Assessing Fairness in Open-Source Face Mask Detection Algorithms

or, how (not) to design and deploy your hybrid AI model responsibly

Marco Zullich

Lecturer

University of Groningen (NL)

Giovanni Santacatterina

PhD Student

University of Trieste (IT)

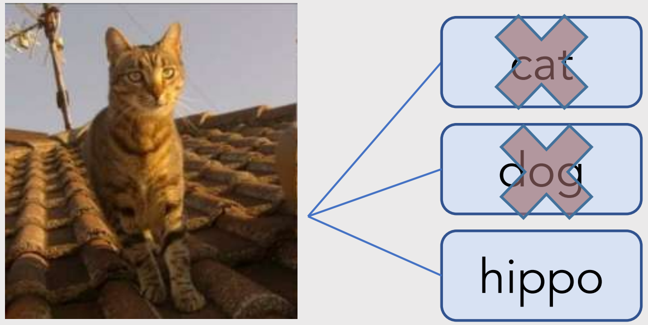
university of groningen

Computer Vision & Object Detection

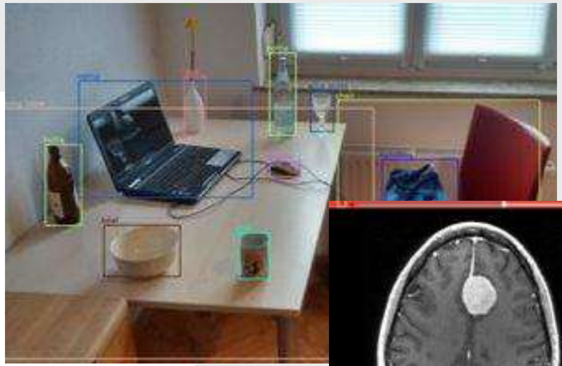
Computer vision

"Teaching" the computer to "see"

Image classification



Object detection



Instance/Semantic segmentation

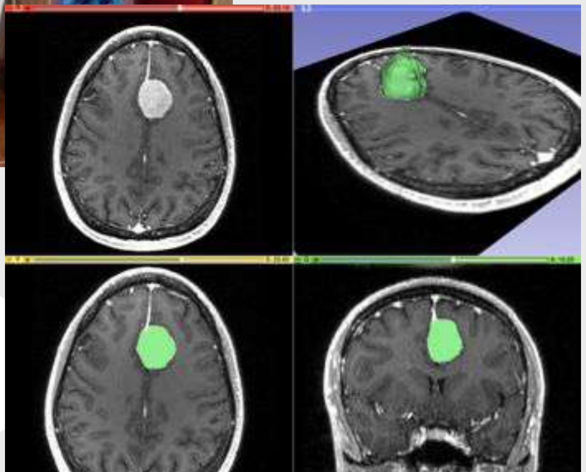
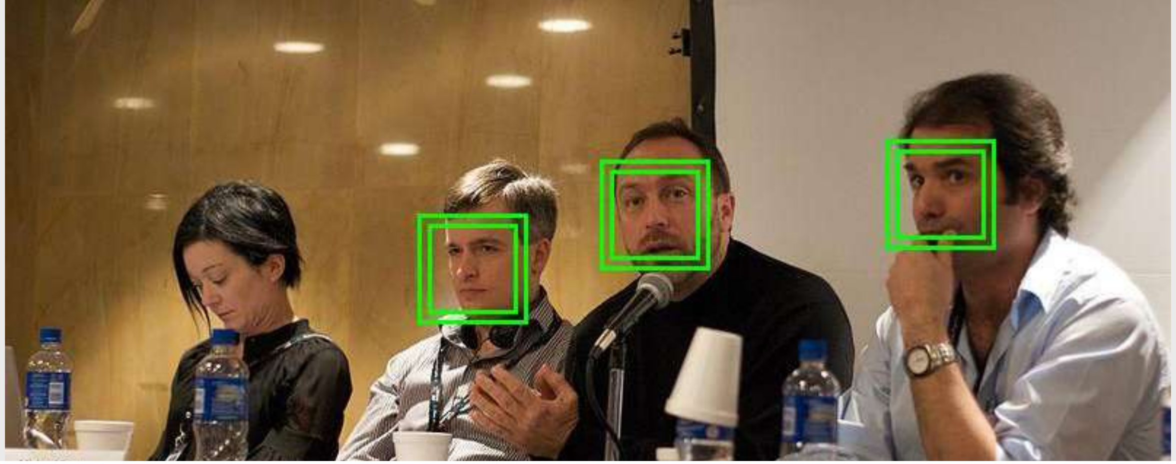
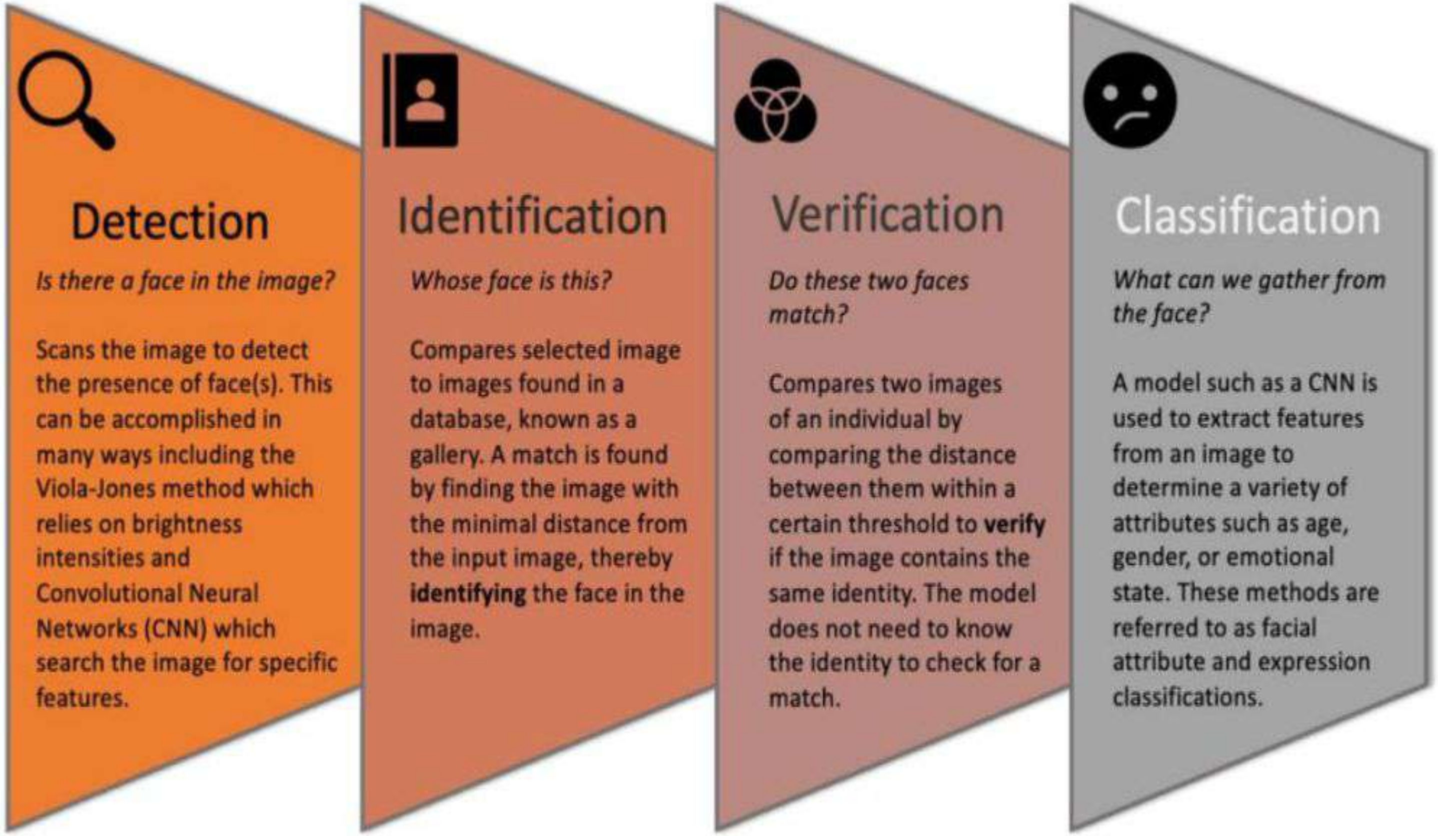


Image credits:
Detected-with-YOLO by MTheiler on Wikimedia Commons (CC-BY-SA)
MeningiomaMRISegmentation by Tatianakashunis on Wikimedia Commons (CC-BY-SA)

Face recognition and detection (FD)





Face detection coarsely localize instances of (human) faces in images

Image credits:
Face detection, by Sylenius on Wikimedia Commons (CC-BY-SA)
The Alan Turing Institute, as adapted by the European Parliament's analysis on Facial Recognition regulation in the context of the "AI Act" (2023)

Face mask detection (FMD)

Coarse localization of faces

Identify whether mask is (not) worn

Image credit: modification of one picture from the "Face-Mask-Detection" dataset, provided with MIT license.

Why FMD?

COVID19 pandemic

Face mask mandates

Need to dedicate personnel to check compliance

Can use a computer program to check compliance in a (semi)automatic way

Image by Herbfagus, CC-BY-SA 4.0 https://commons.wikimedia.org/wiki/File:Raspberry_Pi_3_Model_B.png

Our previous experience

«YOLO-based face mask detection on low-end device using pruning and quantization» [1]

Goal: produce a small model for running 24/7 on inexpensive hardware, focus not only on accuracy

«Can we assess our model beyond speed of inference and predictive accuracy?»

Authors' images used with consent of the people depicted. Pictures are not reusable.

Issues with FD (inspired from [3])

Fairness / Bias

Different predictive accuracy across different demographic variables (protected attributes)

High FPs

Accuracy is only one side of the spectrum

What is an acceptable level of False Positives?

Issues with FD (continued)

Privacy

Datasets are often constructed without the express consent of the people depicted in them

Some applications of FD are heavily restricted or banned under the new EU AI Act [2]

Bias in FDs

[5]

Face Recognition Performance: Role of Demographic Information

Brendan F. Klare, *Member, IEEE*, Mark J. Burge, *Senior Member, IEEE*, Joshua C. Klontz, Richard W. Vorder Bruegge, *Member, IEEE*, and Anil K. Jain, *Fellow, IEEE*

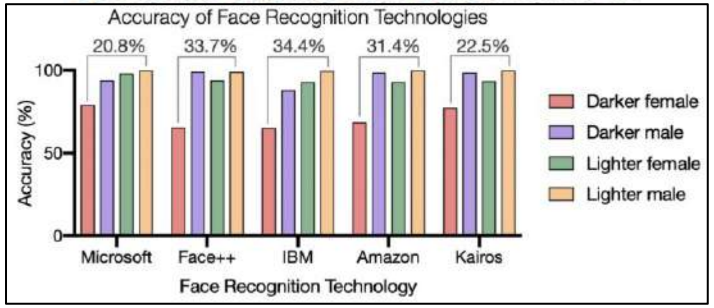
50–70 years old). Experimental results demonstrate that both commercial and the nontrainable algorithms consistently have lower matching accuracies on the same cohorts (females, Blacks, and age group 18–30) than the remaining cohorts within their demographic. Additional experiments investigate the impact of

[4]

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of



[6]

diri noir avec banan @jackyalcine · Jun 29
Google Photos, y'all [redacted] My friend's not a gorilla.
813 394

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By *Alistair Barr* Follow
Updated July 1, 2015 3:41 pm ET

[7]

TOM SIMONITE BUSINESS JAN 11, 2018 7:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

[8]

Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's.

Eight years after a controversy over Black people being mislabeled as gorillas by image analysis software — and despite big advances in computer vision — tech giants still fear repeating the mistake.

By Nico Grant and Kashmir Hill

May 22, 2023

Which are the culprits?

Hand-crafted features

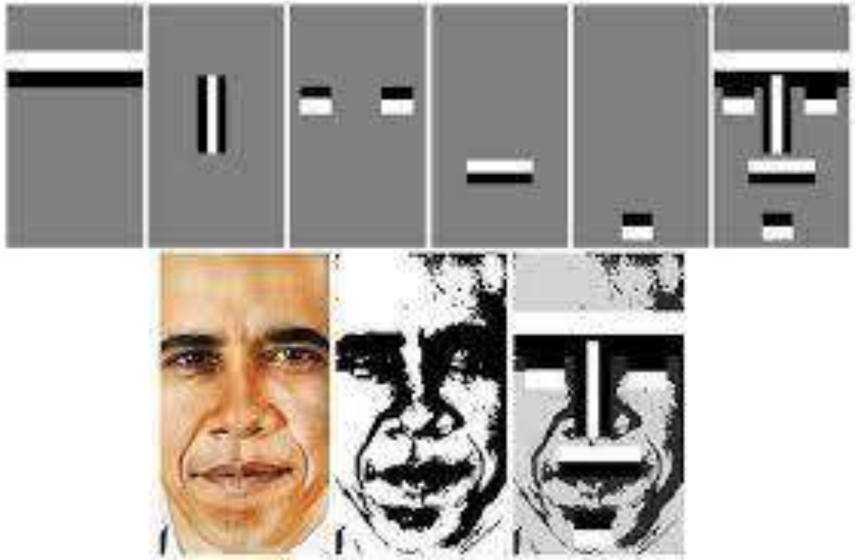


Image from Kadir, Kushsairy, et al. "A comparative study between LBP and Haar-like features for Face Detection using OpenCV." 2014 4th International conference on engineering technology and technopreneuship (ICE2T). IEEE, 2014.

Data

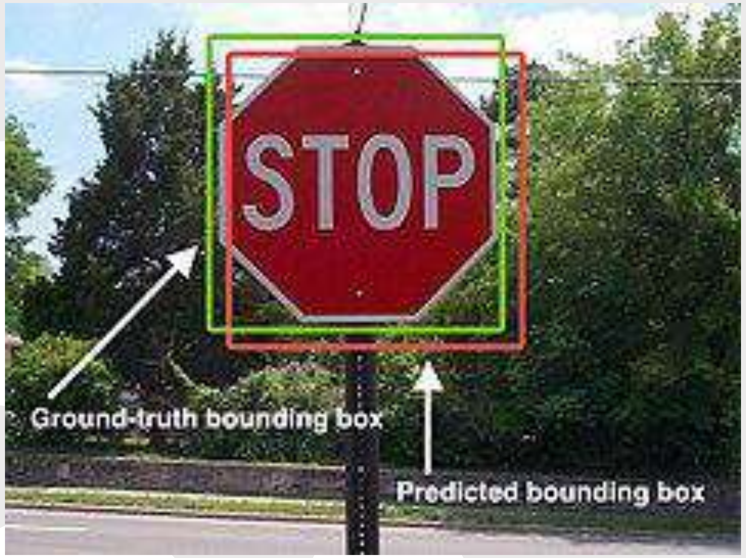


Image credits: Intersection over Union - object detection bounding boxes by Pmigdal on Wikimedia Commons (CC-BY-SA)

Does it translate to FMD?

We don't know...

[9]

Boosting Fairness for Masked Face Recognition

Jun Yu
University of Science and Technology of China
Hefei, China
harryjun@ustc.edu.cn

Xinlong Hao*
University of Science and Technology of China
Hefei, China
haoxl@mail.ustc.edu.cn

Zeyu Cui
University of Science and Technology of China
Hefei, China
mg980806@mail.ustc.edu.cn

Peng He
University of Science and Technology of China
Hefei, China
hp0618@mail.ustc.edu.cn

Tongliang Liu
Trustworthy Machine Learning Lab, The University of Sydney
Sydney, Australia
tongliang.liu@sydney.edu.au

Figure 2: Some samples of MS1M dataset. Face images without a mask (left) and with a mask (right).

[10]

Bias-Aware Face Mask Detection Dataset

Alperen Kantarcı[a,*], Ferda Ofli[b], Muhammad Imran[b], Hazım Kemal Ekenel[a]

[a] *Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey*
[b] *Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar*

6. Conclusions

We studied the problem of face mask detection during the COVID-19 pandemic with particular focus on dataset bias. Face mask detection problem has been an understudied sub-problem of face and object detection. In order to help society during the COVID-19 pandemic, many researchers across the world rapidly focused on the problem. However, majority of the earlier work has simply focused on training new architectures with the limited number of face occlusion datasets.

In this work, we introduced a novel face mask detection dataset named as Bias-Aware Face Mask Detection (BAFMD) dataset. To the best of our knowledge, it is the first face mask detection dataset that has been collected with a focus on mitigating demographic bias. Unlike most publicly available datasets, our dataset contains real-world face mask images with a more balanced distribution across different demographics, e.g., gender, race and age.

The setting

1 Survey publications introducing face mask detectors

Identify those with a publicly available implementation

2 Identify possible datasets for assessing demographic fairness

3 Carry out a statistically rigorous analysis of fairness

Survey publications introducing face mask detectors

15 claim free implementation

5 readily available

Reasons for rejection
- Parameters of models missing
- Link dead or repo empty
- Unspecified dependencies of software versions
- Requires additional data- or time-intensive setup

+1 non-published open-source face mask detector

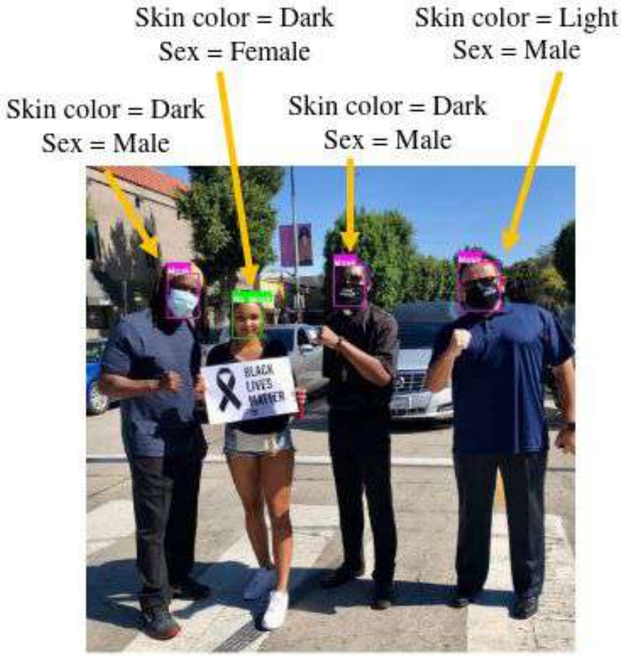
The models

Table 1. List of relevant works with publicly accessible code and model parameters which we identified and used in the present work. (*) indicates that a work is not part of a scientific publication, but it is released solely as a GitHub repository. (**) for MOXA, we make use of the YOLOv3 implementation. For additional information on the implementation details, see Section 3.1.

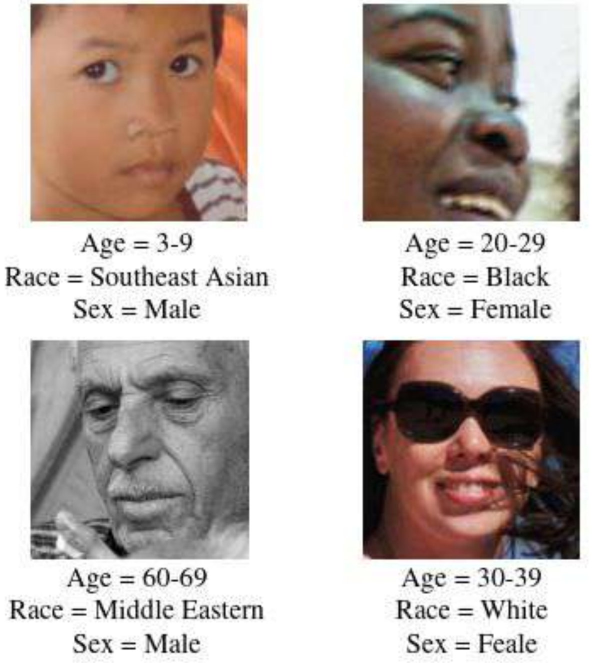
Name	Ref.	Implementation details	Language/library
Face-Mask-Detection (FMD)(*)	[21]	CNN using pre-trained face detector	TensorFlow
Maskd	[22]	CNN using pre-trained face detector	TensorFlow
Modified-Yolov4Tiny-RaspberryPi (MYTR)	[16]	YOLOv4-tiny adapted for low-end device	PyTorch + TFLite
MOXA(**)	[20]	YOLOv3, YOLOv3-tiny, SSD, Faster-RCNN	Darknet
RHF	[12]	Faster-RCNN	PyTorch
waittim-mask-detector (waittim)	[34]	custom YOLO	PyTorch

Identify possible datasets for assessing demographic fairness

Bias Aware Face Mask Detection Dataset (BAFMD)



FairFace



F2LA

Image credits: Adaptation from [10]

Adaptation from [11]

Adaptation from [12]

Fairness [13]

Fairness >> Equal treatment >> Equal probabilities

Binary support

- Target variable Y
- Prediction $\hat{Y}$
- Protected attribute A

$$P(\hat{Y}=0|A=0,Y=0)=P(\hat{Y}=0|A=1,Y=0)$$

$$P(\hat{Y}=1|A=0,Y=1)=P(\hat{Y}=1|A=1,Y=1)$$

Carry out a statistically rigorous analysis of fairness

Evaluation of object detectors

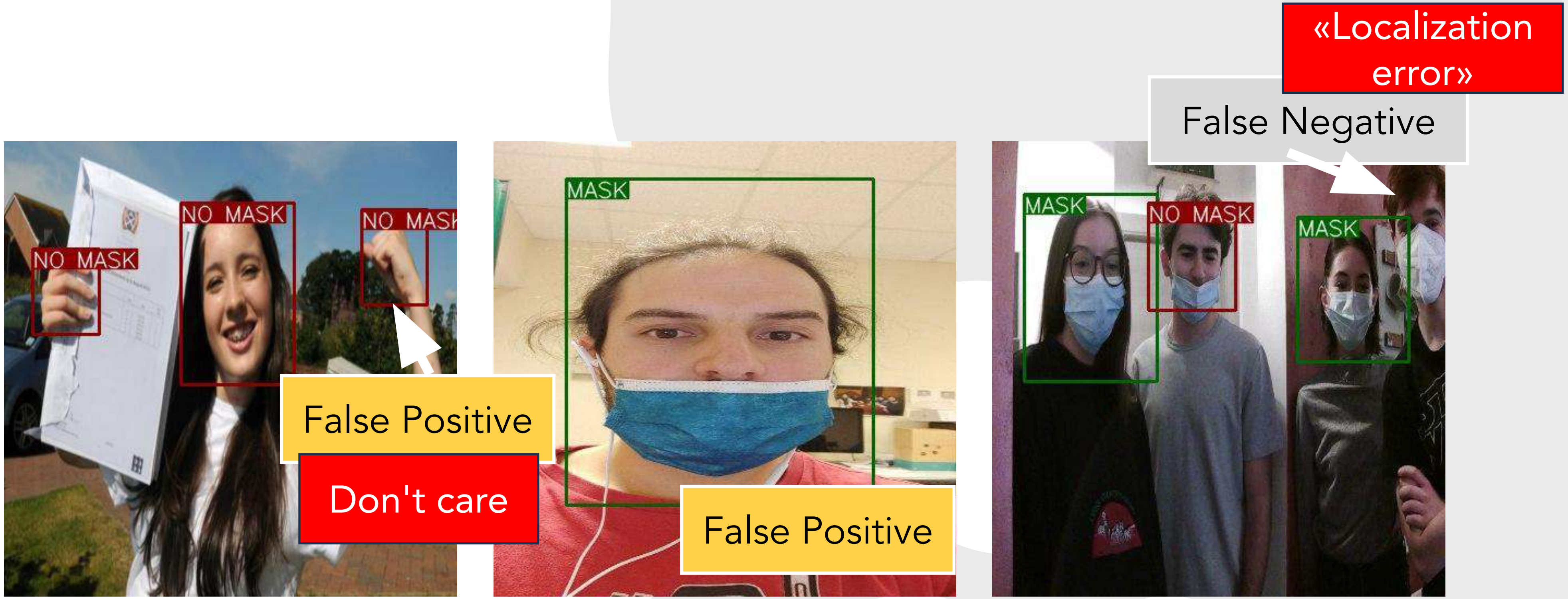
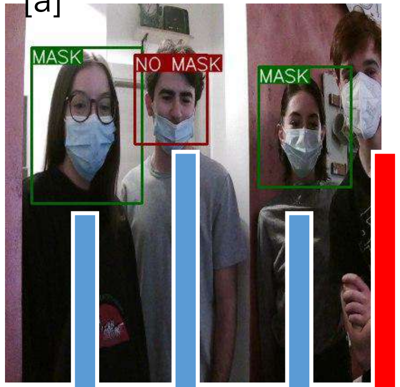


Image credit: modification of one picture from the "Face-Mask-Detection" dataset, provided with MIT license.

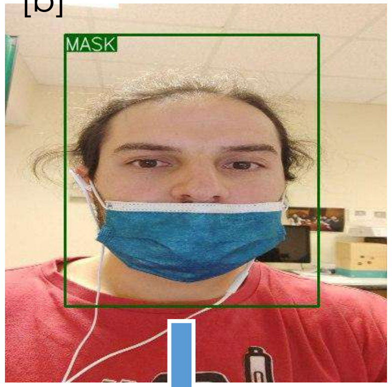
Indicators to study

Localization rate

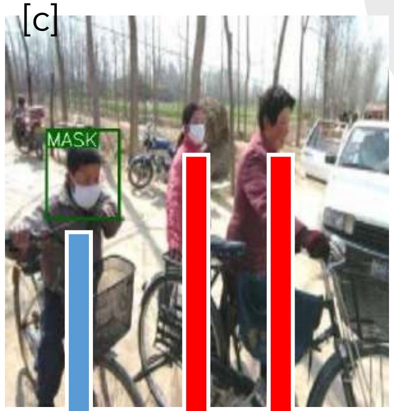
[a]



[b]



[c]



correct

wrong

True positive rate

[d]

[e]

True negative rate

[f]

[g]

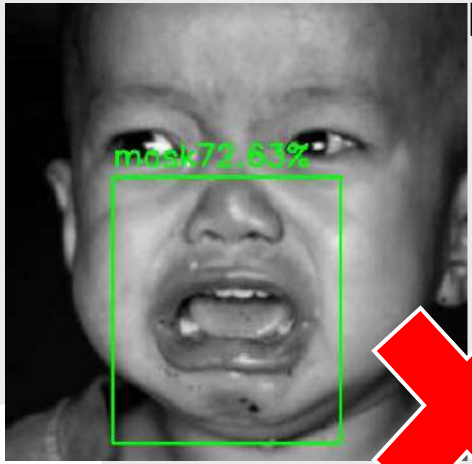


Image credits: [c], [d]: adaptations from the "Face-Mask-Detection" dataset, provided with MIT license. [e]: adaptation from [10], [g]: adaptation from [11]. Rest: own work

Statistical analysis (frequentist)

Our indicator are rates

Binomial A/B testing

$$\begin{cases} H_0 : r_1 = r_2 \\ H_1 : r_1 \neq r_2 \end{cases}$$

Difference in rates between two populations: is it significant?

$$z^\star = \frac{\hat{r}_1 - \hat{r}_2}{\sqrt{\frac{n_1\hat{r}_1 + n_2\hat{r}_2}{n_1+n_2}\left(1 - \frac{n_1\hat{r}_1 + n_2\hat{r}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Effect size

Quantifies *magnitude* of difference between groups

Cohen's h

$$h = \text{abs}(2\arcsin\sqrt{\hat{r}_1} - 2\arcsin\sqrt{\hat{r}_2})$$

h ~ 0.2	Small
h ~ 0.5	Medium
h ~ 0.8	Large

Comparison between more than 2 groups

One VS One

	A	B	C
A			
B			
C			

One VS Rest

	Population \ group
A	vs B & C
B	vs A & C
C	vs A & B

Results – FairFace - localization

Rate
Group size
p-value
Effect size

	MYTR				MOXA				RHF			
Sex	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
Female	0.9922	5162	**0.0011**	0.0629	0.9872	5162	0.1531	0.0275	0.8807	5162	**0.0000**	0.1924
Male	0.9857	5792			0.9839	5792			0.8116	5792		
Race	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
Black	0.9826	1556	**0.0133**	0.0616	0.9826	1556	0.3124	0.0266	0.7584	1556	**0.0000**	0.2561
East Asian	0.9910	1550	0.3757	0.0255	0.9903	1550	0.0857	0.0511	0.8865	1550	**0.0000**	0.1433
Indian	0.9855	1516	0.1912	0.0342	0.9875	1516	0.4869	0.0198	0.8127	1516	**0.0003**	0.0977
Latino/Hispanic	0.9975	1623	**0.0003**	0.1271	0.9889	1623	0.2113	0.0355	0.8823	1623	**0.0000**	0.1294
Middle Eastern	0.9917	1209	0.3008	0.0337	0.9793	1209	0.0575	0.0535	0.8528	1209	0.3818	0.0269
Southeast Asian	0.9866	1415	0.4003	0.0230	0.9894	1415	0.1871	0.0401	0.8848	1415	**0.0000**	0.1355
White	0.9871	2085	0.4072	0.0195	0.9808	2085	**0.0476**	0.0457	0.8374	2085	0.3445	0.0229
Age	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
0-2	0.9849	199	0.6033	0.0345	1.0000	199	0.0840	0.2438	0.9347	199	**0.0004**	0.2993
3-9	0.9904	1356	0.5399	0.0185	0.9956	1356	**0.0009**	0.1201	0.8990	1356	**0.0000**	0.1858
10-19	0.9924	1181	0.2128	0.0417	0.9865	1181	0.7685	0.0092	0.8704	1181	**0.0084**	0.0839
20-29	0.9885	3300	0.8518	0.0038	0.9867	3300	0.4971	0.0143	0.8479	3300	0.4818	0.0147
30-39	0.9854	2330	0.0825	0.0386	0.9850	2330	0.8179	0.0053	0.8283	2330	**0.0175**	0.0547
40-49	0.9882	1353	0.8239	0.0063	0.9800	1353	0.0739	0.0484	0.8012	1353	**0.0000**	0.1296
50-59	0.9912	796	0.4984	0.0264	0.9799	796	0.1713	0.0467	0.8204	796	0.0544	0.0689
60-69	0.9938	321	0.3884	0.0558	0.9688	321	**0.0114**	0.1176	0.7913	321	**0.0080**	0.1417
70+	0.9915	118	0.7753	0.0283	0.9576	118	**0.0110**	0.1757	0.8051	118	0.2393	0.1041

Three models (FMD, Maskd, waittim) not working correctly

Performance on rates seems good

Some differences are significant and effect size noticeable

Results – FairFace – True Negative Rate

	MYTR				MOXA				RHF			
Sex	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
Female	0.2087	5122	0.0906	0.0326	0.9939	5096	0.0709	0.0352	0.9996	4546	0.2757	0.0233
Male	0.2221	5709			0.9909	5699			0.9989	4701		
Race	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
Black	0.2276	1529	0.2250	0.0332	0.9908	1529	0.4783	0.0189	0.9992	1180	0.9037	0.0037
East Asian	0.1816	1536	**0.0004**	0.0992	0.9935	1535	0.5696	0.0162	1.0000	1374	0.2689	0.0596
Indian	0.2430	1494	**0.0059**	0.0753	0.9953	1497	0.1505	0.0442	0.9992	1232	0.9402	0.0023
Latino/Hispanic	0.2508	1619	**0.0002**	0.0979	0.9913	1605	0.6073	0.0135	1.0000	1432	0.2572	0.0599
Middle Eastern	0.1910	1199	**0.0270**	0.0691	0.9907	1184	0.5037	0.0198	0.9990	1031	0.7920	0.0082
Southeast Asian	0.2249	1396	0.3728	0.0254	0.9943	1400	0.3646	0.0276	0.9992	1252	0.9540	0.0017
White	0.1934	2058	**0.0061**	0.0682	0.9907	2045	0.3569	0.0218	0.9983	1746	0.1049	0.0367
Age	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
0-2	0.1633	196	0.0713	0.1366	0.9799	199	**0.0430**	0.1116	1.0000	186	0.7045	0.0556
3-9	0.1906	1343	**0.0167**	0.0712	0.9919	1350	0.8363	0.0059	0.9992	1219	0.9312	0.0026
10-19	0.2048	1172	0.3327	0.0302	0.9914	1165	0.7112	0.0112	1.0000	1028	0.3492	0.0584
20-29	0.2220	3262	0.3047	0.0214	0.9920	3256	0.8167	0.0048	0.9996	2798	0.3574	0.0232
30-39	0.2121	2296	0.6309	0.0113	0.9939	2295	0.3262	0.0241	0.9990	1930	0.6160	0.0121
40-49	0.2117	1337	0.6970	0.0114	0.9940	1326	0.4612	0.0227	0.9982	1084	0.1657	0.0364
50-59	0.2522	789	**0.0097**	0.0931	0.9923	780	0.9991	0.0000	0.9985	653	0.4555	0.0254
60-69	0.2539	319	0.0927	0.0929	0.9871	311	0.2892	0.0535	1.0000	254	0.6565	0.0558
70+	0.2991	117	**0.0275**	0.1935	1.0000	113	0.3469	0.1765	1.0000	95	0.7874	0.0553

Three models (FMD, Maskd, waittim) not working correctly

MYTR – performance is terrible

Bias seems overall better w.r.t. localization

Results - BAFMD

One model (waittim) not working correctly

Performance ranges a lot

Table 7.: Results concerning the localization rate, true positive rate, and true negative rate on the dataset BAFMD. $\hat{\pi}_i$ is the rate achieved by the model on a specific group, n_i indicates the size of the group in the dataset, while p is the p-value corresponding to the unpaired binomial test; h refers to the Cohen's h, measuring the effect size. p-values and effect sizes are shown only once per attribute since they all have binary support, and are hence the same for both groups. p-values smaller than 0.05 are shown in **boldface**—they indicate a significant difference with respect to the other groups of the same attribute. The effect size is also indicated in bold when the difference is significant and the h-number is larger than 0.2, denoting a *severe* bias (ref. Section 6). As introduced in Section 7 and Section 7.1, the models FMD, Maskd, and waittim fail to produce valid outputs on FairFace, and hence do not appear in this table.

		FMD				Maskd				MYTR				MOXA				RHF			
LOCALIZATION	Sex	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	F	0.5116	346	0.9678	0.0032	0.5751	346	0.3864	0.0657	0.1531	346	0.9549	0.0044	0.8382	346	0.9580	0.0041	0.9682	346	**0.0082**	**0.2053**
	M	0.5100	349			0.6074	349			0.1547	349			0.8367	349			0.9226	349		
	Skin Color	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	Dark	0.4600	250	**0.0447**	0.1588	0.5440	250	0.0569	0.1501	0.1560	250	0.9109	0.0089	0.8000	250	**0.0451**	0.1557	0.9320	250	0.2469	0.0896
	Light	0.5393	445			0.6180	445			0.1528	445			0.8584	445			0.9528	445		
TRUE POSITIVES	Sex	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	F	0.9618	126	0.5004	0.0843	0.9732	149	0.3780	0.1021	0.5217	23	**0.0111**	**0.7373**	0.9876	241	0.3173	0.0935	0.8111	270	0.3321	0.0855
	M	0.9763	124			0.9872	156			0.8519	27			0.9958	240			0.8434	249		
	Skin Color	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	Dark	0.9753	81	0.6921	0.0547	0.9870	97	0.4213	0.0826	0.6667	18	0.6997	0.1130	0.9938	160	0.7246	0.0355	0.7416	178	**0.0002**	**0.3317**
	Light	0.9661	177			0.9760	208			0.7188	32			0.9907	321			0.8710	341		
TRUE NEGATIVES	Sex	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	F	0.9347	46	0.8017	0.0508	0.9000	50	0.5019	0.1318	1.0000	30	1.0000	0.0000	0.8571	49	0.6776	0.0829	1.0000	65	0.0984	0.4083
	M	0.9216	51			0.8571	56			1.0000	27			0.8269	52			0.9589	73		
	Skin Color	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h	$\hat{\pi}_i$	n_i	p	h
	Dark	0.9706	34	0.2319	0.2827	0.8974	39	0.6307	0.0983	1.0000	21	1.0000	0.0000	0.8000	40	0.3540	0.1863	0.9880	55	0.3376	0.1644
	Light	0.9048	63			0.8657	67			1.0000	36			0.8689	61			0.9636	83		

Notable bias in some cases (exp. RHF for skin color)

Dataset size too small for in-depth eval

Wrapping up

Fairness in FMD algorithms for guaranteeing equal treatment in protected groups

FD algorithms have been proven to be biased exp. towards race, sex, age

Gather 6 (out of 170+ publications) open-source implementations of FMDs

Assess performance on localization and true positives/negatives rates across demographic variables on two datasets

Main results

- Bias exists but not excessive
- Performance range a lot, unacceptablee in many cases
- Too few open implementations

Limitations & Future work

Image credit: adapted from [14]

More data

Datasets specific for FMD

Datasets designed for fairness test in FD

Artificially increase data using "Mask The Face" tool

Extension of protocol

Use library «Fairness360»

Check for combination of attributes

Use Bayesian analysis instead of frequentist

References

[1] Liberatori, Benedetta, et al. "Yolo-based face mask detection on low-end devices using pruning and quantization." 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2022.
[2] Madiega, Tambiama and Mildebrath, Hendrik. “Regulating facial recognition in the EU.” EPRS | European Parliamentary Research Service, 2021
[3] Najibi, Alex. “Racial Discrimination in Face Recognition Technology.” Science in the News Harvard Blog, Special Edition on Science Policy and Social Justice. 2021
[4] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.
[5] Klare, Brendan F., et al. "Face recognition performance: Role of demographic information." IEEE Transactions on information forensics and security 7.6 (2012): 1789-1801.
[6] Barr, Alistair. “Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms.” The Wall Street Journal, 2015.
[7] Simonite, Tom. “When It Comes to Gorillas, Google Photos Remains Blind.” Wired, 2018.
[8] Grant, Nico and Hill, Kashmir. “Google’s Photo App Still Can’t Find Gorillas. And Neither Can Apple’s.” The New York Times, 2023.
[9] Yu, Jun, et al. "Boosting fairness for masked face recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[10] Kantarcı, Alperen, et al. "Bias-Aware Face Mask Detection Dataset." arXiv preprint arXiv:2211.01207 (2022).
[11] Kärkkäinen, Kimmo, and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age." arXiv preprint arXiv:1908.04913 (2019).
[12] Mittal, Surbhi, et al. "Are Face Detection Models Biased?." 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2023.
[13] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).
[14] Anwar, Aqeel, and Arijit Raychowdhury. "Masked face recognition for secure authentication." arXiv preprint arXiv:2008.11104 (2020).

Thanks for the attention!

Questions?

m.zullich@rug.nl

www.zullich.it

@marco_zul