Fairness is in the details: Face Dataset Auditing

Valentin Lafargue[1,2,3], Emmanuelle Claeys[4], Jean-Michel Loubes[2,3]

valentin.lafargue@math.univ-toulouse.fr
1: IMT, 2: INRIA, 3: ANITI, 4: IRIT, France

I - Problem

- Auditing: verifying compliance
- Discrimination: Gender, Age, Ethnicity using Fitzpatrick classes [1]
- Image dataset (biases & attribute-laden)
- Bias definition : distribution equivalence for:
  1. Parity test (one sensitive variable): $\mathcal{H}_0 : \mathcal{D}(\{G\}) \sim B(p)$, compared with expected distribution (i.e., gender : $p = \frac{1}{2}$)
  2. Equal representation test (two sensitive variables): $\mathcal{H}_0 : \mathcal{D}(\{\text{Fitz}|G = \text{Male}\}) \sim \mathcal{D}(\{\text{Fitz}|G = \text{Female}\})$
- Ensuring result : statistical guarantees
  1. χ^2 test
  2. CLT-based test
  3. Wasserstein-based test [3]
- Scale -> Prediction -> ML -> Error-aware robustness

II - Dataset exhibition: GAN & CelebA

Age distribution	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
Real / Expected	2	12	17	17	15	13	11	8	7
Predicted CelebA	0	0.3	3	48	26	11	7	4	0.1
Predicted GAN	0.3	4	6	61	22	4	2	0.3	0.05

Table 1: Age distribution: real estimated & predicted [3] (%)

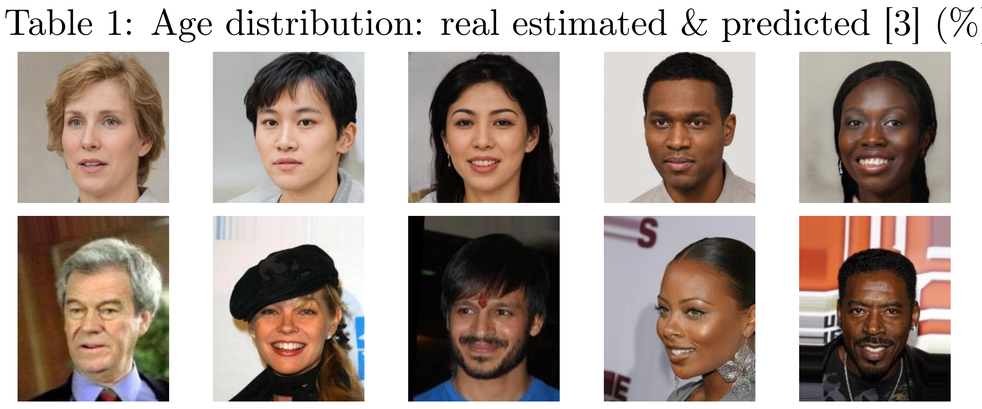


Figure 1: Fitzpatrick classification (5 class) from the left to right Phototype I, II, III-IV, V, VI. The first row is GAN dataset, the second is the CelebA dataset.

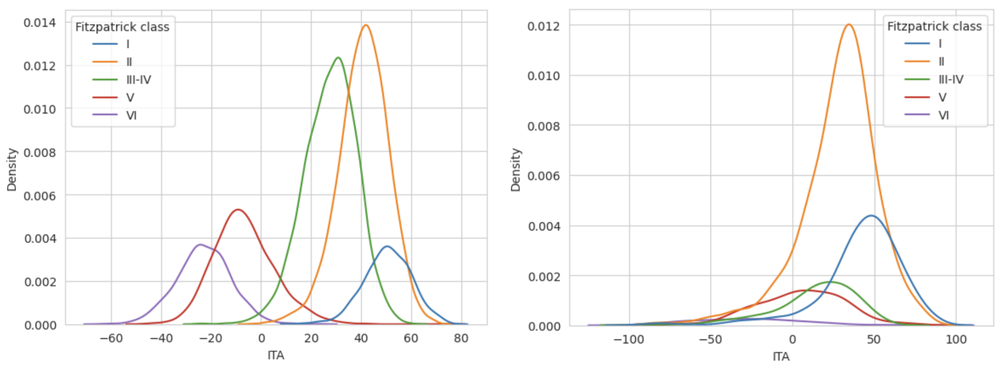


Figure 2: Probability density of ITA given the Fitzpatrick class. Gan left, CelebA right.

III - Achitecture framework: Fitzpatrick classification

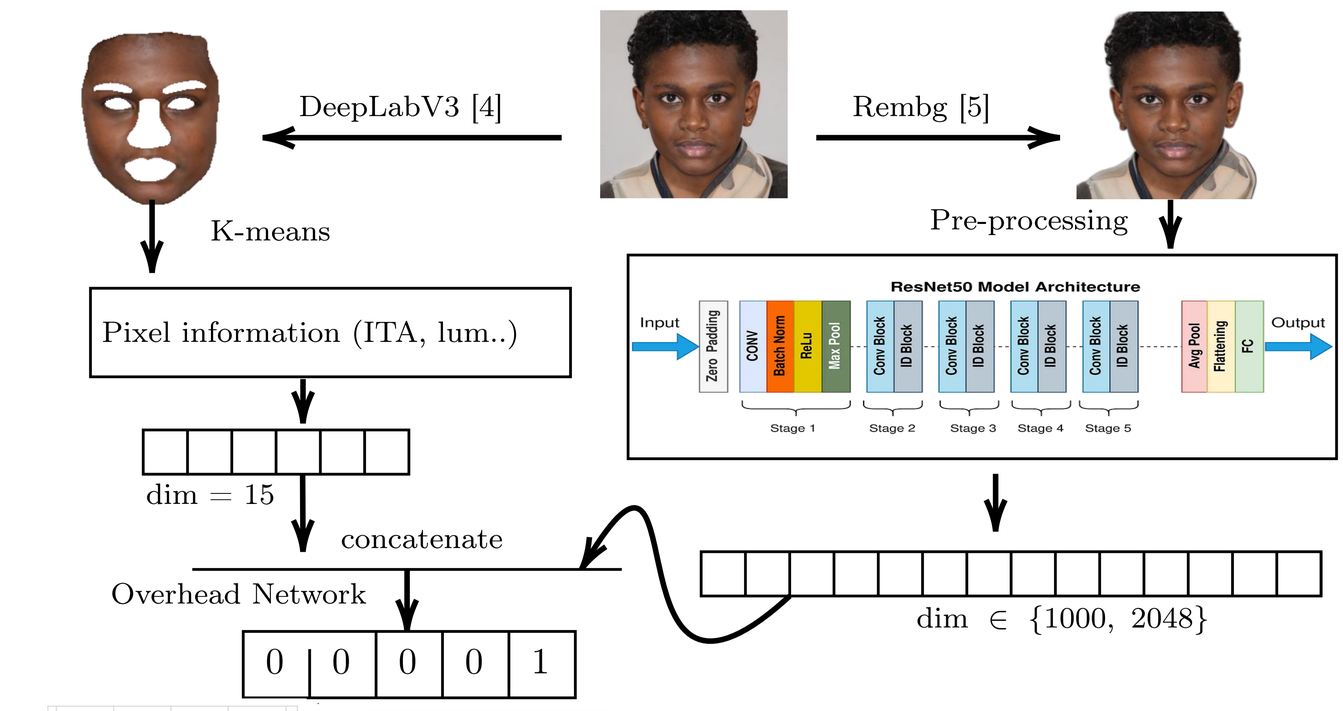


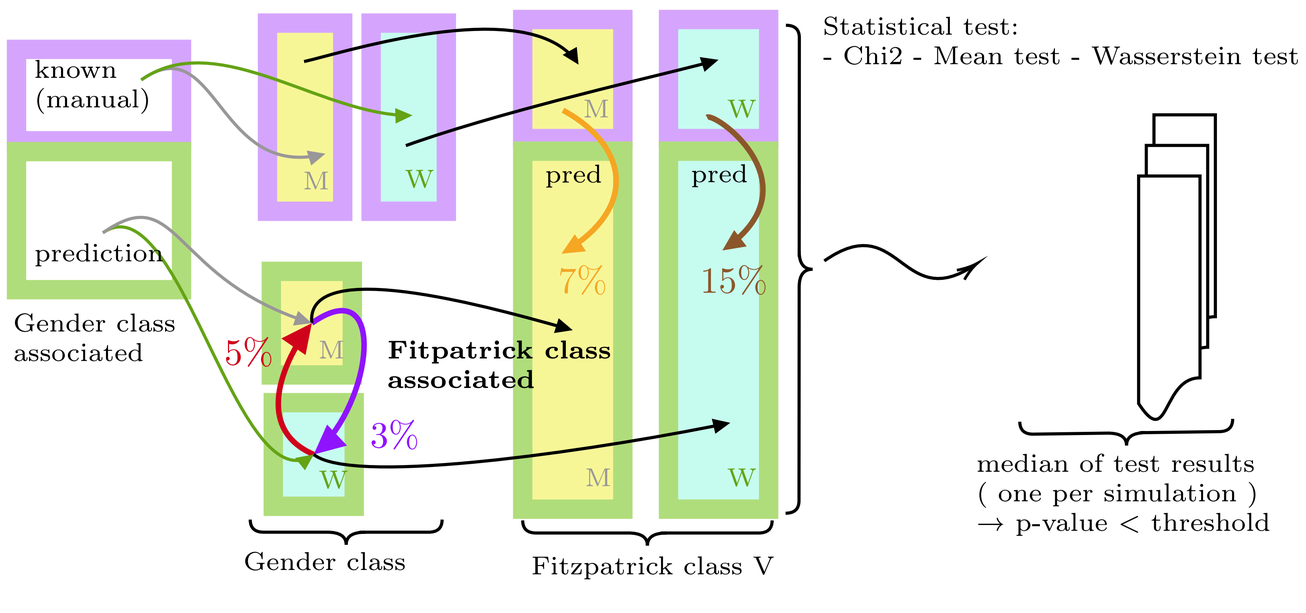
Figure 3: Learning sample size impact on the Network accuracy (Fitzpatrick classification). The by-one-accuracy includes predictions for the true class as well as the directly adjacent classes. (Left GAN, Right CelebA)

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

IV - Uncertainty aware statistical tests

	Accuracy / Precision (%)	
Gender (G)	Men (M) : 97 % $\rightarrow$3% permutation with W	Women (W) : 95 % $\rightarrow$ 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 $x \sim \mathcal{U}(X_{Fitz._V}|X_G = W, X_{pred} = False)$



V - Auditing Results

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	×	×	×	×	×

Sample Size 100	500	1000	1200
×	×	×	×
✓	✓	✓	✓
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×
✓	✓	✓	✓
✓	✓	✓	✓
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	✓
×	×	×	$\checkmark_{2/3}$

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

Sample Size 100	500	1000	1200
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
×	×	×	×
✓	$\checkmark_{2/3}$	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
$\checkmark_{2/3}$	$\checkmark_{2/3}$	×	×
×	×	×	×
✓	✓	✓	✓
×	×	$\checkmark_{2/3}$	✓
✓	×	$\checkmark_{2/3}$	$\checkmark_{2/3}$

References

[1] Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types i through vi. Arch Dermatol (1988)

[2] 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. (2019)

[3] Kärkkäinen, K., Joo, J., FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. (2021)

[4] Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017)

[5] Gatis, D.: Rembg on GitHub (2022)