Assessing Fairness in Open-Source Face Mask Detection Algorithms

Marco Zullich[1,*], Giovanni Santacatterina[2]

[1]Department of AI, University of Groningen, Nijenborgh 9, 9747AG, Groningen, the Netherlands
[2]Department of Mathematics and Earth Sciences, University of Trieste, via Weiss, 2, 34128 - Trieste (TS), Italy

Abstract

In response to the breakout of the CoViD-19 pandemic and the resulting face mask mandates, interest has surged in the development of face mask detection algorithms for automatic checking of the compliance with these mandates. Despite the large amount of software and publications connected to this topic, little interest has been paid to ethical facets that the deployment of these systems poses. Face detection models have been noted in the past for showing widely different performances across some demographic attributes, potentially amplifying discrimination which may already exist in certain societies. While a minority of publications raised similar concerns for face mask detection systems, no practical analyses have been carried out to investigate the fairness of these algorithms. In the present work, we aim at filling this gap. After surveying the literature on face mask detection, we uncover a small set of 6 open-source algorithms. We assess their fairness by comparing their performance across demographics such as sex, race, and age. In contrast to the aforementioned concerns, we do not uncover consistent and substantial bias over these attributes but in one model. We, though, find that some algorithms generalize very poorly to new datasets, thus raising concerns over their application to real-life scenarios. We conclude by highlighting that the small number of publicly-available implementations is concerning, as it creates a lack of transparency, which could potentially conceal from the end users issues like biases or poor generalization. The shortcomings which we found in the implementations we were able to test, further emphasize the need for more transparency in the development of these algorithms.

Keywords

Face mask detection, object detection, algorithmic fairness, bias measurement, gender bias, racial bias, fair AI, responsible AI

1. Introduction

The CoViD-19 pandemic has left a strong mark on societies all over the world. In addition to strict lockdown policies, many governments implemented social distancing and face mask mandates, rendering mandatory the use of face masks to cover mouth and nose in indoor (and sometimes also outdoor) spaces and requiring that people keep a minimum distance between each other [1]. As a result, facilities with a large attendance, such as shops, supermarkets, hospitals, *etc.*, had to dedicate staff to verify the compliance of these rules by the public. In turn,

HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26—27, 2023, Munich, Germany
*Corresponding author.
marco.zullich@gmail.com (M. Zullich); giovanni.santacatterina@phd.units.it (G. Santacatterina)
https://zullich.it/ (M. Zullich)
0000-0002-9920-9095 (M. Zullich); 0000-0001-8535-2379 (G. Santacatterina)

this has sparked the necessity for having automatic systems to relieve human staff from these tasks.

While automatic face mask detectors can greatly help public facilities in monitoring the compliance of CoViD-19 mandates, the speed with which these many researches were published calls for additional analyses concerning ethical aspects; these concerns can be, by and large, the same of face detection algorithms, which have been around as early as the '70s [2], and that, by the start of the new millennium, already had a very large number of works surveyed [3]. These concerns can be linked to algorithmic bias, whereas due to design flaws in the algorithm or in the data the algorithm is trained on, the detector commits substantially different error rates across attributes such as gender, race, and age [4]. This can be particularly concerning when these algorithms are connected to, e.g., police control systems [5]. Despite these concerns being cited also in the literature on face mask detection [6, 7], to the best of our knowledge, there does not seem to be a work testing in practice whether these bias are present also in these algorithms. A development of face mask detection systems which does not encompass possible considerations on fairness might have negative consequences: in case these systems were to be deployed without human supervision, people could be, for instance, denied entrance to shops because these models might be inaccurate on faces of a specific age group or ethnicity. In the ill-advised situation in which one of these model might be employed by police forces for enforcement of face mask mandates, people might even incur in fines for errors attributable to unfair algorithms. In the present work, we aim at assessing the fairness of face mask detectors, thus addressing the aforementioned unsupported claims of bias. Via the aid of recently published reviews, we survey the literature in search for publications presenting open-source freely-accessible implementations of face mask detection algorithms, finding 6 of them. We then make use of two publicly-accessible datasets which are designed to boast a variability in race, sex, or age, to test the performance of these algorithms across different demographics. We find that claims of unfairness are generally unsupported except for one model. We do however notice that some of these algorithms showcase very poor generalization, thus making their deployment in the wild potentially hazardous. In conclusion, we provide a consideration regarding the issue that, despite there being a very large number of works implementing face mask detection system—more than 150—published in the last few years, only a very small minority of these release an open-source and functioning implementation of their system. This greatly hampers reproducibility, a hot topic in the Artificial Intelligence community [8], and renders extremely hard—if not impossible—the assessment by independent researchers of aspects such as fairness of these algorithms.

Contributions Summarizing, the main contribution of the present work is the following: we provide a thorough and reproducible statistical analysis on fairness in open-source face mask detectors, which was not previously conducted in the literature. While previous claims of bias in face mask detectors have been made in the literature, this paper adds empirical evidence to the discussion.

The code for reproducing our analyses is available at the link https://github.com/face-mask-detection-algos/.

2. Related work

The matter of fairness in face detection algorithms is a topic which has been debated as early as 2002: Furl et al. [9] already identified differences in performance of these tools with respect to ethnicity. Klare et al. [10] showed that face recognition algorithms available in 2012 were consistently underperforming when evaluated on images of young black females. More recently, the GenderShades project [11] benchmarked a face detection algorithm on two popular datasets, concluding that the algorithm showcased substantially higher error rates for dark-skinned women than other demographic groups. In a popular media case from 2015, Google Photo's recognition algorithms erroneously classified two black-skinned men as "gorillas"[1], a problem which apparently seems yet to be fixed[2].

For what concerns the specific task of face mask detection, there exist works [6, 7] claiming that these biases are present also in face mask detection algorithms, although these claims are not backed by relevant experiments. Rather, the authors use these assertions to introduce two different datasets focused on high variability across race/ethnicity. We incorporate the one by Kantarcı et al. [7] in our analysis, while the one proposed by Yu et al. [6] seems not to be publicly available.

3. Methods for Face Mask Detection

The problem of face mask detection is a specialization of object detection (OD), a popular Computer Vision (CV) task. Given a number of categories to recognize, OD works by identifying and coarsely localizing instances of said objects in images. In the case of face mask detection, usually there are two categories: mask not worn and mask worn. In some instances [12, 13], datasets are designed using more than two categories, e.g., mask not worn, mask correctly worn, mask incorrectly worn.

The earliest approaches for OD use feature engineering for finding instances of known objects within images. This is the case, for instance, with the Viola-Jones algorithm for face detection [14], which combines the responses of multiple *weak* feature detectors, based on the intensity differences in small rectangular areas of images, to identify instances of frontal faces. This approach has also been adopted in face mask detection, as in the work by Dewantara et al. [15]. Nonetheless, these "classical" CV approaches have recently fallen into disuse in favor of more effective techniques based on DNNs, which all the algorithms used in our analysis make use of. DNN-based OD systems can be further split in two classes, one-shot and two-shot detectors [16].

One-shot detectors: these algorithm employ DNNs to perform identification and localization of objects in one shot. Thus, their output will contain information for both the categories of the objects and their localization within the image. One-shot detectors are usually fast, but tend to be less accurate than their two-shot counterparts. Examples of these techniques include You Only Look Once (YOLO) [17] and Single-Shot Detector (SSD) [18], used extensively for face mask detection (e.g., [19, 20, 21] and many others).

[1] https://www.wsj.com/articles/BL-DGB-42522, retrieved on April 13th, 2023
[2] https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html, retrieved on April 13th, 2023

Two-shot detectors: these algorithms employ an initial phase of localization, where they propose regions in which the objects are to be found; subsequently, another module performs the classification on these regions. Due to their two-phase detection, they tend to be slower than one-shot detectors, although they might enjoy a better accuracy. An example of this technology is Faster-R-CNN [22], used, for instance, in [23, 13]. For face mask detection, another two-shot approach is to use a face detector for identifying the relevant regions, then perform mask classification on these regions [24, 25].

3.1. Frameworks for Implementing Face Mask Detectors

In addition to custom implementations, there exists a large number of frameworks for implementing OD algorithms. The most common are PyTorch [26] and TensorFlow [27], two powerful open-source frameworks for Deep Learning. A lesser used framework, yet worthy of mention, is Darknet[3], written in C. For instance, the original implementation of YOLO was released in Darknet. The library Tensorflow-Lite (TFLite) [28], now part of TensorFlow, is a framework for deploying neural networks onto low-end devices. It uses C as a destination language. Other frameworks include the programming language Matlab[4], the Caffe [29] platform, and many others.

4. Background on Fairness

Fairness, when considering the act of decision-making, is defined as «absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics» [30]. Fairness may be defined in terms of *parity of treatment* between individuals, regardless of their differences in sensitive attributes. We suppose to have a model outputting a prediction $\hat{Y}$, the corresponding ground truth being identified as Y. Let A be a *protected attribute* on which we want to calculate a possible bias. Without loss of generality, let us suppose this attribute can assume only two values, 0 and 1. One possible definition of fairness (adapted from [31]) is the following:

$$P(\hat{Y} = j|A = 0, Y = k) = P(\hat{Y} = j|A = 1, Y = k), \forall j \in \text{supp}(\hat{Y}), k \in \text{supp}(Y) \quad (1)$$

This means that, fixing the ground truth, the model needs to behave similarly across the various groups of the protected attribute (even in case of misclassification). In this sense, we will be assessing for fairness by testing for statistical equality over the output of the models and the corresponding ground truths. The variables on which we will assess fairness, along with the statistical methodology employed, will be discussed in Section 6.

Biased behaviors in algorithms can arise due to flaws in the algorithms themselves or, when these algorithms are data-driven, due to existing bias in the training datasets. The latter is often the case of the data-driven Machine Learning algorithms which are experimented with in the present work. A large bulk of face masks detection datasets were created during the early days of the CoViD-19 pandemic by quickly aggregating existing resources scraped from the web,

[3] https://pjreddie.com/darknet/

[4] https://www.mathworks.com/products/matlab.html

which is a clear indication of non-random sampling, which could result in bias. Kantarcı et al. [7] claim that a large number of such datasets contain an overwhelming majority of Asian, or otherwise light-skinned people, due to the availability of images of people wearing face masks by the time of the creation of these datasets. This could be a source of unfairness, as algorithms trained on these data may fail to recognize, e.g., darker-skinned people, due to their under-representation within the dataset: a behavior which has already been noted in face detection tasks [10, 32].

5. Selection of Algorithms for the Analysis

In order to select relevant algorithms for our analysis, we decided to explore the existing literature on the topic of face mask detection and recognition. We mainly made use of three surveys [33, 34, 20] for gathering the main bulk of the researches up to the first half of 2022. For identifying works published after this period, we operated a research in a similar fashion with respect to Hu et al. [20]: we queried Google Scholar, IEEE Xplore Digital Library, Web of Science, and Springer Link with the search term (``face mask'' OR ``facemask'') AND (``detection'' OR ``recognition''). In total, we identified more than 150 publications treating the topic of automatic face mask detection from as early as 2017. As already mentioned by Liberatori et al. [19], the number of works on the topic of face mask recognition with an available implementation is very low: out of all the publications we surveyed, we were able to identify only 15 of them claiming an open-source implementation.

We identified a set of desirable characteristics that these implementations should meet in order to be ready-to-use for our analyses:

(i) A clear list of dependencies or requirements that are needed in order to run the code.

(ii) Availability of parameters for running pre-trained models without re-training phases.

(iii) Possibility of using the proposed models or methods in a plug-and-play fashion, without time-consuming set-ups, like hyperparameter fine-tuning on specific datasets.

Following this analysis, we found that 10 of the 15 works which we originally identified either (a) did not meet the aforementioned criteria, or (b) were linking to nonexistent or empty repositories. In Appendix A we detail the list of these works, specifying the motivation behind their rejection. As a consequence, only 5 works passed this initial scrutiny and were ready for use in our study. Moreover, we included an additional work [24], which is not part of a scientific publication, but is an open-source software cited in other relevant papers in the field of face mask detection (such as [35, 36]) and which has other times been employed as a benchmark for comparing performances with respect to other face mask detection algorithms. Table 1 shows the final list of implementations that we use in our analysis. As an additional note for what concerns MOXA [23]: the authors present four different architectures with different sets of weights. We made use of YOLOv3, which, according to the authors, is the model which recorded the best performance.

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	Implementation details	Language/library
Face-Mask-Detection (FMD)(*) [24]	CNN using pre-trained face detector	TensorFlow
Maskd [25]	CNN using pre-trained face detector	TensorFlow
Modified-Yolov4Tiny-RaspberryPi (MYTR) [19]	YOLOv4-tiny adapted for low-end device	PyTorch + TFLite
MOXA(**) [23]	YOLOv3, YOLOv3-tiny, SSD, Faster-RCNN	Darknet
RHF [13]	Faster-RCNN	PyTorch
waittim-mask-detector (waittim) [37]	custom YOLO	PyTorch

5.1. Datasets

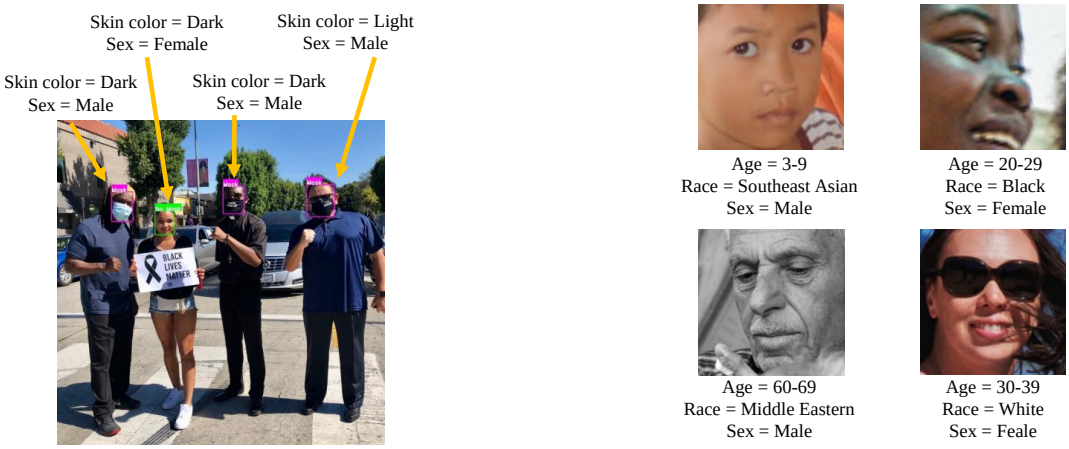
In order to evaluate the fairness of the selected algorithms, we made use of two datasets, which were recently published in an attempt to mitigate algorithmic bias in (face) mask detection systems:

- FairFace [38]: a dataset for face *classification* composed of 108 501 pictures containing one face, centered with respect to the image frame. The labels contain information on age group (0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, and 70+ years of age), gender (female, male), and race (Black, East Asian, Indian, Latino/Hispanic, Middle Eastern, Southeast Asian, White). We ran our experiments on the validation split, which contains 10 954 images. The dataset is not specific for face mask detection and contain only images of faces without a face mask. This means that we could use it only for checking true negatives and false positives in our analyses.

- Bias Aware Face Mask Detection Dataset (BAFMD) [7]: a dataset for face mask detection having more than 13 000 images containing faces with or without face masks. The labels are provided only for the presence/absence of a face mask, thus, to make it usable for our purpose, we manually annotated two attributes, skin color (dark, light) and sex (female, male), on a subset of 319 pictures (695 total faces) extracted from the validation set.

Figure 1 showcases some sample images from the datasets. In Appendix B we provide additional details on the datasets, including a summary of the numerosity for each of the variables of interest in our study.

In addition to these two datasets, we identified a third one, Fair Face Localization with Attributes (F2LA) [32]. Despite submitting a request to access it, the owners never replied to us; we were thus unable to use it in our analysis.



(a) BAFMD (b) FairFace

Figure 1: Examples of images from the datasets introduced in Section 5.1.

6. Fairness assessment

As previously talked about in Section 4, we assess fairness of the models by testing for equality of outputs given the value of ground truth and the value of the protected attribute(s). We identify two variables on which we will evaluate the fairness of the models: localization and classification.

Localization In the problem of object detection, we can assess a model over multiple factors, the first one being the *localization* of an object, regardless of the correctness of the class predicted. We can quantify the *overlap* between predicted bounding box and ground truth with the Intersection-over-Union (IoU), a metric, commonly used in OD benchmarks, for assessing the quality of localization of predicted bounding boxes [39]. Given a predicted bounding box B_{pred} and the corresponding ground truth B_{gt}:

$$\text{IoU}(B_{\text{pred}}, B_{\text{gt}}) \doteq \frac{\text{Area}(B_{\text{pred}}) \cap \text{Area}(B_{\text{gt}})}{\text{Area}(B_{\text{pred}}) \cup \text{Area}(B_{\text{gt}})} \tag{2}$$

As standard in the literature, an IoU larger than 0.5 is considered a match between ground truth and prediction, thus indicating a good localization [40]. Having characterized what constitutes a good localization, we can then define the localization rate:

$$\text{Localization rate} = \frac{\text{Faces correctly identified}}{\text{Total faces in image}}$$

For the problem of localization, we can then assess fairness by requesting the localization rate of the models to be *similar* across the support of the protected attribute(s).

Figure 2: Sample of images from the dataset FairFace: some images have more than one face per picture, thus possibly inducing the models in predicting more than one bounding box.

Classification By considering only cases of *correct localization*, we can assess the classification accuracy by checking whether a model correctly predicts the presence/absence of the face mask within the predicted bounding boxes. By recalling Equation (1), we will then check for fairness in the cases in which the model correctly predicts the presence of a face mask (*true positives*), or correctly predicts the absence of a face mask (*true negatives*). The check on false positives or negatives is redundant as the proportions are complementary with respect to the true negatives and positives, respectively.

Statistical tests In order to evaluate in a statistical fashion the significance of a difference, we will use an unpaired binomial test. Let $\hat{\pi}_i$ be the rate attained by a model (i.e., localization rate, true positive rate, or true negative rate) on a dataset over all instances having protected attribute $A = i$. We define $\hat{\pi}_{\setminus i}$ the rate attained over all the other instances in the dataset. We can see the rates as observed realizations of two binomial distributions with unobserved true rates π_i and $\pi_{\setminus i}$. We use the unpaired binomial test for the null hypothesis $H_0 : \pi_i = \pi_{\setminus i}$ with a level of significance α of 0.05. We accompany each p-value with an estimate of the effect size—namely, Cohen's h [41]—to quantify the *magnitude* of the difference between each pair of ratios. According to Cohen's guidelines, an effect size larger than 0.2 can be considered *small*. We will use this threshold for labeling significant biases as *severe*. We provide additional details on this topic in Appendix C.

Application to the Two Datasets FairFace does not have ground truth encompassing localization of the face within the image. For this dataset, thus, we have to make some assumptions and simplifications to conduct the assessments. We simplify the localization part in this way: if the model predicts *at least* one bounding box (regardless of the predicted category) then we consider the localization correct. Incorrect localizations are, then, cases where the model does not output any bounding box. Since the examples in this dataset are all negatives (i.e., people not wearing face masks), we can only check the fairness in the case of true negatives. We consider true negatives those cases in which the model predicts at least one bounding box where a mask is not worn. These simplifications are also motivated by the presence within the dataset of some images depicting more than one face (see Figure 2) which might pollute the final results.

For the other dataset, BAFMD, we are able to provide the whole picture (i.e., fairness assessment for localization, true positives, and true negatives) since the dataset has (a) information concerning the localization in the ground truth, and (b) images belonging to the *positive* class, i.e., masked faces, both of which are lacking in FairFace.

Table 2
Rates attained by the models subject of our study. "Loc." = localization rate, "TP" = true positive rate, "TN" = true negative rate. As per Section 7, waittim was excluded due to very poor performance on both datasets. Additionally, as per Section 7.1, FMD and Maskd were not able to produce a valid output on FairFace, hence no results are shown in the relevant fields. On FairFace, the true positive rates cannot be computed due to the nature of the dataset, as mentioned in Section 6.

Dataset	**FMD**			**Maskd**			**MYTR**			**MOXA**			**RHF**		
	Loc.	TP	TN	Loc.	TP	TN	Loc.	TP	TN	Loc.	TP	TN	Loc.	TP	TN
FairFace							0.9888		0.2158	0.9855		0.9923	0.8442		0.9992
BAFMD	0.5108	0.9690	0.9278	0.5914	0.9804	0.8773	0.1539	0.7000	1.0000	0.8374	0.9917	0.8416	0.9453	0.8266	0.9783

7. Results

As indicated in Section 6, we assess fairness over localization rate, true positive rate, and true negative rate, on the datasets FairFace and BAFMD. Although these metrics can be employed to measure the performance, in terms of accuracy, of the algorithms, it is important to remark that *they are not the centerpiece of our analysis*, this being more directed towards fairness.

Table 3
Results concerning the **localization rate** on the dataset FairFace. $\hat{\pi}_i$ 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 0.5 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.

	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

7.1. FairFace

Three algorithms seem completely unable to correctly identify faces in this dataset: FMD, Maskd, and waittim. In the first two cases, the models predict the presence of faces lying completely outside of the image frame, while waittim does not predict bounding boxes for almost all the images in the dataset. This leaves us with only three algorithms for this task: MYTR, MOXA, and RHF.

The results concerning the two rates for these models are presented in Table 2. All models seem to behave well (> 80%) on both rates, the only exception being MYTR, which posts an abysmal 21.58% on true negative rate, which means that it very often predicts the presence of a face mask when the subject in the picture is wearing none. The results concerning the

Table 4
Results concerning the **true negative rate** on the dataset FairFace. $\hat{\pi}_i$ 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, the model waittim fails to produce valid outputs on BAFMD, and hence does not appear in this table.

	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

analysis of fairness are instead presented in Table 3 for the localization rate and in Table 4 for the true negative rate. We notice that RHF struggles a lot with localization, as it records several significant differences across almost all demographic groups. Specifically, it also records a severe bias by apparently discriminating against black people (rate of 75.84% against an average of 85.84% for the other races—a Cohen's h of 0.2561). MYTR commits several biases in the true negative rate, although none are severe and the results are quite meaningless considering its very low performance across all demographics.

Table 5: Results concerning the localization rate, true positive rate, and true negative rate on the dataset BAFMD. π_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		

7.2. BAFMD

On BAFMD, we can provide an additional analysis of true positive rates, since this dataset encompasses also cases of people wearing face masks. Again, waittim is unable to recognize faces on this dataset. This greatly undermines its credibility, as it seems to be overtrained on its training dataset distribution, being completely unable to generalize to other situations. Now, FMD and Maskd showcase decent results, and thus we include them in this analysis.

The results concerning the three rates are presented in Table 2, while the fairness analysis is detailed in Table 5. For the localization task, performances range wildly, from the terrible 15.39% of MYTR to the 94.51% of RHF. We also have to note the poor performance of FMD and Maskd (51.05% and 59.14% respectively). On this task, though, only RHF records a severe bias, possibly discriminating men. On the analysis of positive and negative rates, all the models showcase decent performances, although again RHF commits a severe bias, possibly discriminating on dark-skinned people. This behavior was already noticed in the localization task for FairFace, thus creating a strong evidence that RHF could be consistently biased towards specific races or ethnicities. An additional notice on the false negative rates: there is no significant difference to report, although the small sample size—due to the low number of people without face masks in the dataset—does not help in getting robust results; in this sense, a more complete dataset could help in getting clearer results.

7.3. Summary

We can wrap up the results by stating that the performances we observed are quite varied, with two models showcasing good or very good results on both datasets and four performing quite poorly, thus possibly hinting at bad generalization outside of the distribution of the training data. A couple of models were notable for showcasing unfair behavior in many instances—MYTR and RHF. The latter, specifically, despite consistently showing good performance across all rates, recorded a total of four severe biases in separate occasions, with two notable cases discriminating black/dark-skinned people. All in all, apart from this case, there do not seem to be egregious cases of unfairness in the other four models, at least not with the magnitudes reported in the GenderShades project [4].

8. Conclusion and Discussion

In the present paper, we provided an analysis on the fairness of a subset of 6 open-source, ready-for-use face mask detection algorithms. Our work was prompted by the unproven claims of unfairness of face mask detection algorithms [7, 42]. Out of more than 150 papers published presenting face mask detection algorithms, we are able to single out only 6 open-source ready-to-use implementations. We identified 2 datasets for the assessment of fairness over attributes such as sex, age, and race/skin color. We assessed the fairness by testing these models on these datasets over multiple performance indicators. Our analysis seem to suggest that only one model records multiple severe cases of bias (twice on Black/dark-skinned people), that one being RHF [13], while other models, like MYTR [19], commit several of them but of smaller magnitude.

Deployment of Face Mask Detection Models in the Wild The results we obtained point out quite clearly that the deployment of these models in the wild cannot happen without extensive supplementary analyses on additional test data or on bias/fairness with respect to protected attributes. Despite being mostly fair, we show that many of the models we experimented with do not seem ready to be adopted as general-purpose face mask detectors in the wild, as they mostly do not generalize well to real-life scenarios. Three of them (FMD, Maskd, waittim) were unable to produce meaningful outputs on one or both datasets, while another (MYTR) multiple times recorded rates lower than 25%. Analyses on the shortcomings of FMD, Maskd, and MYTR are further displayed in Appendix D; for what concerns waittim, we posit that the bad performances might be due to an extreme overfitting on the domain of the training dataset (from which the test dataset was sampled). All models, though, even the best-performing ones, do showcase either weak points or severe biases, and hence, our opinion is that their deployment must be subject to human supervision to remedy their shortcomings.

Reproducibility of Results of Face Mask Detection Algorithm Another point of discussion is the unavailability of open-source code for more than 95% of the work we surveyed. Reproducibility of the results claimed in a scientific publication is fundamental to verify the reliability and transparency of these findings [8] and allows for the evaluation of aspects such as bias and fairness of the proposed models [43], aspects which might have not been considered in the original researches, and that might, thus, remain concealed from the end users of these applications. A need for transparency is further emphasized by some of the limitations demonstrated by the models tested by us, which raises the question on whether the unavailable implementations might exhibit similar issues.

Limitations and Future Work The analysis we conducted is limited due to the very low number of implementations in the field of face mask detection we were able to attain to. Moreover, our study could greatly benefit by adding more datasets, like the aforementioned F2LA [32]. Additionally, the "MaskTheFace" tool [44] could be used to increase the number of positives in some datasets by artificially drawing face masks on top of faces. For what concerns the fairness study, a future work could encompass a combination of multiple attributes, instead of considering single attributes in isolation, as we did in our analysis, to allow for a fine-grained investigation. In addition, our notion of fairness is limited to the definition given in Equation (1), which has been termed "Equal Opportunity" in a recent survey by Mehrabi et al. [30]. They also include several alternative definitions of fairness which could yield different results if applied in the context of our analysis. Nevertheless, we hope that our work helps in shedding light to the claims of bias towards race, age, or sex, of these algorithms, by showing that the situation is not as bad as other works had discovered for face detection systems [9, 10, 4].

References

[1] Y. Yan, J. Bayham, A. Richter, E. P. Fenichel, Risk compensation and face mask mandates during the covid-19 pandemic, Scientific reports 11 (2021) 3174.

[2] T. Sakai, M. Nagao, T. Kanade, Computer analysis and classification of photographs of human faces, Kyoto University, 1972.
[3] E. Hjelmås, B. K. Low, Face detection: A survey, Computer vision and image understanding 83 (2001) 236–274.
[4] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.
[5] P. Fussey, D. Murray, Independent report on the london metropolitan police service's trial of live facial recognition technology (2019).
[6] J. Yu, X. Hao, Z. Cui, P. He, T. Liu, Boosting fairness for masked face recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1531–1540.
[7] A. Kantarcı, F. Ofli, M. Imran, H. K. Ekenel, Bias-aware face mask detection dataset, arXiv preprint arXiv:2211.01207 (2022).
[8] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program), The Journal of Machine Learning Research 22 (2021) 7459–7478.
[9] N. Furl, P. Phillips, A. J. O'Toole, Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, Cognitive Science 26 (2002) 797–815. doi:https://doi.org/10.1016/S0364-0213(02)00084-8.
[10] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, A. K. Jain, Face recognition performance: Role of demographic information, IEEE Transactions on information forensics and security 7 (2012) 1789–1801.
[11] B. S. Bayu Dewantara, D. Twinda Rhamadhaningrum, Detecting Multi-Pose Masked Face Using Adaptive Boosting and Cascade Classifier, in: 2020 International Electronics Symposium (IES), 2020, pp. 436–441. doi:10.1109/IES50839.2020.9231934.
[12] A. Kumar, A. Kalia, K. Verma, A. Sharma, M. Kaushal, Scaling up face masks detection with YOLO on a novel dataset, Optik 239 (2021) 166744. doi:10.1016/j.ijleo.2021.166744.
[13] S. Wang, X. Wang, X. Guo, Advanced face mask detection model using hybrid dilation convolution based method, Journal of Software Engineering and Applications 16 (2023) 1–19.
[14] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, volume 1, Ieee, 2001, pp. I–I.
[15] B. S. B. Dewantara, D. T. Rhamadhaningrum, Detecting multi-pose masked face using adaptive boosting and cascade classifier, in: 2020 International Electronics Symposium (IES), IEEE, 2020, pp. 436–441.
[16] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, Z. Wang, A review of research on object detection based on deep learning, in: Journal of Physics: Conference Series, volume 1684, IOP Publishing, 2020, p. 012028.
[17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.

[19] B. Liberatori, C. A. Mami, G. Santacatterina, M. Zullich, F. A. Pellegrino, Yolo-based face mask detection on low-end devices using pruning and quantization, in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, 2022, pp. 900–905.

[20] Y. Hu, Y. Xu, H. Zhuang, Z. Weng, Z. Lin, Machine Learning Techniques and Systems for Mask-Face Detection—Survey and a New OOD-Mask Approach, Applied Sciences 12 (2022) 9171. doi:10.3390/app12189171.

[21] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, J. Hemanth, Ssdmnv2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2, Sustainable cities and society 66 (2021) 102692.

[22] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[23] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, T. Das, Moxa: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks, Transactions of the Indian National Academy of Engineering 5 (2020) 509–518.

[24] C. Deb, Face Mask Detection, 2022. URL: https://github.com/chandrikadeb7/Face-Mask-Detection.

[25] H. Goyal, K. Sidana, C. Singh, A. Jain, S. Jindal, A real time face mask detection system using convolutional neural network, Multimedia Tools and Applications 81 (2022) 14999–15015.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: Osdi, volume 16, Savannah, GA, USA, 2016, pp. 265–283.

[28] P. Warden, D. Situnayake, Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers, O'Reilly Media, 2019.

[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675–678.

[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[31] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[32] S. Mittal, K. Thakral, P. Majumdar, M. Vatsa, R. Singh, Are face detection models biased?, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–7.

[33] E. Mbunge, S. Simelane, S. G. Fashoto, B. Akinnuwesi, A. S. Metfula, Application of deep learning and machine learning models to detect COVID-19 face masks - A review, Sustainable Operations and Computers 2 (2021) 235–245. doi:10.1016/j.susoc.2021.08.

001.

[34] A. Nowrin, S. Afroz, M. S. Rahman, I. Mahmud, Y.-Z. Cho, Comprehensive Review on Facemask Detection Techniques in the Context of Covid-19, IEEE Access 9 (2021) 106839–106864. doi:10.1109/ACCESS.2021.3100070.

[35] S. K. Dey, A. Howlader, C. Deb, Mobilenet mask: a multi-phase face mask detection model to prevent person-to-person transmission of sars-cov-2, in: Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, Springer, 2020, pp. 603–613.

[36] I. B. Venkateswarlu, J. Kakarla, S. Prakash, Face mask detection using mobilenet and global pooling block, in: 2020 IEEE 4th Conference on Information & Communication Technology (CICT), 2020, pp. 1–5. doi:10.1109/CICT51604.2020.9312083.

[37] Z. Wang, P. Wang, P. C. Louis, L. E. Wheless, Y. Huo, Wearmask: Fast in-browser face mask detection with serverless edge computing for covid-19, arXiv preprint arXiv:2101.00784 (2021).

[38] K. Karkkainen, J. Joo, 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, pp. 1548–1558.

[39] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[41] J. Cohen, Statistical power analysis for the behavioral sciences, Academic press, 2013.

[42] J. Yu, W. Zhang, Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4, Sensors 21 (2021) 3263. doi:10.3390/s21093263.

[43] A. Lucic, M. Bleeker, S. Jullien, S. Bhargav, M. de Rijke, Reproducibility as a mechanism for teaching fairness, accountability, confidentiality, and transparency in artificial intelligence, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12792–12800.

[44] A. Anwar, A. Raychowdhury, Masked face recognition for secure authentication, arXiv preprint arXiv:2008.11104 (2020).

[45] N. Fasfous, M.-R. Vemparala, A. Frickenstein, L. Frickenstein, M. Badawy, W. Stechele, Binarycop: Binary neural network-based covid-19 face-mask wear and positioning predictor on edge devices, in: 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2021, pp. 108–115. doi:10.1109/IPDPSW52791.2021.00024.

[46] P. Bhalla, S. S. Kundu, S. Deepanjali, G. Vadivu, S. Utomo, Automatic face mask detection using a hide and seek algorithm, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers, Springer, 2022, pp. 430–439.

[47] F. Boutros, N. Damer, F. Kirchbuchner, A. Kuijper, Self-restrained triplet loss for accurate masked face recognition, Pattern Recognition 124 (2022) 108473. URL: https://www.sciencedirect.com/science/article/pii/S003132032100649X. doi:https://doi.org/

10.1016/j.patcog.2021.108473.
[48] I. Javed, M. A. Butt, S. Khalid, T. Shehryar, R. Amin, A. M. Syed, M. Sadiq, Face mask detection and social distance monitoring system for covid-19 pandemic, Multimedia Tools and Applications 82 (2023) 14135–14152.
[49] B. A. Kumar, M. Bansal, Face mask detection on photo and real-time video images using caffe-mobilenetv2 transfer learning, Applied Sciences 13 (2023) 935.
[50] R. K. Shinde, M. S. Alam, S. G. Park, S. M. Park, N. Kim, Intelligent iot (iiot) device to identifying suspected covid-19 infections using sensor fusion algorithm and real-time mask detection based on the enhanced mobilenetv2 model, in: Healthcare, volume 10, MDPI, 2022, p. 454.
[51] N. Ottakath, O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Mohamed, T. Khattab, K. Abualsaud, Vidmask dataset for face mask detection with social distance measurement, Displays 73 (2022) 102235.
[52] A. S. Joshi, S. S. Joshi, G. Kanahasabai, R. Kapil, S. Gupta, Deep learning framework to detect face masks from video footage, in: 2020 12th international conference on computational intelligence and communication networks (CICN), IEEE, 2020, pp. 435–440.
[53] B. Qin, D. Li, Identifying facemask-wearing condition using image super-resolution with classification network to prevent covid-19, Sensors 20 (2020) 5236.
[54] P. Grother, M. Ngan, K. Hanaoka, Face recognition vendor test (fvrt): Part 3, demographic effects, National Institute of Standards and Technology Gaithersburg, MD, 2019.
[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[56] M. Liu, M. Zhu, Mobile video object detection with temporally-aware feature maps, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5686–5695.
[57] S. Hooker, A. Courville, G. Clark, Y. Dauphin, A. Frome, What do compressed deep neural networks forget?, arXiv preprint arXiv:1911.05248 (2019).
[58] M. Paganini, Prune responsibly, arXiv preprint arXiv:2009.09936 (2020).
[59] S. Hooker, N. Moorosi, G. Clark, S. Bengio, E. Denton, Characterising bias in compressed models, arXiv preprint arXiv:2010.03058 (2020).
[60] C. Tran, F. Fioretto, J.-E. Kim, R. Naidu, Pruning has a disparate impact on model accuracy, Advances in Neural Information Processing Systems 35 (2022) 17652–17664.
[61] V. Joseph, S. A. Siddiqui, A. Bhaskara, G. Gopalakrishnan, S. Muralidharan, M. Garland, S. Ahmed, A. Dengel, Going beyond classification accuracy metrics in model compression, arXiv preprint arXiv:2012.01604 (2020).
[62] C. Blakeney, N. Huish, Y. Yan, Z. Zong, Simon says: Evaluating and mitigating bias in pruned neural networks with knowledge distillation, arXiv preprint arXiv:2106.07849 (2021).

A. Implementations Discarded from Our Analysis

In Table 6 we present works claiming an open-source implementation that we were unable to use in our analysis. The causes for this were either an unreachable or empty repository or a non-compliance with the criteria which we identified in Section 5.

Table 6
List of works claiming accessible implementation of face mask detection algorithms, but with issues that prevented us from using them in our analysis.

Ref.	Issue(s)
[45]	Does not meet criterion (iii) (requires additional setup)
[46]	Does not meet criterion (ii) (parameters not provided)
[47]	Does not meet criterion (iii) (requires tuning on specific datasets before usage)
[48]	Repository linked in paper is empty
[49]	Claim code is accessible via contact, but did not reply to email
[50]	Repository linked in paper leads to dead page
[51]	Does not meet criterion (ii) (parameters not provided)
[52]	Does not meet criterion (i) (dependencies not specified)
[35]	Repository linked in paper leads to dead page
[53]	Does not meet criterion (ii) (parameters not provided)

B. Additional Information on the Datasets Used

In this appendix, we provide additional details on the datasets used in our analysis, FairFace [38] and BAFMD [7]. Table 7 summarizes relevant information on these two datasets, while Figure 3 shows the composition of the datasets with respect to the labels (i.e., mask present or absent) and the additional attributes on which we analyze the fairness. Notice that FairFace only contains picture of people without face masks, being a face classification dataset.

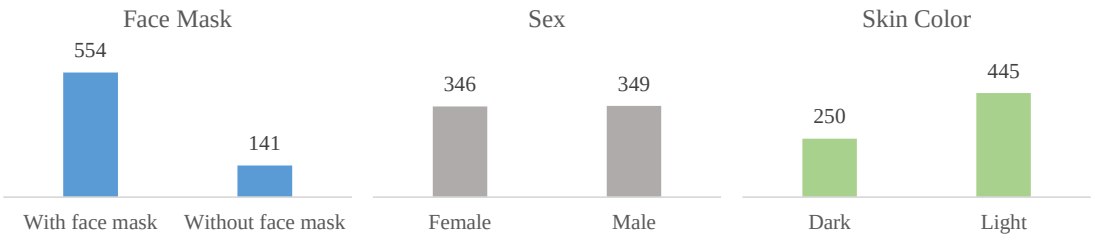
C. Details on Statistical Testing

As mentioned in Section 6, we assess the significance of the difference in ratios between two groups using an unpaired binomial test. The observed ratios, which we denote as $\hat{\pi}_1$ and $\hat{\pi}_2$, can be seen as realizations of two Binomial random variables with number of trials n_1 and n_2 and unobserved success probabilities π_1 and π_2. The numbers of trials coincide with the sizes of the two groups. We can test for the difference between these true unobserved ratios of the two populations using the unpaired binomial test with the following set of hypotheses:

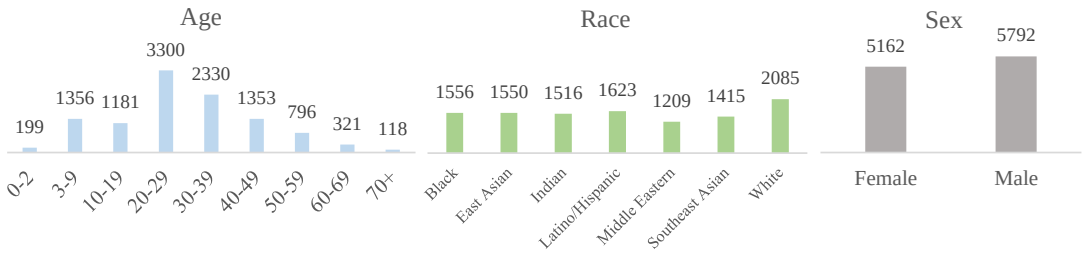
$$\begin{cases} H_0 : \pi_1 = \pi_2 \\ H_1 : \pi_1 \neq \pi_2 \end{cases}$$

Table 7
Summary of the datasets used for assessing the fairness of the identified face mask detection models and introduced in Section 5.1. "Num. labels" refers to the number of faces labelled with a bounding box within the dataset. "N/A" = "not applicable".

Dataset name	Task	Num. images	Image size	Num. labels	Attributes (Support)
BAFMD	Face mask detection	318	variable	695	Sex (Female, Male) Skin color (Dark, Light)
FairFace	Face classification	10 954	224 × 224	N/A	Age (0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+) Race (Black, East Asian, Latino/Hispanic, Southeast Asian, White) Sex (Female, Male)



(a) BAFMD



(b) FairFace

Figure 3: Composition of the two datasets, BAFMD 3a and FairFace 3b, with respect to the attributes subject of our analysis, which we introduced in Section 5.1.

Let $\hat{\pi} \doteq \frac{n_1\hat{\pi}_1+n_2\hat{\pi}_2}{n_1+n_2}$. The test quantity is defined as:

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

The corresponding p-value is computed as $P(|Z| > z^\star)$, Z being a Gaussian random variable with mean 0 and variance 1. We accompany the p-value with an evaluation on the effect size using Cohen's h [41]. The effect size can complement a statistical test by quantifying the *magnitude* of a difference between two populations' aggregates. Cohen's h was designed to introduce a notion of *dissimilarity* between two ratios or proportions. It is calculated as

$$h = \left|2 \cdot \arcsin\sqrt{\hat{\pi}_1} - 2 \cdot \arcsin\sqrt{\hat{\pi}_2}\right|.$$

Cohen introduced a rule of thumb for the interpretation of h, indicating cutoffs at 0.2, 0.5 and 0.8 as reference values for denoting the difference as *small*, *medium*, *large*.

D. Additional Insights on Models and Results

Hereby we offer supplementary insights on three of the six models we made use of.

FMD and Maskd: similarities and invalid outputs on FairFace FMD and Maskd are quite similar in concept: they both make use of a two-stage detector composed of (i) a face detector for identifying a region of interest, and (ii) a mask classifier which acts on one specific region of interest. The usage of a face detector for recognizing masked faces does not seem a good strategy, as indicated by Groher et al. [54]. They noted that these algorithms, despite showing good performance at generically recognizing faces, often failed (around 50% higher error rate) when evaluating images of masked faces. This could motivate the very subpar performance on localization attained on BAFMD of these two models. The connections do not end here: there are obvious similarities in the implementations. In both cases the face detector is the same pre-trained ResNet-10 [55]; the face mask classifier, on the other hand, is a custom Convolutional Neural Network in Maskd and a MobileNetV2 [56] in FMD. Both the works use TensorFlow for training and OpenCV for deployment; in addition, some code looks extremely similar, included the readme file in the GitHub repositories. Given the fact that Face-Mask-Detector is released in an older repository than Maskd, and the latter does not cite the former in any form, we have notified the authors of Face-Mask-Detector on the matter, citing a potential case of plagiarism. The similarities are not limited to the code and architecture; both models showcase the exact same behavior on the dataset FairFace, whereas they consistently output bounding boxes which lie *completely* outside the image frames. It is unclear to us what is causing this pathological behavior, although we assume that the problem is caused by the ResNet-10 composing the first stage of the detection. We do not know whether the issue lies in the architecture itself or in the pre-processing which is applied on the data before being fed into the model. We did however assume that one problem could be the small size of the images of the dataset (224×224). We tried upscaling the images by a factor of 2 and feed them into the models, but the results did not change. We operated no further analysis on the malfunctioning on the two implementations.

MYTR: poor localization on BAFMD and fairness concerns Continuing with another underperforming model, MYTR, we have a motivation for the very poor performance on the localization rate on BAFMD (around 15%). We did expect to record lower results across all the three rates with respect to the other models, since MYTR was heavily pruned and quantized to

Figure 4: Comparison between the bounding box predicted by MYTR (in red) and the corresponding ground truth (in purple). This case is considered a missed localization as the area of the red box is more than double than the one of the purple box, causing the Intersection-over-Union to be smaller than 0.5, thus determining the miss.

be run on low-end devices; however, the outcomes were quite underwhelming. By analyzing the output produced by the model, we realized that the bounding boxes produced by it were much larger than the ones in the ground truth. This deflated the localization rate, as in many cases the Intersection-over-Union between prediction and ground truth was lower than the recognition threshold of 0.5. We can see an example of this in behavior in Figure 4. The reason for this difference in size of bounding boxes can be attributable to (at least) two aspects: (a) large differences in the labeling process for BAFMD and the training dataset of MYTR, or (b) inaccurate set of *anchor boxes*, which are the system used by YOLO (up to version 4) for outputting a fixed dimension of bounding boxes. This issue is not present in FairFace, as there we are missing bounding boxes for the ground truth, thus we used other proxies for determining a good localization (as indicated in Section 6). To check for possible improvements, we tried experimentally lowering the threshold to 0.25 to observe possible changes in the localization rate of MYTR. We did indeed observe an increase in the rate (to around 50%, still a poor result). Despite the better rate, though, we did notice important hints of possible bias in localization, with the rate for light-skinned people at around 55% and for dark-skinned people at around 45% (an effect size of 20.03). This further reinforces our findings that MYTR seems generally to do a poor job in both localization and classification by adding additional fairness concerns. The presence of such biases, in addition to those already mentioned in Section 7.2, are probably to be expected since MYTR is a dataset which has undergone pruning and quantization, both of which have been observed to increase bias towards minority groups [57, 58, 59, 60]. There are several works introducing bias-aware model compression techniques (e.g., [61, 62]) which could be employed to mitigate the biases on this model.