Fairness is in the details : Face Dataset Auditing

Valentin Lafargue[1,2,3] (✉), Emmanuelle Claeys[4], and Jean-Michel Loubes[2,3]

[1] Institut de Mathématiques de Toulouse, France
valentin.lafargue@math.univ-toulouse.fr
[2] Institut national de recherche en sciences et technologies du numérique
[3] Artificial and Natural Intelligence Toulouse Institute 2
[4] Institut de Recherche en Informatique de Toulouse

Abstract. Auditing involves verifying the proper implementation of a given policy. As such, auditing is essential for ensuring compliance with the principles of fairness, equity, and transparency mandated by the European Union's AI Act. Moreover, biases present during the training phase of a learning system can persist in the modeling process and result in discrimination against certain subgroups of individuals when the model is deployed in production. Assessing bias in image datasets is a particularly complex task, as it first requires a feature extraction step, then to consider the extraction's quality in the statistical tests. This paper proposes a robust methodology for auditing image datasets based on so-called "sensitive" features, such as gender, age, and ethnicity. The proposed methodology consists of both a feature extraction phase and a statistical analysis phase. The first phase introduces a novel convolutional neural network (CNN) architecture specifically designed for extracting sensitive features with a limited number of manual annotations. The second phase compares the distributions of sensitive features across subgroups using a novel statistical test that accounts for the imprecision of the feature extraction model. Our pipeline constitutes a comprehensive and fully automated methodology for dataset auditing. We illustrate our approach using two manually annotated datasets.[5]

Keywords: Audit, Images, Bias, Distribution, Statistical test, Uncertainty, Classification, Fitzpatrick, Gender, Age

1 Introduction

The widespread adoption of machine learning (ML) systems in industrial applications has heightened concerns about fairness, transparency, and accountability. The issue of bias in algorithmic decision-making has emerged as a critical concern within the machine learning community. A substantial body of research has examined how such biases can adversely impact algorithmic outcomes, potentially leading to violations of fundamental rights, as highlighted in the European AI Act. This legislation highlights the need to prevent AI systems from perpetuating or exacerbating existing societal inequities through systematic bias

[5] Code and datasets available at github.com/ValentinLafargue/FairnessDetails

analysis. These biases not only compromise fairness but also raise ethical and legal challenges, underscoring the need for rigorous detection through systematic audit processes to ensure accountability and mitigate unintended harms. We refer for instance to [40], [3], [21], [9], [44], [17], [25] or [28]. Beyond decision-making contexts, we know that algorithmic biases often stem from biases present in the training datasets themselves. Auditing an image dataset is a challenge in itself. Firstly, it is necessary to determine which variables to consider and how to extract them from an image. The importance of auditing image datasets is amplified by the fact that every image inherently encodes explicit features. Unlike text or numerical datasets, which can omit or abstract sensitive details, images visually represent specific characteristics, often revealing cues about sensitive features such as ethnicity, age, and gender. Manual labeling of such features is prohibitively expensive when dealing with large-scale datasets or when conducting extensive audits across multiple variables. To address this issue, convolutional neural networks (CNNs) can be employed to predict sensitive features, although they require annotated data for their training (lesser amount). Once trained, the network can predict the sensitive feature of the remaining data in the dataset (with a certain error relative to it). In our context, we define bias in an image dataset as statistically significative difference of distributions (e.g., an ethnicity or an age group under-represented). Statistical tests usually do not take into account the uncertainty of the labels (false predictions). We propose a prediction-aware testing pipeline that evaluates the underlying characteristic of a dataset while accounting for the model's imprecision during statistical analysis. Considering the model's accuracy in our testing pipeline helps minimize the required manual labeling, enabling large-scale auditing. The Section 2 presents the literature review about the sensitive feature extraction method and about the error-robust statistical testing. The Section 3 introduces the datasets used and our manual annotation procedure, the Section 4 explains our feature extraction and classification methodology, the Section 5 presents our error-aware testing protocol, then the Section 6 highlights our results. Section 7 concludes with some perspectives and future work.

2 Related Works

Assessing bias in image datasets requires careful consideration of several aspects. First, the dataset contains potentially sensitive variables. Some features must be extracted to serve as proxy estimates for these variables. Based on these features, an auditing pipeline generates reports on diversity and representativeness using selected metrics or statistical tests. The following subsections provide an overview of general concepts from the literature related to each of these aspects.

2.1 Choice of possible sensitive variables

Ethnic classification refers to the classification of individuals into distinct groups based on perceived physical characteristics, such as skin color, hair texture, and facial shape. Many academic datasets separate images into at least five categories: Latino, Asian, White, Black, and Other. This classification is common in many reference datasets such as the Adult dataset [7] and is derived from the US Census 2000 classification. In [27], the authors criticize methodologies that rely exclusively on race as a variable, arguing that this approach is overly restrictive.

An alternative to the Census 2000 classification is to use medical skin analysis criteria. In [10], the authors presented a method using the ITA (Individual Typology Angle) algorithm [48, 38] to estimate skin tone in the context of classifying skin lesions and to normalize the impact of lighting variations on facial images. Similarly, the Fitzpatrick classification, introduced by [22], classifies individuals based on their skin's reaction to sun exposure. This classification has six classes and takes into account features such as skin color, the presence of freckles, hair and eye color, and reactions to sun exposure (precise definition and the demographic distribution are in the Appendix, Section A). Inspired by [12], we believe that the Fitzpatrick scale is well-defined as it stems from its dermatology origin. This thorough definition paired is with its popularity justify in our opinion its usage in the context of auditing.

The authors of [46] recommend considering ethnicity as a color shade, in particular to use the newly created Monk Skin Tone Scale [39]. However, the wide range of shades makes it challenging to separate groups and, consequently, to identify bias. Once the features are selected, the next step is to automate their extraction from the image dataset.

2.2 Model for dataset labelisation

Depending on the size of the dataset, manual extraction may be time-consuming and challenging, prompting the use of a classifier to automatically annotate part of the dataset. CNNs are particularly well suited to images, as they can extract visually identifiable features. From a face image, CNNs can capture skin tone as a set of pixel colors or as ITA. However, this information alone omits ethnic features [48] such as hair texture or face shape, which can reduce the accuracy of ethnic classification. This highlights the need for image segmentation. Therefore, the chosen architecture should identify areas that contain these features, while excluding irrelevant areas such as the background. Among existing methods [23], the FairFace architecture [33] detects faces and classifies age and gender using a ResNet34 architecture [29]. A variant approach in [43] employs a nested U-Net architecture called U^2-Net. Finally, interest has brewed around understanding and guiding the CNNs by understanding how the networks treat facial characteristics [52]. A segmentation of the skin region can be achieved using DeepLabv3 [15] with a MobileNetV3 Large Backbone model [30] pretrained on Celeb-HQ [35]. An extension proposed by [48, 38] estimates ITA values. More precisely, after smoothing the image and applying a skin mask, the authors applied a K-means clustering on the pixels values and kept the one with the highest luminosity to extract the ITA values. Finally, [2] trained a CNN from scratch to classify skin pixel shades into 10 classes. However, none of these methods explore the impact of training dataset size which is crucial when auditing, as underlined in [16], or provide specific configurations for the Fitzpatrick classification.

2.3 Metrics and statistical tests

Once features have been extracted from the dataset and transformed into variables, they are used to group individuals based on these variables. The fairness auditing process then evaluates whether certain groups are over- or under-represented in comparison to predefined parameters, which may include equal or official proportions. This parameter ensures the preservation of the so-called *diversity* [18], such as maintaining almost equal frequencies between different groups. Consider a dataset $\mathcal{D}$ of observations composed of p variables : $X^0, \ldots, X^{p-1}$. Let X^0 be a variable that may convey bias (e.g., ethnicity or age), and X^j be a variable that may induce disparity or the bias representation (e.g., gender). We focus on the conditional distribution of X^0 given X^j, denoted as $\mathcal{L}(X^0|X^j)$ or when no ambiguity is possible $\{X^0|X^j\}$.

The first measure of fairness aims at quantifying the diversity in the dataset. For this, a diversity loss is introduced in [51]. Given classes $\{1, \ldots, k\}$ with target frequencies $f_i \left(\sum_{i=1}^{k} f_i = 1\right)$, and real frequencies $f'_1, \ldots, f'_k$, the diversity loss Δ is defined as $\Delta := 1 - \inf_{f_i > 0} f'_i / f_i$. Hence, it computes a ratio $\Delta \in [0, 1]$ where a value of Δ close to 1 means that at least one group is highly under-represented. Unfortunately, this metric focuses solely on one unrepresented group. For discrete categorical variables, diversity can be evaluated using Conditional Shannon entropy. The Conditional Shannon entropy distribution $C(S)$ of a subset $S \subseteq X^0$ is defined as:

$$C(S) = -\sum_{i=1}^{k} f_i \log f_i$$

where x_i^j is a possible modality of X^j and $s_i = \frac{|S| \cap |X^j = x_i^j|}{|S|}$ is the probability to observe S according $X^j = x_i^j$. Equally distributed entropy according to X^j corresponds to good diversity. Both of the aforementioned metrics cannot be extended to cases where the space of conditional observations is large and are not related to a statistical test [13].

The second main measure of fairness for such problems comes from a volumetric perspective comparison. Actually, Geometric diversity [18] provides a meaningful similarity measure for observations in multiple dimensions. Consider each data point of the dataset $x \in X$, represented by a variable vector v_x. The geometric diversity of a subset $S \subseteq X$ is defined as the n-volume of the parallelotope spanned by the p variable vectors $\{v_x : x \in S\}$, where $n = |S|$ is the size of the subset. Denoting the data matrix of the subset S as $\mathbf{D} \in \mathbb{R}^{p \times n}$, the (squared) n-volume of the n parallelogram embedded in p dimensional space can be computed by means of the determinant of the Gramian matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ (with variable vectors as columns in $\mathbf{D}$). Thus, the geometric diversity can be measured by :

$$G(S) = \sqrt{\mathrm{Det}(\mathbf{D}^T\mathbf{D})}$$

The larger $G(S)$, the more diverse is S in the variable space. However, Geometric Diversity cannot be applied if one aims to compare variable distributions using statistical hypothesis testing [14]. The Disparate Impact (DI) is one of the most used fairness metric, defined for a binary model $\hat{Y} = f(X)$ by the ratio

$$DI(f, S) := \frac{\min\big(\mathbb{P}(\hat{Y} = 1 \mid S = 0), \mathbb{P}(\hat{Y} = 1 \mid S = 1)\big)}{\max\big(\mathbb{P}(\hat{Y} = 1 \mid S = 0), \mathbb{P}(\hat{Y} = 1 \mid S = 1)\big)}$$

This quantity is equal to 1 when there is probabilistic independence between the model's decision $\hat{Y}$ and the sensitive variable S. The smaller the DI is, the more discrimination towards the minority class exist. Hence, several norms or regulations impose that a model should have its disparate impact greater than 0.8 as detailed in [26] or [50].

This metric generally used to evaluate the discrimination of a model, can be applied to evaluate the probabilistic bias of two sensitive variables. We choose to include it only in the Appendix (1) not to confuse the reader and make him think that we evaluate our model's fairness (for the instance the bias in the CNN predicting the Fitzpatrick Class), (2) homogeneity between the parity test (about one sensitive variable) where the DI is not applicable and the equal representation test (about two sensitive variables) where one might use the DI and (3) while relevant, we believe that testing a null hypothesis with multiple statistical tests is a more robust approach ; see Section D.3 in the Appendix.

Rather than relying on high-level metrics or aggregated scores, our approach evaluates biases by directly comparing the distributions of sensitive variables across subgroups.To quantify the distance between distributions with large, high-dimensional samples, one may measure the general Wasserstein distance, given by:

$$W_{\tilde{p}}(\mu, \nu) = \big(\inf_{\pi \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^{\tilde{p}} d\pi(x, y)\big)^{1/\tilde{p}}$$

where $\tilde{p} \geq 1$, $W_{\tilde{p}}$ is the $\tilde{p}^{\text{th}}$ Wasserstein distance, $\Gamma(\mu, \nu)$ denote all joint distributions π that have marginals μ and ν, $d()$ is the distance function between points x and y that matched and M is a given metric space. Using the Wasserstein distance, a classical distance-based test, such as the two-sample test (i.e. variables following the same distribution), can be applied following the tests proposed in [5] using the limit distributions developed in [4] or [6] and [47]. Other statistical tests, such as those based on averages or conditional averages, may also provide insights into the proximity of variable distributions [18]. Traditional statistical tests, such as Pearson's R, the t-test, and ANOVA, are commonly used. For non-normal data distributions, non-parametric tests, such as the χ^2 test, the Kolmogorov-Smirnov (KS) test, or the Central Limit Theorem (CLT) based test, serve as an alternative.

The previously mentioned metrics and tests do not account for the classification accuracy of feature extraction. Since feature extraction is performed automatically, as highlighted by [1], who evaluated how varying levels of label error (simulated through label flipping) affected the disparity metrics, it is essential to consider the model's accuracy in the bias detection task. Permutation

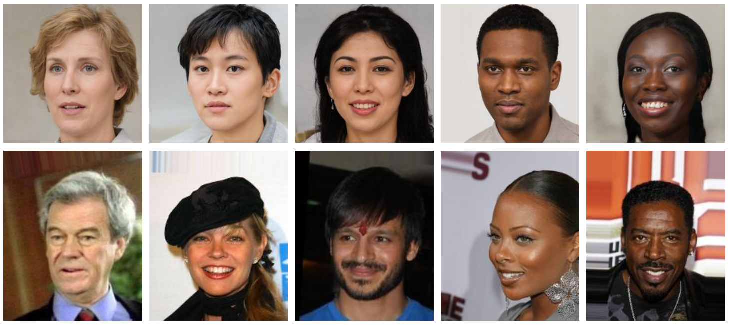


Fig. 1: Fitzpatrick classification (5 class) from the left to right Phototype I, II, III-IV, V, VI. The first row is from the GAN dataset, while the second is from the CelebA dataset, both dataset are released to the community.

methods have long been used in pursuit of robustness, as demonstrated by [20], who introduced a permutation-based fairness framework with labelled data. Although an extensive body of work has addressed label errors in the training set, to the best of our knowledge, no specific test for bias detection, that accounts for errors in the automated extraction of variables, has been proposed .

Our contribution, therefore, is to propose a full pipeline that starts with a variable extraction step and extends to robust statistical tests designed to consider the fairness according to the accuracy of the model's annotations. The following section details our complete methodology and introduces robust statistical techniques to highlight biases in images datasets.

3 Datasets and manual annotation

3.1 Datasets

To illustrate our pipeline, we rely on two datasets as guidelines, using them as the starting point of our process. These datasets are academic benchmark datasets of two different types. Generated Photos dataset [24, 11] is a synthetic images dataset sourced from a commercial platform and are generated using a GAN-based model [32]. The dataset intentionally encompasses a broad range of demographic features, including gender and ethnicity, with the GAN's hyperparameters calibrated to represent individuals with appearances associated with diverse ethnicities. Each image has been generated with Census 2000 labeling, ensuring an almost equal proportion of Caucasian, Asian, Hispanic or Latino, and Black populations. For our work, we utilized the academic version of this dataset, which contains 10,000 generated facial images. We also work on a well-known benchmark: The CelebA dataset [36] for comparative analysis. The CelebA dataset has approximately 200,000 celebrity images sourced from the Internet, annotated with multiple facial features. From this dataset, we randomly sampled 1,500 images to assess how our test performs on a smaller dataset.

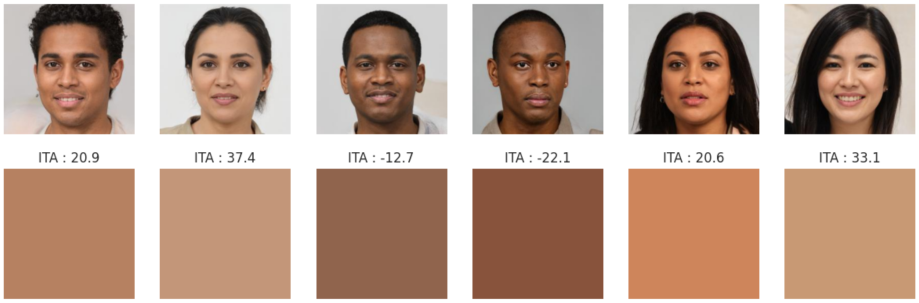
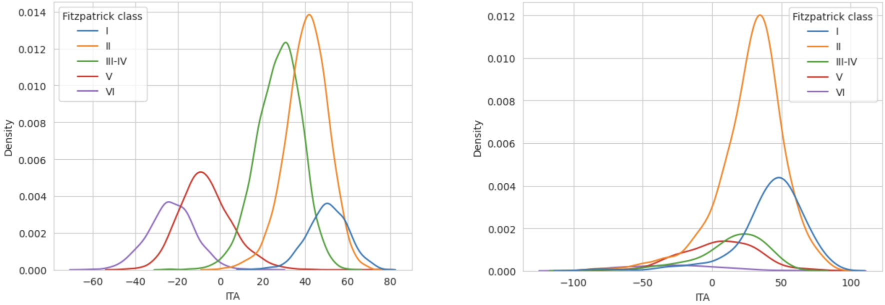


Fig. 2: Skin color extracted and Individual Typology Angle (ITA) of the GAN dataset



(a) GAN dataset

(b) Subset of the CelebA dataset

Fig. 3: Probability density of ITA given the Fitzpatrick class

3.2 Manual annotation

All images were manually labeled by three non-expert individuals according to the Fitzpatrick classification. This manual annotation helps assess how well our model aligns with a fully manual annotation. However, Phototypes III and IV are hardly distinguishable for non-experts and rarely reach full agreement. Thus, we chose to merge them. Fig. 1 gives some examples of our manual classification. As with ethnicity, a person's gender is determined by the majority vote of our three annotators. Even when considering the reflected gender, our dataset did not adequately represent the transgender or the bi-gender community, just to name a few, leading the annotators to classify the portrayal of gender to the limited view of gender binary notion that includes only men and women. In this regard, we view our work as part of initial studies towards auditing gender representation, which should further be extended in this direction in the future.

4 Sensitive variable classification using Neural Network

Manual auditing is not cost-effective for high- or medium-level auditing of large-scale datasets. This underscores the need for neural networks to predict sensitive variables accurately. Numerous manually labelled image datasets exist for binary gender classification - although we regret the lack of datasets with more diverse gender representations - facilitating the use of highly effective pre-trained networks for gender estimation.

4.1 Individual Typology Angle (ITA) estimation and link with Fitzpatrick

To improve the accuracy of the Fitzpatrick classification, we add an Individual Typology Angle (ITA) estimation step to our model, which can be seen as an enrichment of pixel information. ITA values are computationnaly derived from skin regions (isolated by pre-trained DeepLabv3). The estimate of the ITA value is based on 2 colorimetric parameters: the luminance $L*$ and the yellow/blue component $b*$. The ITA is defined as follows:

$$\text{ITA} = \arctan\left(\frac{L* - 50}{b*}\right) \times \frac{180}{\pi} \tag{1}$$

where a perceptual lightness at value 50 corresponds to a maximum chroma. We extracted the mean and standard deviation of the ITA values and of other colometric parameters. Fig. 2 presents examples of extracted ITA.

Fig. 3 gives the ITA distribution according the Fitzpatrick class and confirm the clear correlation between the Fitzpatrick classes and ITA values. Higher Fitzpatrick class numbers correspond to lower mean ITA values. However, as a single ITA score can be assigned to several Fitzpatrick classes, there is no one-to-one correspondence between the two. We further research the difference between the ITA and the Fitzpatrick class in the Appendix, Section B.

4.2 Gender, age and Fitzpatrick scale classification

We used the FairFace [31] method to classify gender and age, as its architecture provides the best results for these tasks. Despite certain limitations —such as detecting undesirable background faces- it achieved the best accuracy (see in the Appendix, Section E.2). Since our classification task relies primarily on facial features, especially skin, we studied the effect of applying masks to images before training our fine-tuned CNN. We tested three approaches (see in the Appendix, Fig. 9): (1) using the original images, (2) removing the background, and (3) isolating only the segmented skin region. The extracted ITA and skin-related information are incorporated as additional features in the latent layer of the neural network architecture.

In most of our experiments, we added a custom classification head to ResNet-50 or ResNet-101 embeddings [29] and fine-tuned these models on our labeled

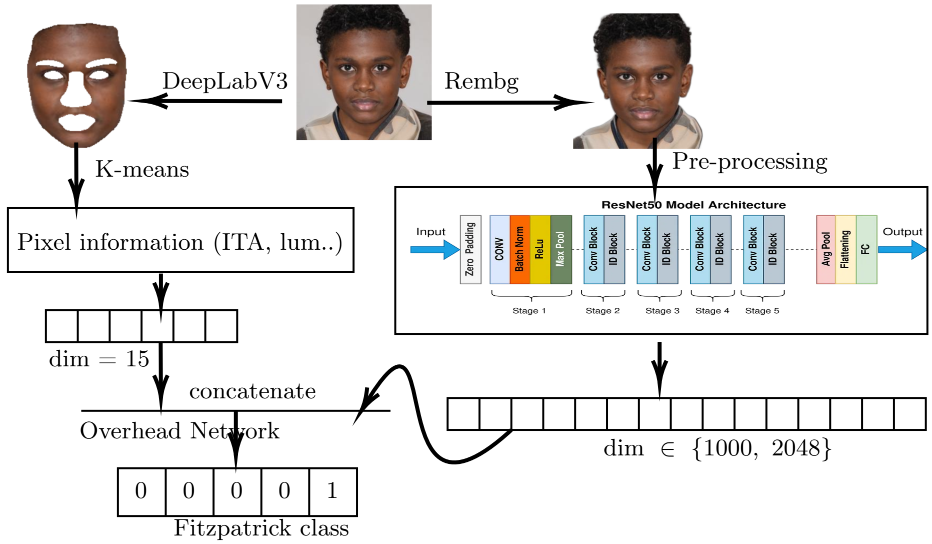


Fig. 4: Pipeline for our Fitzpatrick classification

dataset, eliminating the need for full CNN training. The impact of training set size is analyzed in Section 6. We use two feature extraction configurations: (1) the final dense layer's output (1,000-dimensional) and (2) the preceding layer's output (2,048-dimensional). Fig. 4 present an overview of our classification pipeline. More details on the architecture, optimizer, early stopping, compute time, and transfer learning method are provided in the Appendix, Section C.1. We created a Neural Network architecture to accurately predict the Fitzpatrick class with as few manually labeled image as possible, however, this part was not mandatory to our auditing framework thanks to the following section, which explains how we calibrate our testing pipeline given a model's accuracy.

5 Uncertainty aware statistical test

Statistical test used We find rejection based on variable distributions more meaningful, hence, $\mathcal{H}_0$ assumes that both groups are drawn from the same underlying distribution. In our proposed methodology, we use a modified version of well-known statistical tests to compare two distributions, including the χ^2 test, the CLT-based mean test and the Wasserstein-based test. Note that categorical multimodal variables are treated as binary variables in a one-vs-all approach, which can be considered a limitation to our work. Two tests are presented here: the parity test and the equal representation test.

Parity test (one sensitive variable) To audit the bias according to a tested variable X^0 (gender, age or Fitzpatrick), it is necessary to compare the observed distribution of X^0 with its expected distribution. When X^0 is *gender*, this process

Accuracy / Precision (%)		
Gender (G)	Men (M) : 97 % →3% permutation with W	Women (W) : 95 % → 5% permutation with M
Fitzpatrick class V given G $(Fitz._V)$	V knowing predicted M : 93 % $\{\widehat{X}_{Fitz._V}\|\widehat{X}_G = M\} \leftarrow$ 7% replacement $x \sim \mathcal{U}(X_{Fitz._V}\|X_G = M, X_{pred} = False)$	V knowing predicted W : 85 % $\{\widehat{X}_{Fitz._V}\|\widehat{X}_G = W\} \leftarrow$ 15% replacement $x \sim \mathcal{U}(X_{Fitz._V}\|X_G = W, X_{pred} = False)$

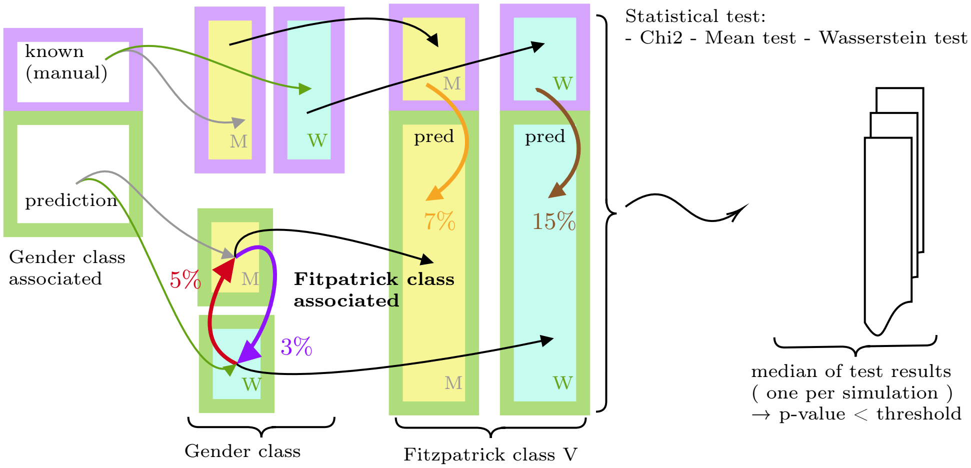


Fig. 5: Diagram explaining our error-aware testing pipeline in an equal representation test of the Fitzpatrick class V conditioned by the gender.

involves comparing X^0 observed values with a Bernoulli distribution of $p = \frac{1}{2}$ (i.e. testing $H_0 : X^0 \sim \mathcal{B}(p)$). While the assumption of a uniform distribution for the binary gender may appear reasonable, it is important to recognize the limitations of such an assumption when considering age groups or the Fitzpatrick class. To this end, a chosen parameter reflecting a real mondial distribution was utilized for comparison (see in the Appendix, in Section D.2, Table 8 and Table 9 to observe the recorded parameters). Hence, we test, respectively for when X^0 is the age ($H_0 : X^0 \sim RealDistr(Age)$) and for when X_0 is the Fitzpatrick class ($H_0 : X^0 \sim RealDistr(Fitzpatrick)$).

Equal representation (two sensitive variables) We test whether the distribution of a variable X^0 (a Fitzpatrick skin type or age interval) differs significantly given another variable X^j (gender 'men' or 'women'). Let's consider $x^0_{i'} \in 1, \cdots, K'$ the K' modalities of X^0 and $x^j_i \in 1, \cdots, K$ the K modalities of X^j. To perform this analysis, we first partition the dataset based on X^j (one-versus-all according to x^j_i) and then compare the distribution of X^0 across the two resulting partitions of X^j. We define the variable $W^0 \in \{0,1\}^{K'}$ such as $W^0 = (W^0_1, \cdots, W^0_{K'})$ and the $W^0_{i'}$ are defined as followed:

$$\forall i' \in 1, \cdots, K' \quad W^0_{i'} := \begin{cases} 1 & \text{if } X^0 = x^0_{i'} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

a condition notation of $W^0_{i'}$ on a subspace S according X^j is given by :

$$\forall i \in 1, \cdots, K \quad W^{0,j}_{i',i} := \begin{cases} 1 & \text{if } X^0 = x^0_{i'} \text{ in } S \in \{X^j = x^j_i\} \\ 0 & \text{if } X^0 \neq x^0_{i'} \text{ in } S \in \{X^j = x^j_i\} \end{cases} \tag{3}$$

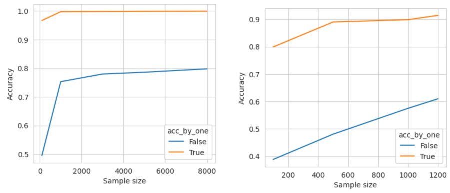
$$W^{0,j}_{i',\overline{i}} := \begin{cases} 1 & \text{if } X^0 = x^0_{i'} \text{ in } S \in \{X^j \neq x^j_i\} \\ 0 & \text{if } X^0 \neq x^0_{i'} \text{ in } S \in \{X^j \neq x^j_i\} \end{cases} \tag{4}$$

We test the following assumption on the distributions, $H_0 : W^{0,j}_{i',i} \sim W^{0,j}_{i',\overline{i}}$.

Uncertainty aware To ensure reliability, the auditing process must be robust to variations in model annotation accuracy. Consequently, the test must be robust to prediction errors and minimize false negatives for the null hypothesis, $\mathcal{H}_0$. As permutation tests, we randomly invert the automatic annotation modality of some predicted variables while keeping manual annotations unchanged. This procedure serves to reduce the discrepancy between distributions and minimize false negatives in test decisions. The model prediction is denoted by $\widehat{W^0_i}$, and the true value by W^0_i. These tests are modified according to the following methodology:

- Parity test:
  1. We calculate the model's accuracy $A^{W^0_{i'}} := \mathbb{P}[\widehat{W^0_{i'}} = W^0_{i'}]$ for each of the estimated variables $\widehat{W^0_{i'}}$.
  2. We randomly replace $100 \times (1 - A^{W^0_{i'}})\%$ of $\widehat{W^0_{i'}}$ by values simulated according to the expected parameter (for example $\mathcal{B}(\frac{1}{2})$ for gender).
- Representation test:
  1. We calculate the prediction's precision $P^{W^j_i} := \mathbb{P}[W^j_i = 1|\widehat{W^j_i} = 1]$ for each of the variables $\widehat{W^j_i}$.
  2. We randomly permute $100 \times (1 - P^{W^j_i})\%$ between $\widehat{W^j_i}$ and $\widehat{W^j_{\overline{i}}}$ values (e.g., transforming predicted women into predicted men and vice versa)
  3. We compute $A^{W^0_{i'}, W^j_i = 1} := \mathbb{P}[\widehat{W^0_{i'}} = W^0_{i'}|W^j_i = 1]$ which represents the classification model's accuracy for the modality $x^0_{i'}$ conditioned on x^j_i.
  4. We randomly replace $100 \times (1 - A^{W^0_{i'}, W^j_i = 1})\%$ of the automatically annotated variables $\widehat{W^{0,j}_{i',i}}$ and $\widehat{W^{0,j}_{i',\overline{i}}}$ respectively with manually annotated variables $W^{0,j}_{i',i}$ and $W^{0,j}_{i',\overline{i}}$.

Note that the accuracies and precisions above are calculated on the validation set. We conduct multiple statistical tests across several simulations and aggregate the results by taking the median p-value of all simulations. This framework is illustrated in Fig. 5.



(a) GAN dataset (b) CelebA dataset

Fig. 6: Learning sample size impact on the Network accuracy (Fitzpatrick classification). The by-one-accuracy includes predictions for the true class as well as the directly adjacent classes.

Hyperparameter	**GAN**	**CelebA**
Skin information (ITA..)	+0.6	+0.2
Overhead network choice	+3.7	+4.2
Skin mask	-1.7	+1.2
Removing background	+0.1	+1.8
Latent space size (2048)	+1.9	+2.7

Table 1: Impact of hyperparameters and architecture for neural network designed to classify the Fitzpatrick scale on the GAN and CelebA datasets.

6 Results

We report results on the two datasets described in Section 3.1. Our auditing process consists of: (1) assessing whether there is a significant difference in the observed proportions across three sensitive attributes—gender, age, and Fitzpatrick classification; and (2) evaluating whether significant differences exist in the observed proportions of age and Fitzpatrick classification, **conditioned on gender**.

6.1 Ablation study of the Fitzpatrick classification

As shown in Fig.6, the size of the manually annotated training dataset has a significant influence on model accuracy. Using the GAN dataset, our model achieves a 76% correct Fitzpatrick classification prediction rate on the test set, with at least 12.5% manual annotation. However, with CelebA, the accuracy of the model does not exceed 65%, despite hyperparameter tuning. It highlights the necessity of accounting for model errors in the tests to avoid the need for a full retraining of the CNN. Age and gender are provided by a trained FairFace model, so there is no training step, and the model achieved respectively a 97.17% and a 94.46% accuracy on *CelebA* and *GAN* for the gender classification. Without age labels, to verify the consistency of FairFace's age prediction, we compared its prediction with another network prediction and obtained a 72.77% classification similarity for the GAN dataset (More details in the Appendix, see Table 7). Table 1 provides an ablation study examining the impact of each hyperparameter in the CNN architecture for Fitzpatrick classification. A high-dimensional latent space, the incorporation of ITA into the model, and the removal of image backgrounds significantly enhance the learning process. However, the use of a skin mask has a negative effect on the results. This can be explained by the removal of features such as hair color, which are essential for Fitzpatrick classification. The addition of ITA showed a notable improvement for small training datasets

Table 2: Statistical test on the parity of Gender ($\mathcal{H}_0 : p = 0.5$) and the real distribution for the Fitzpatrick class and Age ($\mathcal{H}_0$: marginal and global distributions are equivalent). The test used were the Wasserstein test, the Mean test and the χ^2 test. ✓, $✓_{2/3}$ and × respectively means that 0, 1 or at least 2 tests rejected $\mathcal{H}_0$. Colored cell means that the test result are different because of the error-aware protocol: 🟩 means that, thanks to the error-aware method, less test rejected $\mathcal{H}_0$.

(a) GAN

Sensitive Variable	**Sample Size**				
	100	**500**	**1000**	**3000**	**8000**
Gender	×	×	×	×	×
Age					
0-2	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓
3-9	×	×	×	×	×
10-19	×	×	×	×	×
20-29	×	×	×	×	×
30-39	×	×	×	×	×
40-49	×	×	×	×	×
50-59	×	×	×	×	×
60-69	×	×	×	×	×
70+	×	×	×	×	×
Fitzp. class					
I	×	×	×	×	×
II	×	×	×	×	×
III- IV	×	×	×	×	×
V	×	×	×	×	🟩 ✓
VI	×	×	×	×	×

(b) CelebA

Sensitive Variable	**Sample Size**			
	100	**500**	**1000**	**1200**
Gender	×	×	×	×
Age				
0-2	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓
3-9	×	×	×	×
10-19	×	×	×	×
20-29	×	×	×	×
30-39	×	×	×	×
40-49	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓
50-59	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓
60-69	×	×	×	×
70+	×	×	×	×
Fitzp. class				
I	×	×	×	×
II	×	×	×	×
III- IV	×	×	×	×
V	×	×	×	🟩 ✓
VI	×	×	×	🟩 $✓_{2/3}$

with an embedding size of 1000 dimensions. We believe that skin-related features are sufficiently captured for the embedding of size 2048 but lost during dimensionality reduction (embedding of size 1000).

6.2 Ablation study of the error-aware method

To study the impact of the error-aware approach, we evaluated the statistical tests with and without taking into account the imprecision of the neural network's predictions. The colored cell on Table 2 and Table 3 show the effect of adding the uncertainty aware corrections to the statistical tests. For the parity test (1), for over the 135 tests, the aggregation of test accepted $\mathcal{H}_0$ 20 times with the error-aware approach, and only six times without it. For the equal representation test (2), the error-aware method affected the result of 88 of the 126 aggregations of tests: for 84 out of the previously mentioned 88, the uncertainty aware corrections made the audit result more tolerant. we provide in the Appendix the Table 2 and Table 3 without corrections (Table 10 and Table 11).

6.3 Sample size impact on statistical test results

Here, we assess whether both test methodologies produce the same conclusions as those obtained from the fully annotated dataset, which serves as the ground

Table 3: Equal representation statistical test on the for the Fitzpatrick class and the Age, with respect to each reflected closest binary gender subgroup. $\mathcal{H}_0$: The Fitzpatrick or Age distribution of the reflected men subgroup is the same as the reflected women subgroup. The tests used were the Wasserstein test, the Mean test, and the χ^2 test. $\checkmark$, $\checkmark_{2/3}$ and $\times$ respectively means that 0, 1 or at least 2 tests rejected $\mathcal{H}_0$. Colored cell means that the test result are different because of the error-aware protocol:🟩 and respectively 🟥 mean that, because of the error-aware method, more test or respectively less test rejected $\mathcal{H}_0$

(a) GAN

Sensitive Variable	Sample Size				
	100	500	1000	3000	8000
Age					
0-2	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
3-9	🟩 $\checkmark$	🟩 $\checkmark$	🟥 $\times$	🟩 $\checkmark$	🟩 $\checkmark$
10-19	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
20-29	🟩 $\checkmark_{2/3}$	🟩 $\checkmark$	$\times$	🟩 $\checkmark$	🟩 $\checkmark_{2/3}$
30-39	🟩 $\checkmark$	$\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
40-49	$\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
50-59	🟩 $\checkmark$	$\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
60-69	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
70+	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
Fitz. class					
I	🟩 $\checkmark_{2/3}$	$\checkmark_{2/3}$	$\times$	$\times$	$\times$
II	🟥 $\times$	$\times$	🟩 $\checkmark_{2/3}$	🟩 $\checkmark_{2/3}$	🟩 $\checkmark$
III- IV	$\times$	$\times$	$\times$	$\times$	🟩 $\checkmark_{2/3}$
V	$\times$	$\times$	🟩 $\checkmark$	🟥 $\times$	$\checkmark_{2/3}$
VI	$\times$	$\times$	$\times$	$\times$	$\times$

(b) CelebA

Sensitive Variable	Sample Size			
	100	500	1000	1200
Age				
0-2	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
3-9	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
10-19	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
20-29	$\times$	$\times$	$\times$	$\times$
30-39	🟩 $\checkmark$	$\checkmark_{2/3}$	🟩 $\checkmark$	🟩 $\checkmark$
40-49	$\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
50-59	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
60-69	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
70+	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$
Fitz. class				
I	🟥 $\checkmark_{2/3}$	🟩 $\checkmark_{2/3}$	$\times$	$\times$
II	$\times$	$\times$	$\times$	$\times$
III- IV	🟩 $\checkmark$	🟩 $\checkmark$	🟩 $\checkmark$	$\checkmark$
V	$\times$	$\times$	$\checkmark_{2/3}$	🟩 $\checkmark$
VI	🟩 $\checkmark$	$\times$	🟩 $\checkmark_{2/3}$	🟩 $\checkmark_{2/3}$

truth without annotation errors, given different amounts of manually annotated data. The parity test (1), for the *GAN* dataset resulted in only four false rejections out of the 75 tested hypotheses, with the error associated with Fitzpatrick category V (Table 2a). For the *CelebA* dataset, we observed two modalities with false rejection out of the 15 tested (see Table 2b). In both datasets, our parity test demonstrates robustness to the *sample size* effect. The equal representation test (2) is more sensitive to sample size effect. For the *GAN* dataset, it produced eleven false rejections out of 70 tests (Table 3a). It seems that sample size$\geq$ 1000 is enough to get stabilized results. For the *CelebA* dataset, Our equal representation test (Table 3) produced three false negative and three false positives among 56 tests.

6.4 Auditing result

Parity test results (1) For the *GAN* dataset, the parity tests of our audit (Table 2a) reveal that the observed proportions for gender, age, and Fitzpatrick scale features do not align with the proportions recorded in the general population. For the *CelebA* dataset, the population aged 40 to 59 is the only age group representative of the recorded parameter.

Equal representation test result (2) For the *GAN* dataset, our auditing reveals a strong gender-related bias with ethnicity, for example, a woman is 1.37 times more likely to be in the Fitzpatrick class I compared to a man. Contrariwise, a man is 1.41 times more likely to be in the Fitzpatrick class VI (For all values, see in the Appendix, the Table 9). No gender-related biases with age are present in the *Gan* dataset. In *CelebA*, the age group 20-29 is overrepresented among women (73% of women are in this age group against 36% for men). There also exist a strong Fitzpatrick-gender bias : while 66% of women are of Fitzpatrick class II, 52% of men are. On the contrary, men are 12.5 times more likely to be in the Fitzpatrick class I.

7 Conclusion

We have proposed a new bias auditing method that minimizes the need for manual annotation (requiring between 100 and 1000 annotations) and is robust to errors in automated annotation. When rejecting the null hypothesis, data fluctuations mean that not all statistical tests necessarily yield the same conclusions. For instance, the χ^2 test appeared more lenient compared to the Wasserstein test. To address this, we consider the majority vote of the results from our three tests. We are also aware that our method tends to accept the null hypothesis ($\mathcal{H}_0$) more readily in the equal representation test. However, we aim to avoid discouraging users from adopting our method due to an excessive number of false rejections. Our primary goal is to encourage users to utilize our tool rather than to achieve a high recall rate (sensitivity). We emphasize the importance of assessing the representativity of individuals before using an image dataset for training, in order to mitigate potential discrimination. In the context of the AI Act, which will require companies to certify the compliance of training data for machine learning models, we hope that this audit will serve as a first tool at their disposal. Our future work will aim to monitor bias in generative models in online mode.

Acknowledgments. The authors are partially supported by the AI Interdisciplinary Institute and ANR Regulia , which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-23-IACL-0002. We extend our thanks to Diallo Mohamed, Oumou Hawa Bah and Djiguinée Mamady for their efforts in the labeling process.

Disclosure of Interests. The authors have no competing interests.

References

1. Adebayo, J., Hall, M., Yu, B., Chern, B.: Quantifying and mitigating the impact of label errors on model disparity metrics (2023)
2. Alifia, R.V., Ayu, M.A.: Preprocessing with skin segmentation to improve monk skin tone (mst) classification. In: 11th International Conference on EECSI (2024)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: Limitations and opportunities (2023)
4. del Barrio, E., González-Sanz, A., Loubes, J.M.: Central limit theorems for general transportation costs (2024)
5. del Barrio, E., Gordaliza, P., Loubes, J.M.: A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. A Journal of the IMA (2019)
6. del Barrio, E., Loubes, J.M.: Central limit theorems for empirical transportation cost in general dimension. Annals of Probability (2019)
7. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996)
8. Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.M.: Confidence intervals for testing disparate impact in fair learning (2018), https://arxiv.org/abs/1807.06362
9. Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.M., Risser, L.: A survey of bias in machine learning through the prism of statistical parity (2022)
10. Bevan, P.J., Atapour-Abarghouei, A.: Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification (2022)
11. Boddeti, V.N., Sreekumar, G., Ross, A.: On the biometric capacity of generative face models (2023)
12. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: 1st Conference on FAccT (2018)
13. Celis, L.E., Deshpande, A., Kathuria, T., Vishnoi, N.K.: How to be fair and diverse? CoRR (2016)
14. Celis, L.E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., Vishnoi, N.K.: Fair and diverse dpp-based data summarization. CoRR (2018)
15. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017)
16. Chen, M., Zhang, Z., Wang, T., Backes, M., Zhang, Y.: FACE-AUDITOR: Data auditing in facial recognition systems. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 7195–7212. USENIX Association, Anaheim, CA (Aug 2023), https://www.usenix.org/conference/usenixsecurity23/presentation/chen-min
17. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data (2017)
18. Clemmensen, L.K.H., Kjærsgaard, R.D.: Data representativity for machine learning and ai systems. ArXiv (2022)
19. Defazio, A., Yang, X., Mehta, H., Mishchenko, K., Khaled, A., Cutkosky, A.: The road less scheduled (2024)
20. DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., Agarwal, D.: Evaluating fairness using permutation tests. In: 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020)
21. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact (2015)
22. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types i through vi. Arch Dermatol. (6 1988)

23. Gatis, D.: Rembg (2022), https://github.com/danielgatis/rembg
24. Generated, photos, team: Generated dataset (2024), https://generated.photos/solutions/academic-research
25. Gordaliza, P., del Barrio, E., Fabrice, G., Loubes, J.M.: Obtaining fairness using optimal transport theory. In: International conference on machine learning (2019)
26. Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., Strait, A.: Auditing work: Exploring the new york city algorithmic bias audit regime. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 1107–1120 (2024)
27. Hanna, A., Denton, E., Smart, A., Smith-Loud, J.: Towards a critical race methodology in algorithmic fairness. In: 2020 Conference on FAccT (2020)
28. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems (2016)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
30. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3 (2019)
31. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021)
32. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
33. King, D.E.: Max-margin object detection (2015)
34. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
35. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on CVPR (2020)
36. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (December 2015)
37. Liu J, Bitsue HK, Y.Z.: Skin colour: A window into human phenotypic evolution and environmental adaptation. (2024)
38. Merler, M., Ratha, N., Feris, R.S., Smith, J.R.: Diversity in faces (2019)
39. Monk, E.: The monk skin tone scale (May 2023)
40. Oneto, L., Chiappa, S.: Fairness in machine learning (2020)
41. ourworldata: Our world in data (2022), https://ourworldindata.org/
42. Parra, E., Kittles, R., Shriver, M.: Implications of correlations between skin color and genetic ancestry for biomedical research. Nature Genetics (2004)
43. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition (2020)
44. Risser, L., Sanz, A., Vincenot, Q., Loubes, J.M.: Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization (2022)
45. Robbins, H., Monro, S.: A Stochastic Approximation Method. The Annals of Mathematical Statistics (1951)
46. Schumann, C., Olanubi, G.O., Wright, A., Jr., E.M., Heldreth, C., Ricco, S.: Consensus and subjectivity of skin tone annotation for ml fairness (2024)
47. Taskesen, B., Blanchet, J., Kuhn, D., Nguyen, V.A.: A statistical test for probabilistic fairness. In: 2021 ACM Conference FAccT (2021)
48. Thong, W., Joniak, P., Xiang, A.: Beyond skin tone: A multidimensional measure of apparent skin color (2023)

49. worldbankdata: World bank data (2022), https://data.worldbank.org/
50. Wright, L., Muenster, R.M., Vecchione, B., Qu, T., Cai, P., Smith, A., Investigators, C..S., Metcalf, J., Matias, J.N., et al.: Null compliance: Nyc local law 144 and the challenges of algorithm accountability. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 1701–1713 (2024)
51. Zameshina, M., Teytaud, O., Teytaud, F., Hosu, V., Carraz, N., Najman, L., Wagner, M.: Fairness in generative modeling: do it unsupervised! In: GECCC (2022)
52. Zhang, Q., Wang, W., Zhu, S.C.: Examining cnn representations with respect to dataset bias (2017), https://arxiv.org/abs/1710.10577