UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] On the Reproducibility of "FairCal: Fairness Calibration for Face Verification"

Marga Don[1], Satchit Chatterji[1], Milena Kapralova[1], and Ryan Amaudruz[1]

[1]University of Amsterdam, Amsterdam, The Netherlands

Edited by
(Editor)

Received
–

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility — This paper aims to reproduce the study *FairCal: Fairness Calibration for Face Verification* by Salvador *et al.*[1], focused on verifying three main claims: FairCal (introduced by the authors) achieves state-of-the-art (i) global accuracy, (ii) fairness-calibrated probabilities and (iii) equality in false positive rates across sensitive attributes (i.e. predictive equality). The sensitive attribute taken into account is ethnicity.

Methodology — Salvador *et al.* provide partial code via a GitHub repository [2]. Additional code to generate image embeddings from three pretrained neural network models were based on [3] and [4]. All code was refactored to fit our needs, keeping extendability and readability in mind. Two datasets were used, namely, *Balanced Faces in the Wild* (BFW) [5] and *Racial Faces in the Wild* (RFW) [6]. Additional experiments using Gaussian mixture models instead of K-means clustering for FairCal validate the use of unsupervised clustering methods. The code was run on an AMD Ryzen 7 2700X CPU and NVIDIA GeForce GTX1080Ti GPU with a total runtime of around 3 hours for all experiments.

Results — In most cases, we were able to reproduce results from the original paper to within 1 standard deviation, and observe similar trends. However, due to missing information about image pre-processing, we were unable to reproduce the results exactly.

What was easy — The original paper is clear and understandable. Furthermore, the authors provided a mostly working version of the code. Though the datasets are not freely available to the public, their authors supplied these to us swiftly after contacting them.

What was difficult — While most of the code worked with slight changes, it was assumed there were files containing image embeddings available for both datasets, which the authors neither provided nor gave details about. We therefore pre-processed and generated embeddings independent of the authors, which makes it more difficult to judge the overall reproducibility of their method. Additionally, we encountered difficulties while improving the efficiency and extendability of the code.

Copyright © 2023 M. Don et al., released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to Marga Don (margajdon@gmail.com)
The authors have declared that no competing interests exist.
Code is available at https://github.com/margajdon/reproduction-FAIRCAL – DOI 10.5281/zenodo.7941382. – SWH swh:1:dir:875537f11cad3f77fcd8fc7b313d27605118a634.
Open peer review is available at https://openreview.net/forum?id=uVHUy7CWCL.

Communication with original authors — We emailed the first author of the paper twice. First at the beginning of our undertaking, they were enthusiastic about our attempt, and clarified a few initial doubts about their implementation, the embeddings, and missing files. As per the writing of this paper, they have not responded to the second email.

1 Introduction

In recent years, facial recognition (FR) systems have become increasingly important [7]. One sub-area of FR, *face verification*, attempts to determine whether two faces belong to the same person. However, these methods were repeatedly shown to be biased against certain demographic subgroups defined by attributes such as ethnicity, age or gender [8, 9]. This was especially frequent when considering how often a face in incorrectly interpreted as a match, i.e. false positive rate (FPR). This has serious ethical implications in contexts such as law enforcement, security and privacy.
Often, these methods use a deep neural network to generate representations of images called *embeddings*. The images are said to be of the same person if the cosine similarity of their embeddings exceeds a certain threshold. However, as Salvador *et al.* [1] show, this 'baseline' method does not provide the same accuracy for all ethnic subgroups. One approach to reduce the bias has been to simply learn less biased representations [10, 11], which generally also results in lower accuracy [12], requires re-training of the model, and/or requires the information about the sensitive attribute.
In an effort to alleviate this bias, Salvador *et al.* propose the *FairCal* calibration method. FairCal is a post-processing method that clusters image embeddings using K-means clustering and calibrates the cosine similarity scores between the images based on their cluster membership. A more detailed description of FairCal can be found in section 3.1. The authors claim that FairCal is fairly calibrated, achieves equal false positive rates (FPRs) across subgroups (i.e. fewer incorrect matches), while achieving state-of-the-art (SOTA) accuracy, all without the need for retraining or knowledge about sensitive attributes such as ethnicity. This paper aims to reproduce their results.

2 Scope of reproducibility

In this work, we run several experiments to explore the authors' findings and attempt to verify their claims, namely:

- FairCal achieves SOTA accuracy for face verification on two large datasets,
- FairCal outputs SOTA fairness-calibrated probabilities without knowledge of the sensitive attribute,
- FairCal reduces the gap in FPRs across sensitive attributes compared to the baseline method.

The authors also introduce another method named *Oracle*, which works similarly to FairCal, but uses explicit knowledge of the sensitive attributes to group images instead of unsupervised clustering. We report our results for Oracle and other benchmarking methods used by the authors in Appendix B, but focus on FairCal here. We subjectively decided a result to be *reproducible* if they did not deviate greatly from the authors' values.

3 Methodology

The authors provide an open-source implementation of their setup on GitHub [2]. Though the code for the implementation of FairCal and Oracle required minimal debugging, it had room for improvement in readability and efficiency – thus, it was refactored keeping this in mind. Furthermore, code files were added to generate the image embeddings and the files describing the pairs in the datasets (see section 3.2 for details). For the rest of this section, we followed the authors' implementations unless otherwise specified.

3.1 Model descriptions

Salvador *et al.* propose FairCal, a post-processing method to ensure face verification which is fairly calibrated across subgroups, and exhibits predictive equality. The authors define a model to be *calibrated* if the true probability of a match is equal to the model's confidence output. Thus, a binary classifier is fairly-calibrated if it is calibrated when conditioned on each subgroup. For a pair of images (x_1, x_2) with corresponding embeddings $(f(x_1), f(x_2))$, calibration can be seen as a function μ that maps the cosine similarity score, $s(x_1, x_2) = \frac{f(x_1).f(x_2)}{|f(x_1)||f(x_2)|}$, to probabilities. This map can then be used to define a binary classifier, where the score threshold s_{thr} can be computed. Additionally, the authors define a binary classifier to exhibit predictive equality for subgroups g_1 and g_2 if the classifier has equal FPRs for each subgroup.
Mathematical details of this method can be found in the original paper. Briefly, the authors describe FairCal as having three steps:

1 Apply K-means to the image embeddings, forming K clusters Z_k. Use these to create K calibration sets S_k^{cal}, which contain similarity scores of embedding pairs where either x_1 or x_2 belong to Z_k.

2 Use a post-hoc calibration method on S_k^{cal} to compute a map μ_k from cosine similarities to cluster-conditional probabilities. The authors use beta calibration [13].

3 Compute their calibrated score for each image pair (x_1, x_2) in the test set. Now, if the embeddings $f(x_1)$ and $f(x_2)$ belong to clusters k_1 and k_2 respectively, their overall calibrated score is defined as a linear combination of their respective mapped calibrated scores $\mu_{k_1}(s(x_1, x_2))$ and $\mu_{k_2}(s(x_1, x_2))$, weighted by the relative population fraction of the two calibration sets $S_{k^1}^{cal}$ and $S_{k^2}^{cal}$.

The authors use three pretrained models to generate embeddings:

- **'Facenet (VGGFace2)'**: An Inception Resnet model trained on the VGGFace2 dataset.
- **'Facenet (Webface)'**: An Inception Resnet model trained on CASIA-Webface dataset.
- **'ArcFace'**: An ArcFace model trained on the refined version of MS-Celeb-1M.

To pre-process the images, they employed a Multi-Task Convolutional Neural Network (MTCNN), removed images for which it failed to identify a face, and cropped and aligned those for which is did identify a face. This pipeline was attained from [3] and [4], where default hyperparameters were used. To create the functions that map the cosine similarity scores to probabilities and determine the thresholds of those probabilities, the authors used beta calibration. The authors used 100 clusters for their results.
To benchmark FairCal, the authors define a *baseline* method. This method simply calculates the cosine similarity of two image embeddings, and determines them to be a genuine pair if the cosine similarity exceeds a certain threshold determined using beta calibration. Beta calibration works by fitting a logistic regression model with respect to the cosine similarities and ground truth labels. Though logistic regression has no closed-form solution [14], repeated experiments with the baseline method on the same set of embeddings gave the exact same results (validated experimentally by us). The previous SOTA mentioned in the original paper is called *Fair Score Normalization* (FSN) [15], has also been reproduced in our tables and figures for comparison. The authors implemented several more methods to compare against FairCal, and our comparison of these can be found in Appendix B.
As an extension to the original paper, we experimented with using Gaussian mixture models (GMMs) [14] as a clustering method instead of K-means for FairCal, in an attempt to validate the use of unsupervised clustering methods in general to achieve the claims of Salvador *et al.* Additionally we hypothesized that the clusters found by this approach may be able to model the underlying structure of the embeddings as well as, if not better,

than clusters that arise from K-means, which tend to be roughly isotropic and sensitive to outliers [16]. Thus, the embedding space was partitioned into discrete regions, with each data point being assigned to the cluster corresponding to the Gaussian component with the highest probability. Additional experiments were done with different numbers of Gaussian components, which can be found in Appendix E. For the sake of comparison with the K-means version of FairCal, 100 Gaussian components were used for the results mentioned in sections below (with this method referred to as *FairCal-GMM*).

3.2 Datasets

The authors used two datasets: the *Balanced Faces in the Wild* (BFW) dataset [5] and *Racial Faces in the Wild* (RFW) [6]. Both datasets label their images by ethnicity (African, Asian, Caucasian or Indian) and BFW also labels by gender (Female, Male). In RFW, all image pairs are same-ethnicity images, while BFW includes both mixed-gender and mixed-ethnicity pairs. The authors report that RFW is made up of images from MS-Celeb-1M, and BFW is made up of images from VGGFace2. These were used to train the ArcFace and Facenet (VGGFace2) models respectively. Thus, the BFW dataset could only be used with the FaceNet (Webface) and ArcFace models, while RFW could only be used with the two FaceNet models, to ensure the models were tested on unseen examples.
The MTCNN pipelines provided by [3] ('Facenet-MTCNN') and [4] ('Arcface-MTCNN') were used to preprocess images. After the Facenet-MTCNN pre-processing, there remained 23,903 image pairs for RFW dataset. The authors reported using 23,541 pairs, meaning we used 1.5% more pairs. For the BFW dataset, 891,622 image pairs were available after Facenet-MTCNN pre-processing and 888,833 after Arcface-MTCNN, while the authors reported using 890,347 pairs, meaning we used 0.14% more and 0.17% fewer pairs respectively. For both datasets, we found the same approximate ratio of genuine/imposter pairs as the authors, namely 1:1 for RFW and 1:3 for BFW. The authors used folds given alongside each dataset to perform five-fold cross validation.
Since Salvador *et al.* do not provide their implementation for the MTCNN pipeline, it is possible that our implementations differed, which led to the differing amount of images for each dataset. Thus, we cannot be sure that the pairs the authors used are included in our version and vice versa, meaning differences between results can occur.
The authors' code base refers to CSV files relating to each of the datasets, which provide essential information to replicate the author's experiments: namely, the pairing of images and their embeddings' cosine similarities. However, these files were not provided in the author's repository, nor were there instructions on how to find them. The paper's first author generously provided us with a format for each file during correspondence, which, combined with metadata available on the datasets themselves, allowed us to create compatible versions of these files.

3.3 Hyperparameters

100 clusters were used for the K-means clustering, the same as Salvador *et al.*, and for FairCal-GMM. In Appendix C, we show the results for varying amounts of clusters. Additionally, we use the default MTCNN hyperparameters specified in [3] and [4].

3.4 Experimental setup and code

Our setup was based on existing code by Salvador *et al.* in a GitHub repository [2]. We improved the experimental setup by making the code more efficient, primarily by using more efficient data structures. In the author's code, the time to run one experiment (i.e. one dataset for one model) using FairCal was approximately 2 hours. We were able to reduce this to around 54 seconds on average (≈133x speedup). For the sake of reproducibility, we set the random seed to 42, since pseudo-random initialization was used for both K-means and GMM clustering.

3.5 Computational requirements

The authors specify several approaches in their original paper in addition to FairCal. Most are computationally light and are thus able to run on a CPU in relatively little time. For the results provided in this paper, an AMD Ryzen 7 2700X CPU was used. To speed up the K-means and GMM algorithms, we used a GPU implementation [17], which was run on a NVIDIA GeForce GTX1080Ti. An additional benchmarking method, AGENDA (used in Table 5), requires the training of a multilayer perceptron, which was also done using the above GPU. The runtimes for the main experiments can be found in Table 1.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Preprocessing	Embeddings	2237	2198	284	4085
	Cosine sims	11	11	1	1
Experiments	Baseline	1	1	12	15
	FSN	53	47	45	47
	FairCal	54	54	53	57
	FairCal-GMM	247	108	68	576

Table 1. The runtimes for the pre-processing of the dataset and the experiments in seconds. The row *Embeddings* refers to the generation of image embeddings, and the row *Cosine sims* refers calculating cosine similarities for all image embedding pairs.

4 Results

4.1 Results reproducing original paper

Claim 1 (Accuracy) — The first claim we aimed to reproduce was that FairCal achieves SOTA accuracy for face verification on two large datasets. As can be seen in Table 2, we were able to reproduce most of the author's results regarding this over all datasets and models. Thus, we have successfully reproduced this claim.

		AUROC		TPR @ 0.1% FPR		TPR @ 1% FPR	
RFW		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	Baseline	88.26±0.19	89.97±0.58	18.42±1.28	25.27±6.51	34.88±3.27	39.92±2.40
	FSN	90.05±0.29	91.30±0.35	23.01±2.00	26.79±4.63	40.21±2.09	44.52±2.91
	FairCal	90.58±0.29	92.17±0.40	23.55±1.82	26.93±5.23	41.88±1.99	49.68±2.40
	FairCal-GMM	–	92.46±0.43	–	29.88±4.34	–	50.86±3.42
FaceNet (Webface)	Baseline	83.95±0.22	84.46±0.47	11.18±3.45	11.14±5.34	26.04±2.11	26.45±4.90
	FSN	85.84±0.34	86.24±0.63	17.33±3.01	17.98±5.74	32.90±1.03	31.68±2.02
	FairCal	86.71±0.25	86.97±0.72	20.64±3.09	19.23±3.64	33.13±1.67	33.82±4.55
	FairCal-GMM	–	87.19±0.53	–	20.49±5.36	–	33.44±4.09
BFW							
FaceNet (Webface)	Baseline	96.06±0.16	94.62±0.17	33.61±2.10	27.93±2.02	58.87±0.92	52.79±1.74
	FSN	96.77±0.20	94.84±0.22	47.11±1.23	37.87±0.98	69.92±1.01	59.86±1.23
	FairCal	96.90±0.17	95.67±0.13	46.74±1.49	37.68±0.87	69.21±1.19	60.21±1.09
	FairCal-GMM	–	95.48±0.15	–	35.39±1.46	–	58.49±1.57
ArcFace	Baseline	97.41±0.34	97.34±0.36	86.27±1.09	84.75±1.26	90.11±0.87	89.51±0.98
	FSN	97.35±0.33	97.32±0.35	86.19±1.13	84.77±1.20	90.06±0.84	89.49±0.98
	FairCal	97.44±0.34	97.37±0.35	86.28±1.24	84.95±1.32	90.14±0.86	89.55±1.01
	FairCal-GMM	–	97.35±0.37	–	84.78±1.21	–	89.51±1.00

Table 2. Global accuracy measured by AUROC, TPR at 0.1% FPR threshold and TPR at 1% FPR threshold. Higher is better. If we successfully reproduced a result, our result is colored green. Entries indicate mean ± 1 standard deviation over 5-fold validation.

Claim 2 (Fairness Calibration) — Secondly, we aimed to reproduce the authors' claim that FairCal outputs SOTA fairness-calibrated probabilities without knowledge of the sensitive attribute. The authors measure this by computing the Kolmogorov-Smirnov (KS) score per subgroup and comparing the mean KS across subgroups, the average absolute deviation from the mean (AAD), maximum absolute deviation from the mean (MAD) and standard deviation from the mean (STD). By construction, FairCal does not take into account the sensitive attribute. Our results can be seen in Table 3. We find these to be partly reproducible, due to the fact that we are able to reproduce the difference in mean KS between baseline and FairCal, but the authors find a larger difference than we do. Furthermore, the AAD, MAD and STD we attained are not consistently similar to the authors' results.

		Mean		AAD		MAD		STD	
RFW		Authors	Ours	Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	Baseline	6.37	6.29	2.89	2.60	5.73	5.10	3.77	3.84
	FSN	1.43	3.67	0.35	0.59	0.57	1.07	0.40	0.79
	FairCal	1.37	1.67	0.28	0.47	0.50	0.93	0.34	0.66
	FairCal-GMM	–	1.77	–	0.60	–	0.98	–	0.78
FaceNet (Webface)	Baseline	5.55	5.55	2.48	2.31	4.97	4.60	2.91	3.34
	FSN	2.49	3.90	0.84	0.54	1.19	1.05	0.91	0.75
	FairCal	1.75	1.74	0.41	0.48	0.64	0.92	0.45	0.67
	FairCal-GMM	–	1.75	–	0.48	–	0.82	–	0.62
BFW									
FaceNet (Webface)	Baseline	6.77	4.72	3.63	2.83	5.96	7.62	4.03	3.50
	FSN	2.76	3.45	1.38	1.40	2.67	4.60	1.60	1.90
	FairCal	3.09	3.06	1.34	1.19	2.48	2.59	1.55	1.45
	FairCal-GMM	–	3.43	–	1.40	–	3.64	–	1.80
ArcFace	Baseline	2.57	2.17	1.39	1.24	2.94	3.30	1.63	1.57
	FSN	2.65	2.91	1.45	1.29	3.23	4.26	1.71	1.72
	FairCal	2.49	1.94	1.30	1.16	2.68	3.09	1.52	1.46
	FairCal-GMM	–	1.58	–	0.85	–	2.71	–	1.13

Table 3. Fairness Calibration measured by KS score across sensitive subgroups. Measured by mean, the average absolute deviation from the mean (AAD), maximum absolute deviation from the mean(MAD) and standard deviation from the mean (STD). Lower is better in all cases. If we successfully reproduced a result, our result is colored green.

Claim 3 (Gap in FPRs) — Thirdly, we aimed to reproduce the authors' claim that FairCal reduces the gap in FPRs across sensitive attributes compared to the baseline method. The results can be seen in Figure 1 and Table 4. From this, we conclude that the authors' results are reproducible in most cases. We were able to decrease the gap in FPRs across sensitive attributes, but not for all experiments and not as strongly as the authors' results show – thus, we see the same trends despite having different absolute results.

4.2 Results beyond original paper

Additional Result 1 — As was described in section 3.4, we extended the original paper by using Gaussian mixture models (GMMs) instead of the K-means algorithm to cluster the image embeddings. These results can be found in Tables 2, 3 and 4. Concerning accuracy, FairCal-GMM achieves the same or even slightly higher results than FairCal, the KS score same or lower than FairCal, and comparable deviation in subgroup FPRs.

Additional Result 2 — In their original paper, the authors compare the group FPRs for different values of global FPR. The authors claim that if these lines are closer together, it reflects better fairness (cf. figure 1 in Salvador *et al.* [1]). Thus, in a perfectly fair setting,

Global FPR: 1%		AAD		MAD		STD	
RFW		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	Baseline	0.68	0.74	1.02	0.94	0.74	0.90
	FSN	0.37	0.42	0.68	0.77	0.46	0.58
	FairCal	0.28	0.59	0.46	0.95	0.32	0.77
	FairCal-GMM	–	0.44	–	0.74	–	0.59
FaceNet (Webface)	Baseline	0.67	0.68	1.23	1.21	0.79	0.91
	FSN	0.35	0.33	0.61	0.58	0.40	0.45
	FairCal	0.29	0.33	0.57	0.53	0.35	0.44
	FairCal-GMM	–	0.32	–	0.62	–	0.47
BFW							
FaceNet (Webface)	Baseline	2.42	1.99	7.48	6.30	3.22	2.85
	FSN	0.87	0.72	2.19	1.71	1.05	0.96
	FairCal	0.80	0.60	1.79	1.52	0.95	0.81
	FairCal-GMM	–	0.66	–	1.54	–	0.89
ArcFace	Baseline	0.72	0.70	1.51	1.58	0.85	0.92
	FSN	0.55	0.60	1.27	1.45	0.68	0.80
	FairCal	0.63	0.62	1.46	1.34	0.78	0.80
	FairCal-GMM	–	0.70	–	1.54	–	0.91

Global FPR: 0.1%		AAD		MAD		STD	
RFW		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	Baseline	0.10	0.17	0.15	0.31	0.10	0.22
	FSN	0.10	0.19	0.18	0.29	0.11	0.24
	FairCal	0.09	0.18	0.14	0.33	0.10	0.24
	FairCal-GMM	–	0.19	–	0.37	–	0.25
FaceNet (Webface)	Baseline	0.14	0.14	0.26	0.27	0.16	0.19
	FSN	0.11	0.16	0.23	0.26	0.23	0.20
	FairCal	0.09	0.17	0.16	0.23	0.10	0.21
	FairCal-GMM	–	0.15	–	0.21	–	0.19
BFW							
FaceNet (Webface)	Baseline	0.29	0.25	1.00	0.83	0.40	0.36
	FSN	0.09	0.09	0.20	0.17	0.11	0.11
	FairCal	0.09	0.08	0.20	0.19	0.11	0.10
	FairCal-GMM	–	0.10	–	0.28	–	0.14
ArcFace	Baseline	0.12	0.12	0.30	0.27	0.15	0.16
	FSN	0.11	0.11	0.28	0.23	0.14	0.14
	FairCal	0.11	0.10	0.31	0.24	0.15	0.13
	FairCal-GMM	–	0.13	–	0.27	–	0.16

Table 4. Predictive equality, measured by the deviation in subgroup FPRs in terms of average absolute Deviation (AAD), maximum absolute deviation (MAD), and standard deviation (STD). Lower is better in all cases. The top and bottom tables report results for a global FPR of 0.1% and 1% respectively. If we successfully reproduced a result, our result is colored green.

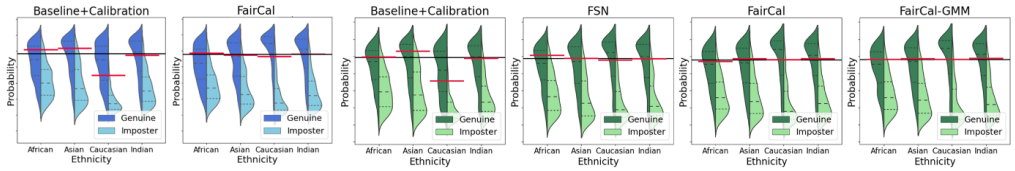


Figure 1. Illustration of bias in FPR rates globally and across subgroups, with the FaceNet (Webface) feature model. Plots from the original paper are in blue (their figure 2) and plots from the current paper are in green. False positives are indicated by imposter pairs above the decision threshold value (horizontal lines). Black horizontal lines indicate thresholds which achieve a global FPR of 5%. Red horizontal lines indicate thresholds which achieve a FPR of 5% for a specific subgroup. Greater distance of red horizontal lines from the black horizontal line indicates greater bias. This shows that the baseline method is biased. Furthermore, global calibration (based on cosine similarity alone) does not reduce the bias. Lastly, we see that the FairCal method reduces bias in both the original and the present paper.

the lines are identical. As such, *un*fairness may be quantified by measuring the extent to which the lines differ. In an attempt to do so, we fitted a line using linear regression to all points in subgroup-data and calculated the sum of squared error residuals from each subgroup to the fitted line. These results can be seen in Figure 2. Note that the points included in this analysis are limited to Global FPR ≤ 0.1, similarly to the authors. From this, we can see that FairCal brings improvements over the baseline, and performs comparably to FairCal-GMM.

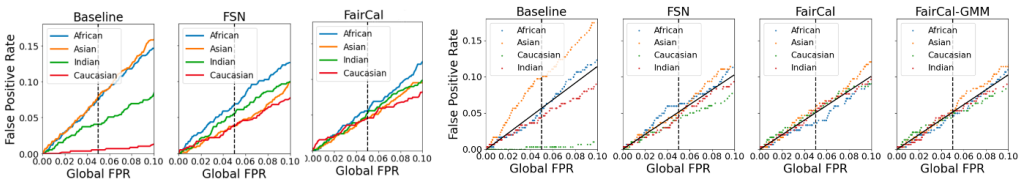


Figure 2. Illustration of quantified reduction in bias (improved fairness) measured by the FPRs evaluated on intra-ethnicity pairs on the RFW dataset with the FaceNet (Webface) feature model. The two leftmost plots are from the original paper (their figure 1) and the three rightmost plots are from the current paper. The best fit line for our experiments are indicated in black. Lines closer together indicate better fairness. Residual (SSE) values for the corresponding methods are: Baseline (0.499), FSN (0.037), FairCal (0.026), FairCal-GMM (0.019).

5 Discussion

5.1 Discussion of the results

Overall, we found the authors' results to be mostly reproducible. While we largely saw the same patterns as the authors, we were unable to consistently reproduce their specific values. This is particularly remarkable for the baseline method, since it is deterministic. We verified this by using the embeddings *we* generated and attained the same results for repeated runs of the baseline method for all datasets and models. As was explained in section 3.2, we believe that the discrepancy in image embeddings is also the primary basis for the differences in the results.
Additionally, since FairCal is non-deterministic due to using K-means (specifically in the initialization of cluster centroids), the authors may have had a very different initialization than we did. Thus, a future study may validate these results further by analysing several experiments with known seeds.

We also find that GMM clustering is at least as good as using K-means, and in some cases, outperforms it. This fits well with the theoretical underpinning of GMMs, since they may produce more flexible, non-isotropic clusters. This also validates the underlying motivation of FairCal, which is that information about sensitive attributes is not required to have fair calibration, and that unsupervised clustering methods are a suitable replacement for sensitive attribute labels. However, the difference is usually under one standard deviation for each metric. Thus, more experimentation and analysis is required to confirm any additional improvement. We recommended this as a direction for further research.

5.2 What was easy

We were very fortunate to be able to start with the author's implementation of the code, which required only minimal debugging. Once we had our datasets and embeddings in the correct format, we were able to run the code without issues. Additionally, though the RFW and BFW datasets are not freely available to the public, they are to researchers – thus, contacting their authors allowed us to swiftly attain them. Furthermore, the authors' description of FairCal and how it implements fairness is very intuitive and clear in their paper. As a result, we were able to understand their methods without prior knowledge in the field of fairness in face verification.

5.3 What was difficult

There were mainly three difficulties: first, the environment that the authors provided with their code was not project-specific, contained conflicting dependencies, and assumed that the user had a specific operating system installed. Additionally, environments built from scratch often had issues with dependency conflicts depending on OS and hardware. Moreover, since the code expected the user to already possess files with image embeddings, several packages needed to generate them were not listed.
Second, the pipeline of how to generate the image embeddings was not clear. For example, the authors report that they used an MTCNN for pre-processing; however, the Facenet-MTCNN repository has a different implementation from the one containing the Arcface-MTCNN. Thus, it was ambiguous which MTCNN was used for which model.
Third, the original authors' code was challenging to understand and optimize, especially since some of the functions or column names were renamed without adjusting downstream references. Fortunately, fixing these naming errors allowed us to run the code from start to end, albeit with a high runtime. We noted that some areas of the code could benefit from more efficient data structures, and the refactored code can be found in our repository.
Finally, in our opinion, certain design decisions of the original authors need further verification, perhaps as future research directions. For example, though they do refer back to another paper's suggestions, it was not clear to us why 100 clusters should be the best performing specifically for FairCal. In addition, though KS scores are mentioned as a strong metric, others, such as ECE and Brier scores, may be more useful in different settings, and could provide insight into how these calibration methods work.

5.4 Communication with original authors

We emailed the first author of the paper twice. First at the beginning of our undertaking, they were enthusiastic about our attempt, and clarified a few initial doubts about their implementation, the embeddings, and missing files. As per the writing of this paper, they have not responded to the second email.

References

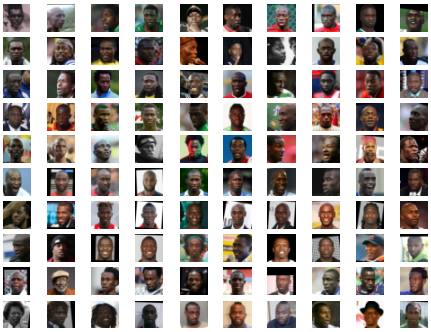
1. T. Salvador, S. Cairns, V. Voleti, N. Marshall, and A. M. Oberman. "FairCal: Fairness Calibration for Face Verification." In: **International Conference on Learning Representations**. 2022. URL: https://openreview.net/forum?id=nRj0NcmSuxb.
2. T. Salvador. **GitHub Repository: FairCal: Fair Calibration for Face Verification**. https://github.com/tiagosalvador/faircal. 2021.
3. T. Esler. **Face recognition using pytorch**. https://github.com/timesler/facenet-pytorch. 2021.
4. A. Sharma. **ONNX Model Zoo: Arcface**. https://github.com/onnx/models/tree/main/vision/body_analysis/arcface. 2021.
5. J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. "Face recognition: too bias, or not too bias?" In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops**. 2020.
6. M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. "Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network." In: **The IEEE International Conference on Computer Vision (ICCV)**. Oct. 2019.
7. I. Masi, Y. Wu, T. Hassner, and P. Natarajan. "Deep face recognition: A survey." In: **2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)**. IEEE. 2018, pp. 471–478.
8. J. Buolamwini and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In: **Proceedings of the 1st Conference on Fairness, Accountability and Transparency**. Ed. by S. A. Friedler and C. Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.
9. M. Alvi, A. Zisserman, and C. Nellaker. **Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings**. 2018. DOI: 10.48550/ARXIV.1809.02169. URL: https://arxiv.org/abs/1809.02169.
10. J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu. **Additive Adversarial Learning for Unbiased Authentication**. 2019. DOI: 10.48550/ARXIV.1905.06517. URL: https://arxiv.org/abs/1905.06517.
11. A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. "Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data." In: **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. 2019, pp. 2261–2268. DOI: 10.1109/CVPRW.2019.00279.
12. S. Gong, X. Liu, and A. K. Jain. **Jointly De-biasing Face Recognition and Demographic Attribute Estimation**. 2019. DOI: 10.48550/ARXIV.1911.08080. URL: https://arxiv.org/abs/1911.08080.
13. M. Kull, T. Silva Filho, and P. Flach. "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers." In: **Artificial Intelligence and Statistics**. PMLR. 2017, pp. 623–631.
14. C. M. Bishop and N. M. Nasrabadi. **Pattern recognition and machine learning**. Vol. 4. 4. Springer, 2006.
15. P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. "Post-comparison mitigation of demographic bias in face recognition using fair score normalization." In: **Pattern Recognition Letters** 140 (2020), pp. 332–338. DOI: https://doi.org/10.1016/j.patrec.2020.11.007. URL: https://www.sciencedirect.com/science/article/pii/S0167865520304128.
16. Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little. "What to do when K-means clustering fails: a simple yet principled alternative algorithm." In: **PloS one** 11.9 (2016), e0162259.
17. O. Borchert. **PyCave**. https://pycave.borchero.com/. 2022.
18. P. Dhar, J. Gleason, H. Souri, C. D. Castillo, and R. Chellappa. "An adversarial learning algorithm for mitigating gender bias in face recognition." In: **CoRR** abs/2006.07845 (2020). arXiv:2006.07845. URL: https://arxiv.org/abs/2006.07845.
19. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. **On Calibration of Modern Neural Networks**. 2017. DOI: 10.48550/ARXIV.1706.04599. URL: https://arxiv.org/abs/1706.04599.
20. M. H. DeGroot and S. E. Fienberg. "The Comparison and Evaluation of Forecasters." In: **Journal of the Royal Statistical Society. Series D (The Statistician)** 32.1/2 (1983), pp. 12–22. URL: http://www.jstor.org/stable/2987588 (visited on 01/30/2023).

6 Appendix

6.1 Appendix A

In this appendix we present the visualisation of some clusters obtained by the K-means and GMM algorithms. Results can be found in Figures 3 and 4 respectively. Similarly to Figure 3 from the original paper, we conclude that our clusters have clear semantic meaning, both for K-means and GMMs.

(a) Indian women with dark hair



(b) Male African sportsmen

(c) Young Asian women with red hair and bangs

(d) Caucasian men with short dark hair

Figure 3. Examples of clusters obtained with the K-means algorithm (K=100) on the RFW dataset based on the feature embeddings computed with the FaceNet (Webface) model, labelled with potential semantic classes (labelled by us).

(a) Older Asian men (b) Blonde Caucasian women

(c) Indian men with mustache (d) Young Indian women

Figure 4. Examples of clusters obtained with the GMM algorithm (k = 100) on the RFW dataset based on the feature embeddings computed with the FaceNet (VGGFace2) model, labelled with potential semantic classes (labelled by us).

6.2 Appendix B

We present additional figures of the experiments on the other methods used by the original authors: the Adverserial Gender De-biasing algorithm (AGENDA) [18], the Fair Score Normalization (FSN) method [15], and Oracle (introduced by the original authors). For clarity, we also include the results for Baseline, FairCal with K-means and FairCal with GMM. Results can be found in Tables 5, 6 and 7.

6.3 Appendix C

In this section, we present the performance of fairness-calibration for the Expected Calibration Error (ECE) [19] and Brier [20] scores. The original authors mention these scores in their paper, but do not show their results. Here, we show that the FairCal-GMM method outperforms regular FairCal in most cases. Results can be found in Tables 8 and 9.

6.4 Appendix D

In this section, we present further results on the performance of FairCal-GMM as compared to the baseline and original FairCal methods. We compare the performance of these methods by reporting the deviation in subgroup FNRs in terms of average absolute Deviation (AAD), maximum absolute deviation (MAD), and standard deviation (STD). Results can be found in Table 10.

RFW		AUROC		TPR @ 0.1% FPR		TPR @ 1% FPR	
		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	AGENDA	76.83±0.57	79.12±0.67	8.32±1.86	13.33±2.87	18.01±1.44	21.63±2.56
	Baseline	88.26±0.19	89.97±0.58	18.42±1.28	25.27±6.51	34.88±3.27	39.92±2.40
	FairCal	90.58±0.29	92.17±0.40	23.55±1.82	26.93±5.23	41.88±1.99	49.68±2.40
	FSN	90.05±0.29	91.30±0.35	23.01±2.00	26.79±4.63	40.21±2.09	44.52±2.91
	Oracle	89.74±0.31	91.26±0.50	21.40±3.54	25.45±6.20	411.83±2.98	47.98±4.92
	FairCal-GMM	–	92.46±0.43	–	29.88±4.34	–	50.86±3.42
FaceNet (Webface)	AGENDA	74.51±0.94	72.79±1.36	6.38±0.78	6.89±3.27	14.98±1.11	13.97±2.58
	Baseline	83.95±0.22	84.46±0.47	11.18±3.45	11.14±5.34	26.04±2.11	26.45±4.90
	FairCal	86.71±0.25	86.97±0.72	20.64±3.09	19.23±3.64	33.13±1.67	33.82±4.55
	FSN	85.84±0.34	86.24±0.63	17.33±3.01	17.98±5.74	32.90±1.03	31.68±2.02
	Oracle	85.23±0.18	85.72±0.54	16.71±1.98	16.72±2.82	31.60±1.08	32.20±4.37
	FairCal-GMM	–	87.19±0.53	–	20.49±5.36	–	33.44±4.09
BFW							
FaceNet (Webface)	AGENDA	82.42±0.45	76.33±0.76	15.95±1.53	6.84±0.72	32.51±1.24	18.58±1.25
	Baseline	96.06±0.16	94.62±0.17	33.61±2.10	27.93±2.02	58.87±0.92	52.79±1.74
	FairCal	96.90±0.17	95.67±0.13	46.74±1.49	37.68±0.87	69.21±1.19	60.21±1.09
	FSN	96.77±0.20	94.84±0.22	47.11±1.23	37.87±0.98	69.92±1.01	59.86±1.23
	Oracle	97.28±0.13	96.18±0.10	45.13±1.45	35.34±1.02	67.56±1.05	58.99±1.01
	FairCal-GMM	–	95.48±0.15	–	35.39±1.46	–	58.49±1.57
ArcFace	AGENDA	95.09±0.55	93.17±0.54	69.61±2.40	44.97±3.06	79.67±2.06	64.09±2.38
	Baseline	97.41±0.34	97.34±0.36	86.27±1.09	84.75±1.26	90.11±0.87	89.51±0.98
	FairCal	97.44±0.34	97.37±0.35	86.28±1.24	84.95±1.32	90.14±0.86	89.55±1.01
	FSN	97.35±0.33	97.32±0.35	86.19±1.13	84.77±1.20	90.06±0.84	89.49±0.98
	Oracle	98.91±0.12	98.85±0.13	86.41±1.19	84.98±1.24	90.40±0.91	89.87±1.06
	FairCal-GMM	–	97.35±0.37	–	84.78±1.21	–	89.51±1.00

Table 5. Global accuracy measured by AUROC, TPR at 0.1% FPR threshold and TPR at 1% FPR threshold. Higher is better. Entries indicate mean ± 1 standard deviation over 5-fold validation.

RFW		Mean		AAD		MAD		STD	
		Authors	Ours	Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	AGENDA	7.71	8.87	3.11	2.66	6.09	4.96	3.86	3.64
	Baseline	6.37	7.28	2.89	2.37	5.73	4.57	3.77	3.30
	FairCal	1.37	3.30	0.28	0.51	0.50	0.96	0.34	0.71
	FSN	1.43	3.67	0.35	0.59	0.57	1.07	0.40	0.79
	Oracle	1.18	3.50	0.28	0.97	0.53	1.80	0.33	1.33
	FairCal - GMM	–	1.77	–	0.60	–	0.98	–	0.78
FaceNet (Webface)	AGENDA	7.71	8.87	3.11	2.66	6.09	4.96	3.86	3.64
	Baseline	6.37	7.28	2.89	2.37	5.73	4.57	3.77	3.30
	FairCal	1.37	3.30	0.28	0.51	0.50	0.96	0.34	0.71
	FSN	1.43	3.67	0.35	0.59	0.57	1.07	0.40	0.79
	Oracle	1.18	3.50	0.28	0.97	0.53	1.80	0.33	1.33
	FairCal - GMM	–	1.75	–	0.48	–	0.82	–	0.62
BFW									
FaceNet (Webface)	AGENDA	13.21	14.06	6.37	4.94	12.91	14.94	7.55	6.40
	Baseline	6.77	5.80	3.63	2.88	5.96	6.86	4.03	3.49
	FairCal	3.09	3.38	1.34	1.12	2.48	2.43	1.55	1.38
	FSN	2.76	3.45	1.38	1.40	2.67	4.60	1.60	1.90
	Oracle	2.23	2.64	1.15	1.00	2.63	2.76	1.40	1.29
	FairCal - GMM	–	3.43	–	1.40	–	3.64	–	1.80
ArcFace	AGENDA	5.14	17.43	2.48	3.43	5.92	8.85	3.04	4.27
	Baseline	2.57	2.36	1.39	1.15	2.94	3.15	1.63	1.47
	FairCal	2.49	2.17	1.30	1.04	2.68	2.87	1.52	1.32
	FSN	2.65	2.91	1.45	1.29	3.23	4.26	1.71	1.72
	Oracle	1.41	2.54	0.59	0.86	1.30	2.16	0.69	1.08
	FairCal - GMM	–	1.58	–	0.85	–	2.71	–	1.13

Table 6. Fairness Calibration measured by KS score across sensitive subgroups. Measured by mean, the average absolute deviation from the mean (AAD), maximumum absolute deviation from the mean(MAD) and standard deviation from the mean (STD). Lower is better in all cases.

Global FPR: 1%		AAD		MAD		STD	
RFW		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	AGENDA	0.71	0.71	1.14	1.19	0.81	0.94
	Baseline	0.68	0.74	1.02	0.94	0.74	0.90
	FairCal	0.28	0.59	0.46	0.95	0.32	0.77
	FSN	0.37	0.42	0.68	0.77	0.46	0.58
	Oracle	0.40	0.44	0.69	0.75	0.45	0.58
	FairCal-GMM	–	0.44	–	0.74	–	0.59
FaceNet (Webface)	AGENDA	0.73	0.89	1.08	1.26	0.78	1.10
	Baseline	0.67	0.68	1.23	1.21	0.79	0.91
	FairCal	0.29	0.33	0.57	0.53	0.35	0.44
	FSN	0.35	0.33	0.61	0.58	0.40	0.45
	Oracle	0.41	0.41	0.74	0.68	0.48	0.53
	FairCal-GMM	–	0.32	–	0.62	–	0.47
BFW							
FaceNet (Webface)	AGENDA	1.21	0.63	3.09	1.43	1.51	0.81
	Baseline	2.42	1.99	7.48	6.30	3.22	2.85
	FairCal	0.80	0.60	1.79	1.52	0.95	0.81
	FSN	0.87	0.72	2.19	1.71	1.05	0.96
	Oracle	0.77	0.67	1.71	1.43	0.91	0.84
	FairCal-GMM	–	0.66	–	1.54	–	0.89
ArcFace	AGENDA	0.65	0.40	1.78	0.93	0.84	0.52
	Baseline	0.72	0.70	1.51	1.58	0.85	0.92
	FairCal	0.63	0.62	1.46	1.34	0.78	0.80
	FSN	0.55	0.60	1.27	1.45	0.68	0.80
	Oracle	0.83	0.74	2.08	1.84	1.07	1.02
	FairCal-GMM	–	0.70	–	1.54	–	0.91

Global FPR: 0.1%		AAD		MAD		STD	
RFW		Authors	Ours	Authors	Ours	Authors	Ours
FaceNet (VGGFace2)	AGENDA	0.11	0.16	0.20	0.28	0.13	0.20
	Baseline	0.10	0.17	0.15	0.31	0.10	0.22
	FairCal	0.09	0.18	0.14	0.33	0.10	0.24
	FSN	0.10	0.19	0.18	0.29	0.11	0.24
	Oracle	0.11	0.19	0.19	0.36	0.12	0.25
	FairCal-GMM	–	0.19	–	0.37	–	0.25
FaceNet (Webface)	AGENDA	0.12	0.14	0.23	0.23	0.14	0.18
	Baseline	0.14	0.14	0.26	0.27	0.16	0.19
	FairCal	0.09	0.17	0.16	0.23	0.10	0.21
	FSN	0.11	0.16	0.23	0.26	0.23	0.20
	Oracle	0.11	0.15	0.20	0.24	0.13	0.19
	FairCal-GMM	–	0.15	–	0.21	–	0.19
BFW							
FaceNet (Webface)	AGENDA	0.14	0.09	0.40	0.20	0.18	0.11
	Baseline	0.29	0.25	1.00	0.83	0.40	0.36
	FairCal	0.09	0.08	0.20	0.19	0.11	0.10
	FSN	0.09	0.09	0.20	0.17	0.11	0.11
	Oracle	0.12	0.14	0.25	0.37	0.15	0.19
	FairCal-GMM	–	0.10	–	0.28	–	0.14
ArcFace	AGENDA	0.09	0.06	0.23	0.15	0.11	0.08
	Baseline	0.12	0.12	0.30	0.27	0.15	0.16
	FairCal	0.11	0.10	0.31	0.24	0.15	0.13
	FSN	0.11	0.11	0.28	0.23	0.14	0.14
	Oracle	0.12	0.11	0.27	0.25	0.14	0.14
	FairCal-GMM	–	0.13	–	0.27	–	0.16

Table 7. Predictive equality, measured by the deviation in subgroup FPRs in terms of average absolute Deviation (AAD), maximum absolute deviation (MAD), and standard deviation (STD). Lower is better in all cases. The top and bottom tables report results for a global FPR of 0.1% and 1% respectively.

RFW		Mean	AAD	MAD	STD
FaceNet (VGGFace2)	Baseline	6.29	2.60	5.10	3.84
	FairCal	1.67	0.47	0.93	0.66
	FairCal-GMM	1.77	0.60	0.98	0.78
FaceNet (Webface)	Baseline	5.55	2.31	4.60	3.34
	FairCal	1.74	0.48	0.92	0.67
	FairCal-GMM	1.75	0.48	0.82	0.62
BFW					
FaceNet (Webface)	Baseline	4.72	2.83	7.62	3.50
	FairCal	3.06	1.19	2.59	1.45
	FairCal-GMM	3.43	1.40	3.64	1.80
ArcFace	Baseline	2.17	1.24	3.30	1.57
	FairCal	1.94	1.16	3.09	1.46
	FairCal-GMM	1.58	0.85	2.71	1.13

Table 8. Fairness calibration measured by ECE score. Lower is better in all cases.

RFW		Mean	AAD	MAD	STD
FaceNet (VGGFace2)	Baseline	6.29	2.60	5.10	3.84
	FairCal	1.67	0.47	0.93	0.66
	FairCal-GMM-full	1.77	0.60	0.98	0.78
FaceNet (Webface)	Baseline	5.55	2.31	4.60	3.34
	FairCal	1.74	0.48	0.92	0.67
	FairCal-GMM-full	1.75	0.48	0.82	0.62
BFW					
FaceNet (Webface)	Baseline	4.72	2.83	7.62	3.50
	FairCal	3.06	1.19	2.59	1.45
	FairCal-GMM-full	3.43	1.40	3.64	1.80
ArcFace	Baseline	2.17	1.24	3.30	1.57
	FairCal	1.94	1.16	3.09	1.46
	FairCal-GMM-full	1.58	0.85	2.71	1.13

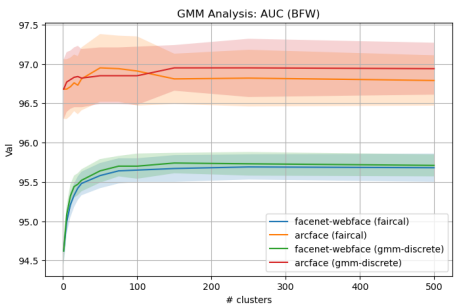
Table 9. Fairness calibration measured by Brier score. Lower is better in all cases.

Global FPR: 1%				
RFW		AAD	MAD	STD
FaceNet (VGGFace2)	Baseline	0.53	0.97	0.72
	FairCal	0.42	0.68	0.55
	FairCal-GMM	0.39	0.69	0.54
FaceNet (Webface)	Baseline	0.44	0.79	0.58
	FairCal	0.47	0.75	0.61
	FairCal-GMM	0.45	0.85	0.62
BFW				
FaceNet (Webface)	Baseline	0.36	0.87	0.48
	FairCal	0.32	0.63	0.41
	FairCal-GMM	0.32	0.63	0.40
ArcFace	Baseline	0.74	1.91	0.99
	FairCal	0.66	1.59	0.87
	FairCal-GMM	0.71	1.85	0.95
Global FPR: 0.1%				
RFW		AAD	MAD	STD
FaceNet (VGGFace2)	Baseline	0.14	0.27	0.18
	FairCal	0.11	0.22	0.15
	FairCal-GMM	0.11	0.22	0.14
FaceNet (Webface)	Baseline	0.09	0.18	0.12
	FairCal	0.07	0.14	0.10
	FairCal-GMM	0.12	0.23	0.16
BFW				
FaceNet (Webface)	Baseline	0.08	0.21	0.11
	FairCal	0.07	0.21	0.10
	FairCal-GMM	0.07	0.16	0.09
ArcFace	Baseline	0.11	0.33	0.15
	FairCal	0.10	0.29	0.14
	FairCal-GMM	0.10	0.30	0.14

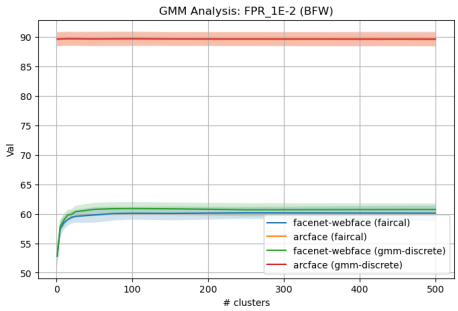
Table 10. Equal opportunity, measured by the deviation in subgroup FNRs in terms of average absolute Deviation (AAD), maximum absolute deviation (MAD), and standard deviation (STD). Lower is better in all cases. The top and bottom tables report results for a global FPR of 0.1% and 1% respectively.

6.5 Appendix E

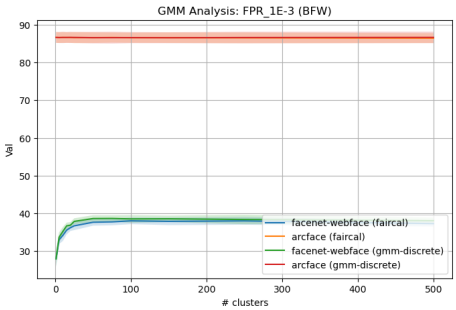
In the original paper, the authors show the robustness of FairCal by showing that its performance does not significantly change for the choice of a wide number amount of clusters. We performed the same analysis and concluded that FairCal-GMM is also robust to differing amounts of clusters, as can be shown in Figure 5.



(a) Comparison of AUROC for different cluster amounts for FairCal and FairCal-GMM



(b) Comparison of FPR @ 1% for different cluster amounts for FairCal and FairCal-GMM



(c) Comparison of FPR @ 0.1% for different cluster amounts for FairCal and FairCal-GMM

Figure 5. Comparison of FairCal and FairCal-GMM on the three accuracy metrics for different cluster amounts. Shaded uncertainties indicate mean ± 1 standard deviation over 5-fold validation.