precision and recall. It offers a balanced measure of the model's accuracy:

$$F1 = 2\frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}.$$

B. DL MODELS AND HYPERPARAMETERS

To establish a baseline for evaluating the performance of the curated MTF data set, we performed assessments using five well-known DL classification models. These models vary in depth and complexity. Table 5 provides an overview of the models chosen, including their number of parameters and the year they were published. Furthermore, the table summarizes the main purpose behind their design and highlights their advantages at time of their release.

TABLE 5. DL models used for evaluation.

Models	Parameters	Year	Purpose
MobileNetV3 [27]	4,509,472	2019	To run in embedded systems. Reduced number of parameters.
AlexNet [28]	57,987,120	2012	Extremely capable DL model at that time, to be trained on the ImageNet [29] data set.
ResNet50 [30]	23,999,792	2016	Reduced number of parameters and matrix multiplications. Enables much faster training of each layer. Uses a stack of three layers rather than two.
VGG16 [31]	135,243,824	2014	Presents the small (3×3) convolution filters. Introduces the deep network with 16 layers.
ConvNeXT [32]	87,812,464	2022	Introduces the Vision Transformer into image processing through depthwise convolution.

To ensure a fair evaluation across the different models, we used the same hyperparameters during both training and evaluation. However, a minor exception was made for the batch size. Due to resource limitations, we adjusted the batch size to be smaller for the models with a large number of parameters, allowing them to run efficiently on the available hardware. Conversely, for the smaller models, we increased the batch size to optimize time efficiency. This approach enabled us to conduct a comprehensive and equitable assessment of the performance of the models.

Specifically, the chosen hyperparameters included the cross-entropy loss function [33], the Adam optimizer [34], and an automated learning rate scheduler [35]. The initial learning rate was set at 0.001, and it decreased by a factor of 0.1 every 20 epochs. The training process was limited to a maximum of 100 epochs, with early stopping implemented if the training loss did not improve for five consecutive epochs.

The batch size for the three smaller models was set to 128 images, while for the two larger models it was reduced to 32 images.

The experiments were carried out on a computer equipped with an AMD Ryzen 5 3600 CPU running at a base speed of 3.6 GHz, 32 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 12 GB of dedicated RAM.

C. RESULTS

We first report on the performance of the five DL models on the curated MTF data set. All experiments were conducted using the predefined models available in the PyTorch [36] library. For each of the four tasks of the data set, we report the results obtained from random guessing, the performance of the five models when trained from scratch on the data set, and the performance of the five models when fine-tuned on the curated MTF data set using pre-trained weights provided by the PyTorch library. These pre-trained weights were obtained by training the model using the ImageNet data set [29].

1) FACE RECOGNITION TASK

Table 6 reports the performance of DL models on the curated MTF data set for the face recognition task, which comprises 240 labels.

TABLE 6. Performance of the DL models on the curated MTF data set for the face recognition task. Boldface figures are the best in each column. S denotes that the model is trained from scratch, while P denotes that the model was pre-trained.

Model	Origin	Accuracy	Precision	Recall	F1 score	Training time (minutes)/ number of epochs
Random guess	N/A	0.16%	0.00%	0.42%	0.00%	N/A
MobileNet_v3	S	2.88%	0.01%	0.42%	0.02%	**94.77** / 70
	P	57.35%	49.77%	46.16%	43.88%	94.99 /71
AlexNet	S	18.93%	8.71%	10.76%	7.97%	95.56 /76
	P	36.66%	27.34%	25.81%	23.94%	101.96 / 82
ResNet 50	S	9.58%	3.13%	4.30%	3.13%	168.24 / 100
	P	58.39%	46.94%	46.01%	44.61%	151.95 / 94
VGG16	S	23.96%	17.89%	16.08%	14.57%	184.13 / 70
	P	34.66%	24.21%	23.63%	21.23%	156.77 / 74
ConvNeXT	S	13.26%	5.13%	6.60%	4.60%	216.9 / 71
	P	**79.87%**	**76.29%**	**73.31%**	**74.07%**	201.74 / 76

As expected, all trained models achieved better results than random guess. However, the MobileNet_v3 model trained from scratch exhibited subpar performance, just slightly above the random baseline (accuracy 2.88% and F1 score 0.02%). Remarkably, its low recall of just 0.42% indicates that this model predicted all images on a single label, suggesting that it was not able to accurately extract image features for precise predictions. This outcome for MobileNet_v3 was not surprising because this model was designed for simple and fast training in embedded systems. The other models trained from scratch exhibited better performance than the aforementioned model, with higher recall scores implying predictions across multiple labels and potentially benefiting from increased training epochs and higher learning rates.

As anticipated, all fine-tuned models with pre-trained weights outperformed their scratch-trained counterparts. This can be attributed to the pre-trained layers of the models, which are adept at identifying important image features and passing them to the model's classifier. The recall scores of the pre-trained models indicate successful predictions across multiple correct labels. Taking into account the uneven distribution of the data, the ConvNeXT model stood out with high results, with an accuracy of 79.78% and an F1 score of 74.07% due to its novelty and ability to tackle complex tasks using the Vision Transformer. Despite this model showing high performance on the FR task, the complexity of this highly multidimensional classification challenge presents significant difficulties for all other models, including the ConvNeXT model when trained from scratch. As a result, the MTF data set emerges as a valuable asset for encouraging and supporting future research in this area.

When examining the runtime required to train the models, we observe that each pre-trained model exhibits a very similar training duration if compared to its counterpart trained from scratch. The slight discrepancy in training times can be attributed to the variation in the number of epochs required to achieve the final model. The training time is contingent on the complexity of the model and the number of parameters. The longer training times of ResNet50 and ConvNeXT are due to the former using residual layers and the latter employing Vision Transformers, which contribute to their higher complexity. Furthermore, the smaller batch size used in the training of VGG16 and ConvNeXT played a role in the extended training time required by these models.

2) RACE CLASSIFICATION

The race classification task involves a four-label classification objective. The relative differences across models follow what was observed in the previous FR task.

TABLE 7. Performance of the DL models on the curated MTF data set for the race classification task. Boldface figures are the best in each column. S denotes that the model is trained from scratch, while P denotes that the model was pre-trained.

Model	Origin	Accuracy	Precision	Recall	F1 score	Training time (minutes)/ number of epochs
Random guess	N/A	15.34%	3.83%	25.00%	6.65%	N/A
MobileNet_v3	S	40.97%	10.24%	25.00%	14.53%	**92.4** / 70
	P	85.62%	83.12%	81.09%	82.01%	98.97 / 72
AlexNet	S	67.01%	60.13%	55.31%	56.23%	92.97 / 73
	P	87.22%	85.82%	84.49%	85.08%	94.22 / 78
ResNet 50	S	60.06%	44.24%	45.46%	43.39%	106.84 / 74
	P	90.18%	88.95%	87.16%	87.98%	103.1 / 71
VGG16	S	78.04%	74.65%	71.36%	72.70%	134.27 / 70
	P	92.33%	90.66%	90.95%	90.78%	177.13 / 74
ConvNeXT	S	61.50%	51.40%	48.43%	48.16%	233.86 / 82
	P	**95.77%**	**94.41%**	**94.94%**	**94.65%**	207.07 / 70

As shown in Table 7, all models demonstrated better performance than the random guess by the end of their training, whether pre-trained or trained from scratch. The training times to reach their final form were similar to those shown in Table 6. The slight differences in training times among tasks were due to varying numbers of training epochs, after which all models reached the stopping condition of non-improved loss for five successive epochs.

The curated data set exhibits an imbalanced nature for this task, with two majority labels: Asian (Chinese/Korean) accounting for 32.38% of the data, and White comprising 40.97% of the data. On the other hand, two minority labels are present: Asian (Indian) making up 15.63% of the data, and Black representing 11.02% of the data. This imbalance

severely affected the MobileNet_v3 model when trained from scratch: its recall was just 25%, which indicates that all images were predicted to belong to a single label. That behavior could be attributed to the simplicity of the model and the low number of parameters, which prevented the model from handling such complex data. Moreover, the other models trained from scratch (ResNet_50, AlexNet, and ConvNeXT) were also affected by the data distribution, with accuracies lower than 70% and recall scores hovering around 50%, which suggest they were primarily classifying images into two labels. The exception was the VGG16 model, which achieved a relatively high accuracy of 78.04% and a recall score of 71.36%, therefore indicating a successful multi-label image classification.

If we examine the pre-trained models, all of them outperformed their trained-from-scratch counterparts, boasting recall scores over 81%. This demonstrates their ability to correctly classify images into the four labels and their resilience to the data set imbalance. The most complex models, ConvNeXT and VGG16, achieved impressively high accuracy and F1 scores, both exceeding 90%. Their accuracies were 95.77% and 92.33%, respectively, with F1 scores of 94.65% and 90.78%.

3) GENDER CLASSIFICATION

The gender classification task in the curated MTF data set involves a binary classification problem with a well-balanced data distribution: approximately 47.12% of the images are labeled as males, and 52.88% are labeled as females in both the training and the test sets. Thus, GC should be an 'easy' problem for most DL models, as the results reported in Table 8 suggest.

TABLE 8. Performance of the DL models on the curated MTF data set for the gender classification task. Boldface figures are the best in each column. S denotes that the model is trained from scratch, while P denotes that the model was pre-trained.

Model	Origin	Accuracy	Precision	Recall	F1 score	Training time (minutes)/ number of epochs
Random guess	N/A	45.61%	40.60%	43.88%	39.18%	N/A
MobileNet_v3	S	52.88%	26.44%	50.00%	34.59%	**80.3** / 59
	P	76.60%	77.94%	75.89%	75.93%	100.74 / 74
AlexNet	S	70.13%	71.35%	69.28%	69.05%	90.33 / 71
	P	88.18%	88.16%	88.11%	88.13%	92.99 / 73
ResNet 50	S	70.85%	70.80%	70.57%	70.62%	142.16 / 78
	P	97.60%	97.59%	97.61%	97.60%	127.24 / 71
VGG16	S	90.65%	90.61%	90.67%	90.63%	181.9 / 76
	P	97.28%	97.29%	97.26%	97.27%	211.4 / 80
ConvNeXT	S	73.80%	73.92%	73.97%	73.80%	209.88 / 71
	P	**98.88%**	**98.86%**	**98.90%**	**98.88%**	219.22 / 74

When MobileNet_v3 was trained from scratch, it exhibited low performance, as it did in the previous tasks: it achieved an accuracy of 52.88%, and recall of 50%, indicating that it labeled all images as females. Thus, the proposed curated MTF data set contains images more complicated than the MobileNet_v3 model can handle even for the easiest task.

However, the remaining trained-from-scratch models performed better, with accuracies exceeding 70% and recall scores over 69%. Among them, VGG16 stood out, achieving impressive results with an accuracy of 90.65% and a recall of 90.67%. The complexity of the VGG16 network and its extensive parameter count allowed it to accurately classify the images.

On the other hand, models fine-tuned with pre-trained weights demonstrated superior performance compared to the previously discussed models. Three of these models achieved accuracy above 97%. Ranging from best to worst accuracy, these models are ConvNeXT, ResNet 50, and VGG16, with accuracies of 98.88%, 97.60%, and 97.28% respectively, and F1 scores of 98.88%, 97.60%, and 97.27%, respectively. These outstanding results can be attributed to the capabilities of these models, including their large number of parameters and their use of novel layers such as residual and transformer layers.

Regarding the training runtimes, they are comparable to those of the previous tasks for each respective model, taking into consideration the number of training epochs needed for each model to converge, as well as the differences in batch sizes used.

4) AGE CLASSIFICATION

The binary age classification task in the curated MTF data set exhibits a significant bias, with 89.94% of the data labeled as 'young' and only 10.06% labeled as 'old' in both the training and test sets.

TABLE 9. Performance of the DL models on the curated MTF data set for the age classification task. Bold figures are the best in each column. S denotes that the model is trained from scratch, while P denotes that the model was pre-trained.

Model	Origin	Accuracy	Precision	Recall	F1 score	Training time (minutes)/ number of epochs
Random guess	N/A	39.06%	51.80%	54.49%	35.27%	N/A
MobileNet_v3	S	90.01%	44.97%	50.00%	47.35%	101.19 / 72
	P	94.81%	91.06%	78.08%	83.03%	96.67 / 70
AlexNet	S	90.50%	86.20%	53.48%	54.06%	**94.67** / 65
	P	95.93%	90.71%	85.75%	88.03%	96.86 / 67
ResNet 50	S	90.02%	78.37%	50.75%	48.92%	110.96 / 70
	P	95.85%	89.14%	87.47%	88.28%	140 / 77
VGG16	S	92.81%	90.26%	66.40%	72.21%	168.09 / 78
	P	96.33%	91.82%	87.03%	89.24%	149.55 / 72
ConvNeXT	S	90.50%	86.20%	53.48%	54.06%	211.75 / 75
	P	**97.60%**	**95.13%**	**91.27%**	**93.09%**	220.09 / 73

In Table 9 we can see that the high accuracy levels reached (above 90%) did not guarantee accurate predictions, as the data imbalance led to biased models. This is evident when analyzing the precision, recall, and F1 score, where biased models had low scores hovering around 50%, revealing the impact of imbalanced predictions. During training from scratch, only the VGG16 model appeared unbiased. This observation suggests that smaller models and models with complex novel layers are more susceptible to bias when trained from scratch.

On the other hand, fine-tuned models with pre-trained weights exhibited better resistance to bias, as evidenced by the recall scores. All pre-trained models reported recall rates

greater than 78%, indicating the accurate classification of images into the two labels. Notably, ConvNeXT stood out with an impressive accuracy of 97.60% and an F1 score of 93.09%, making it the best-performing model among those we tried.

The reported training times were again consistent with those observed during the training of other tasks. This consistency is attributed to the use of the same number of images and models across all tasks.

D. CURATED VS. NON-CURATED DATA

As discussed in Section III-B, the curated MTF data underwent meticulous manual processing whereby a large amount of low-quality, inappropriate, or noisy images were filtered out. To illustrate the benefits of this manual processing, especially for training classification models, in this section we compare the performance of the best-performing DL model (ConvNeXT with pre-training) on the four tasks when trained on the two MTF data sets (non-curated and curated). The non-curated data set was automatically split into a training set encompassing 70% of the total data, and a validation set that includes the remaining 30%. The results of this experiment are presented in Table 10.

TABLE 10. Comparison of ConvNeXT with pre-training on the non-curated MTF data set vs the curated MTF data set.

Task	Training data	Accuracy	Precision	Recall	F1 score	Training time (minutes)/ number of epochs
FR	Non-curated	10.07%	8.85%	8.09%	7.01%	4089.72 / 94
	Curated	79.87%	76.29%	73.31%	74.07%	201.74 / 76
RC	Non-curated	65.38%	59.89%	58.67%	59.18%	3809.95 / 87
	Curated	95.77%	94.41%	94.94%	94.65%	207.07 / 70
GC	Non-curated	69.46%	69.38%	69.32%	69.34%	4094.41 / 93
	Curated	98.88%	98.86%	98.90%	98.88%	219.22 / 74
AC	Non-curated	62.57%	61.70%	61.57%	61.62%	3911.01 / 89
	Curated	97.60%	95.13%	91.27%	93.09%	220.09 / 73

To enable a fair comparison for the FR task, we only considered the 240 identities in the curated MTF data set when training the model on the non-curated version. The ConvNeXT model trained on the non-curated data achieved a mere 10% accuracy. In contrast, the model trained on the curated data exhibited an impressive 79.87% accuracy in the 240-label classification. The disparity in performance is evident from the F1 score, where the model trained on the curated data outperformed the model trained on the non-curated data by an order of magnitude. The observed discrepancy in performance was particularly expected in the FR task, given the intended use of both data sets, the key role of data processing in the curated data to discard a large number of crawled images that did not correspond to the correct identity, and the substantial amount of noise in the raw data. These factors explain the large gap in results between the two data sets. From these results, it is clear that only the curated MTF data set should be used for FR tasks.

We limited the comparison on the non-curated data set to the ConvNeXT model, as it consistently outperformed the other models in the curated data experiments across all tasks. This choice was made to provide a strong upper-bound baseline while avoiding the expensive computational costs associated with retraining multiple models on a much larger dataset.

The reported results are based on a single training run per model and task. While this approach provides a consistent baseline for comparison, it does not capture the variability that may arise due to random weight initialization or mini-batch sampling. Researchers may consider running each experiment multiple times with different seeds and reporting averages and standard deviations to better assess the robustness and stability of their model.

For the remaining three tasks (race, gender, and age classification) we observe smaller differences. The model trained on the non-curated data reported accuracy levels ranging from 62% to 69%, while the model trained on the curated data reported accuracies higher than 95%. Although these three tasks were comparatively easier for the models to learn, and the occurrence of incorrect images crawled from the search engine was low, the impact of manual processing in improving the performance of classification models was evident.

Training runtimes were also proportional to the size of the data. In this case, the non-curated data set resulted in around 20 times higher training costs than the curated data.

In summary, whereas the curated MTF data set should be used when model performance in data classification is the main goal, the non-curated version would be better suited for data intensive and unsupervised tasks, such as building general or generative AI models.

VI. ETHICAL AND LEGAL CONSIDERATIONS

The MTF data sets were compiled with strict adherence to ethical and legal requirements, particularly those outlined in the EU GDPR. To ensure a lawful image collection, we exclusively sourced face images of publicly recognized celebrities who had voluntarily made their images publicly available under permissive licenses. Only images with explicit copyright permissions, public domain or Creative Commons licenses that allow modification and reuse were included.

Furthermore, the use of facial images was approved by the BOEL of the SoBigData++ H2020 project. This ethical oversight ensured that the data collection, processing, and labeling procedures were in accordance with established guidelines for privacy, fairness, and transparency. The curated version of the dataset further emphasizes quality and appropriateness by manually excluding images that were synthetic, misleading, or potentially harmful.

VII. CONCLUSION AND FUTURE WORK

We have presented two Multi-Task Faces (MTF) data sets, one of them being the curated version of the other, and both containing labeled images of human faces. Unlike similar data sets consisting of human faces which are

at risk of being removed from the internet for privacy reasons, MTF data were compiled with a view to longevity thanks to compliance with the legal requirements of the GDPR and strict ethical considerations. MTF data exclusively focus on publicly recognized celebrities' images that were either public-domain or published under Creative Commons licenses granting permission for modification, sharing, and commercial use. In this paper, we have elaborated on the data preparation processes, while also presenting comprehensive statistical information and empirical results that aid to better understand the quality and suitable uses of each data set.

We applied the curated MTF data set to four classification tasks using five well-known deep learning models. The ConvNeXT model achieved the best performance in all tasks, with an accuracy of 98.88% for gender classification, 95.77% for race classification, 97.60% for age classification and 79.87% for face recognition. These results demonstrate that the curated MTF data set offers high-quality labels suitable for training robust AI models. In contrast, the non-curated version showed significantly lower performance, underscoring the importance of careful data curation in building effective and fair machine learning pipelines.

As a future work, we plan to apply and test the MTF data sets to other tasks, such as face anonymization. In particular, the FR task in the curated MTF data set can be used to measure the residual re-identification risk of image anonymization algorithms, whereas the accuracy achieved for the other three tasks on the anonymized images would quantify the degree of utility preserved. In addition, we aim to test the ability of the non-curated MTF data set to train generative models of synthesized human faces.

ACKNOWLEDGMENT

The authors appreciate the effort and generosity of the owners of the images they used to make these public domain or release them under the most flexible Creative Commons license. The owners of the images are acknowledged in a CSV file named ''URL_of_original_images'' and included in the zip file of both MTF data sets that links all the original sources of the images.

REFERENCES

[1] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, and V. Munigala, ''Overview and importance of data quality for machine learning tasks,'' in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3561–3562.

[2] R. Ducato and A. Strowel, ''Limitations to text and data mining and consumer empowerment: Making the case for a right to 'machine legibility,'' *IIC-Int. Rev. Intellectual Property Competition Law*, vol. 50, no. 6, pp. 649–684, 2019.

[3] G. Voyatzis and I. Pitas, ''Protecting digital image copyrights: A framework,'' *IEEE Comput. Graph. Appl.*, vol. 19, no. 1, pp. 18–24, Aug. 1999.

[4] K. A. Philbrick, A. D. Weston, Z. Akkus, T. L. Kline, P. Korfiatis, T. Sakinis, P. Kostandy, A. Boonrod, A. Zeinoddini, N. Takahashi, and B. J. Erickson, ''RIL-contour: A medical imaging dataset annotation tool for and with deep learning,'' *J. Digit. Imag.*, vol. 32, no. 4, pp. 571–581, Aug. 2019.

[5] P. Voigt and A. Von dem Bussche, ''The EU general data protection regulation (GDPR),'' in *A Practical Guide*, vol. 10, 1st ed., Cham, Switzerland: Springer, 2017, pp. 10–5555.

[6] N. A. Smuha, ''The EU approach to ethics guidelines for trustworthy artificial intelligence,'' *Comput. Law Rev. Int.*, vol. 20, no. 4, pp. 97–106, Aug. 2019.

[7] F. Boutros, V. Struc, J. Fierrez, and N. Damer, ''Synthetic data for face recognition: Current state and future prospects,'' *Image Vis. Comput.*, vol. 135, Jul. 2023, Art. no. 104688.

[8] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, ''Deep learning-based facial emotion recognition for human–computer interaction applications,'' *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, Nov. 2023.

[9] S. K. Gupta and N. Nain, ''Review: Single attribute and multi attribute facial gender and age estimation,'' *Multimedia Tools Appl.*, vol. 82, no. 1, pp. 1289–1311, Jan. 2023.

[10] R. Kumar, K. Singh, D. P. Mahato, and U. Gupta, ''Face-based age and gender classification using deep learning model,'' in *Proc. PSIVT Int. Workshops Image Video Technol.*, vol. 235. Cham, Switzerland: Springer, Jan. 2024, pp. 2985–2995.

[11] G. Sanil, K. Prakash, S. Prabhu, V. C. Nayak, and S. Sengupta, ''2D-3D facial image analysis for identification of facial features using machine learning algorithms with hyper-parameter optimization for forensics applications,'' *IEEE Access*, vol. 11, pp. 82521–82538, 2023.

[12] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, ''Attribute-preserving face dataset anonymization via latent code optimization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8001–8010.

[13] T. Mitchell. (Jun. 1999). *CMU Face Images*. [Online]. Available: http://archive.ics.uci.edu/dataset/124/cmu+face+images

[14] Z. Liu, P. Luo, X. Wang, and X. Tang, ''Deep learning face attributes in the wild,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[15] E. Rosati, *Copyright and the Court of Justice of the European Union*. Oxford, U.K.: Oxford Univ. Press, 2019.

[16] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, ''DigiFace-1M: 1 million digital face images for face recognition,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3515–3524.

[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, ''Labeled faces in the wild: A database for studying face recognition in unconstrained environments,'' Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 7-49, Oct. 2007.

[18] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, ''The MegaFace benchmark: 1 million faces for recognition at scale,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.

[19] S. Yang, P. Luo, C. C. Loy, and X. Tang, ''WIDER FACE: A face detection benchmark,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[20] B. Yang, J. Yan, Z. Lei, and S. Z. Li, ''Fine-grained evaluation on face detection in the wild,'' in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–7.

[21] R. Rothe, R. Timofte, and L. Van Gool, ''Deep expectation of real and apparent age from a single image without facial landmarks,'' *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 144–157, Apr. 2018.

[22] T. Karras, S. Laine, and T. Aila, ''A style-based generator architecture for generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405. [Online]. Available: https://github.com/NVlabs/ffhq-dataset

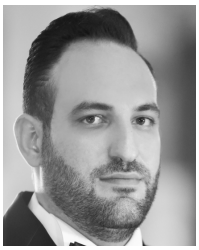
[23] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, ''VGGFace2: A dataset for recognising faces across pose and age,'' in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[24] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, ''UMDFaces: An annotated face dataset for training deep networks,'' in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 464–473.

[25] P. Viola and M. Jones, ''Rapid object detection using a boosted cascade of simple features,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, vol. 1, Dec. 2001, pp. I-511–I-518.

[26] A. Mordvintsev and K. Abid. (2017). *OpenCV-Python Tutorials Documentation: Release 1*. [Online]. Available: https://opencv24-python-tutorials.readthedocs.io//downloads/en/stable/pdf/

[27] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst., 26th Annu. Conf. Neural Inf. Process. Syst.*, vol. 60, Lake Tahoe, NV, USA, May 2017, pp. 84–90.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] Simonyan, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Jan. 2015.
[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.
[33] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Dec. 2014.
[35] C. Kim, S. Kim, J. Kim, D. Lee, and S. Kim, "Automated learning rate scheduler for large-batch training," 2021, *arXiv:2107.05855*.
[36] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," in *Programming With TensorFlow: Solution for Edge Computing Applications*. Springer, 2021, pp. 87–104.



RAMI HAFFAR received the M.Sc. and Ph.D. degrees in computer science from Rovira i Virgili University, Tarragona, Catalonia, in 2019 and 2023, respectively. He is currently a Postdoctoral Researcher with Rovira i Virgili University. His research interests include explainable AI, machine learning, and image anonymization.

DAVID SÁNCHEZ (Senior Member, IEEE) received the Ph.D. degree in computer science from the Technical University of Catalonia, Barcelona, in 2008. He is currently a Full Professor and an ICREA-Academia Researcher with Universitat Rovira i Virgili, Tarragona, Catalonia. He has participated in several national and European funded research projects and has authored several papers and conference contributions. His research interests include data semantics, ontologies, machine learning, and data privacy.

JOSEP DOMINGO-FERRER (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the Autonomous University of Barcelona in 1988 and 1991, respectively. He holds the B.Sc. and M.Sc. degrees in mathematics. He is a Distinguished Full Professor of computer science and an ICREA-Academia Researcher with Universitat Rovira i Virgili, Tarragona, Catalonia, where he also leads CYBERCAT. His research interests include privacy, security, and trustworthy AI.