Bias Mitigation of Face Recognition Models Through Calibration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Face recognition models suffer from bias: for example, the probability of a false positive (incorrect face match) strongly depends on sensitive attributes like ethnicity. As a result, these models may disproportionately and negatively impact minority groups when used in law enforcement. In this work, we introduce the Bias Mitigation Calibration (BMC) method, which (i) increases model accuracy (improving the state-of-the-art), (ii) produces fairly-calibrated probabilities, (iii) significantly reduces the gap in the false positive rates, and (iv) does not require knowledge of the sensitive attribute.

1 Introduction

Face recognition (FR) systems are being increasingly deployed worldwide in a variety of contexts, ranging from policing and border control to providing security for everyday consumer electronics. According to [14], facial images depicting around half of all American adults are searchable in police databases. Some law enforcement agencies can even identify people in real-time using video surveillance. FR systems have been built to maximize overall accuracy [23], achieving impressive results. However, they have also been shown to exhibit significant bias against certain demographic subgroups [4, 31, 2]. For example, many FR systems have much higher false positive rates (FPRs) for non-white faces than for white faces [16]. Therefore, when FR is employed by law enforcement, non-white individuals may be more likely to be falsely detained [1]. Devising means by which to detect and mitigate FR bias, particularly when that bias targets groups that already face disproportionate police scrutiny, is thus of utmost importance.

Bias mitigation in the context of FR models has recently received a great deal of attention. Most efforts have been directed towards learning less biased representations and (hence) decisions [42, 28, 43, 25, 46, 15, 41, 22]. These approaches have enjoyed varying degrees of success: though the bias is reduced, it may still endure [42]. Moreover, they often require retraining the models from scratch, are computationally expensive, and may not always be feasible in practice. For these, and other reasons, an approach which mitigates bias in existing models is valuable.

In this work, we pursue a complementary approach: post-processing bias mitigation. In this case, the model that provides the feature representation remains unchanged (we assume to only have black-box access to it). The goal is to use it to build a fair classifier for the downstream task—in our case the face verification problem, a major subproblem in FR. Given two facial images, we must determine whether they depict the same person. Current state-of-the-art (SOTA) face verification classifiers are based on deep neural networks that embed images into a low dimensional space; a pair of images is deemed a match if the similarity score of their respective embeddings exceeds a particular threshold.

The fairness of face verification classifiers is typically assessed by measuring false positive rates (FPRs) (or false negative rates (FNRs)) across different subgroups, which are defined by a sensitive attribute such as ethnicity, gender, or age. We illustrate the bias of current models in Figure 1. Another

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

notion of fairness across subgroups concerns calibration, which requires a classifier to output an estimate of the true probability (i.e. confidence) that a pair is indeed a genuine match. The classifier is said to be (globally) calibrated if given a set of pairs such that each pair has the predicted probability p of being a match, a p fraction of the pairs truly are matches [17]. From the point of view of fairness, it is crucial for calibration to hold conditionally on each of the subgroups (fairness-calibration, see Definition 2). Otherwise, if the same predicted probability score is known to carry different meanings for different demographic groups, users, including law enforcement officers and judges, may be motivated to take sensitive attributes into account when making critical decisions about arrests or sentencing [34]. While global calibration can be achieved with off-the-shelf post-hoc calibration methods, such methods do not achieve fairness-calibration (see Figure 2).

Achieving all three fairness definitions mentioned above: 1) fairness-calibration, 2) equal FPRs, 3) equal FNRs is, however, impossible, as several authors have shown [20, 6, 24, 34]. We elaborate further on this in section 3, but the key takeaway is that we may aim to satisfy at most two of the three fairness notions. In the particular context of policing, equal FPRs (also known as predictive equality) is considered more important than equal FNRs (equality opportunity), as false positive errors (i.e. false arrests) risk causing significant harm to individuals, a risk that is heightened when those individuals are members of groups already at disproportionate risk for police scrutiny or violence. As a result, our goals are to achieve predictive equality and fair-calibration across subgroups. We further assume that we do not have access to the sensitive attributes, as this may not be feasible in practice due to (i) privacy concerns, (ii) challenges in defining the sensitive attribute (e.g. ethnicity cannot be neatly divided into discrete categories), and (iii) because it may be laborious and expensive to collect the sensitive attribute.

The main contributions of the paper are as follows:

- We introduce a method called **Bias Mitigation Calibration (BMC)**, which achieves state-of-the-art **accuracy** (Table 2) on face verification on two distinct datasets.
- **Fairness-calibration**: Our face verification classifier outputs state-of-the-art fainess-calibrated probabilities without the need for the sensitive attribute (Table 3).
- **Predictive equality:** Our method reduces the gap in FPRs across sensitive attributes (Figure 4). For example, in Figure 2, at a Global FPR of 5% using the baseline method, Black people are 15X more likely to false match than white people. Our method reduces this to 1.2X (while SOTA for post-hoc methods is 1.7X).
- **No sensitive attribute required**: Our approach does not require knowledge of the sensitive attribute, neither at training nor at test time. In fact, it outperforms models that use this knowledge.
- Our method has **no need for retraining** the original feature representation model, since it performs statistical calibration on any given trained model.

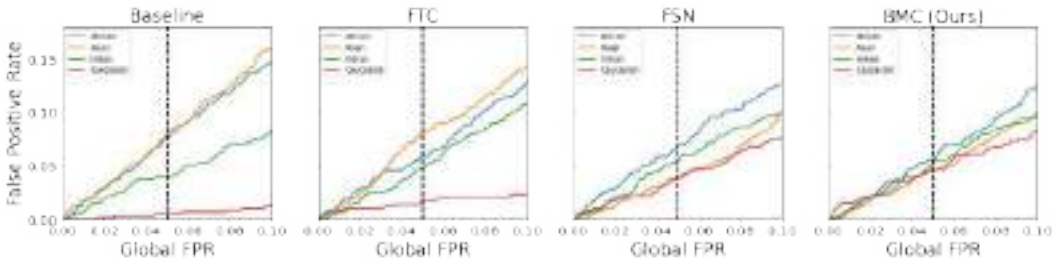

Figure 1: Illustration of bias reduction, as measured by the FPRs evaluated on intra-ethnicity pairs on the RFW dataset with the FaceNet (Webface) feature model. Lines closer together is better, unbiased corresponds to the lines overlapping. The baseline method shows significant subgroup bias, which is reduced with post-processing methods FTC [38], FSN [39], and BMC (ours), BMC method is best.

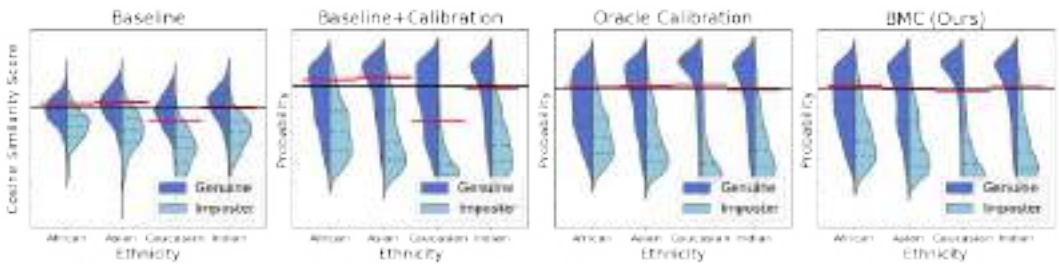

Figure 2: Illustration of bias reduction. False Positives correspond to the imposter pairs above a decision threshold value (y-axis). *Black horizontal line*: threshold which achieves global FPR of 5%; *Red lines*: thresholds which achieve subgroup FPRs of 5%. The deviation of the subgroup FPRs from the global FPR is a measure of bias (smaller is better, red lines closer to black horizontal line is better). The baseline method is biased. Calibration (based on cosine similarity alone) does not reduce the bias of the method. The oracle method reduces bias by using subgroup membership labels for calibration. The BMC method (ours) reduces bias by using feature vectors for calibration.

2 Related work

Bias mitigation. Work on bias mitigation for deep FR models can be divided into two main camps: (i) methods that learn less biased representations, and (ii) post-processing approaches that attempt to remove bias from a pre-trained feature representation model by building fairer decision systems [37, 38, 39]. Several approaches have been pursued in (i), including domain adaptation [42], data augmentation [46, 25], reinforcement learning [41], and adversarial learning [2, 15]. While these methods mitigate bias, this is often achieved at the expense of a strong decrease in recognition accuracy and at high computational cost. The work we present in this paper fits into (ii), and thus we focus on reviewing those approaches. For an in-depth literature review of bias mitigation for FR models see [9].

Srinvas et al. [37] proposed an ensemble approach, exploring different strategies for fusing the scores of multiple models. While this work does not directly measure bias reduction (presenting only the results of applying the method to a protected subgroup), [39] show that while this method is effective in increasing overall accuracy, it fails to reliably mitigate bias.

Terhöst et al. [38] proposed the Fair Template Comparison (FTC) method, which replaces the computation of the cosine similarity score by a shallow neural network trained using cross-entropy loss, with a fairness penalization term and an l_2 penalty term to prevent overfitting. While this method does indeed reduce a model's bias, it comes at the expense of an overall decrease in accuracy. Moreover, it requires tuning parameters related to the shallow neural network and the loss weights. Most importantly, it requires the sensitive attributes during training.

Our proposed method is most closely related to Terhöst et al's [39] Fair Score Normalization (FSN) method, which equalizes a model's FPRs across unsupervised clusters (see subsection 4.1). However, our method differs from FSN in that we convert the cosine similarity scores into calibrated probabilities, and do not rely on a predefined FPR.

Calibration. Calibration is closely related to uncertainty estimation for deep networks [17]. Several post-hoc calibration methods have been proposed such as Platt's scaling or temperature scaling [33, 17], histogram binning [47], isotonic regression [29], spline calibration [19], and beta calibration [27], among others. All of these methods involve computing a calibration function that maps the softmax outputs of a neural network to confidence estimates, i.e. to estimates of the probability of truly belonging to the predicted class. As such, any calibration method can readily be applied to cosine similarity scores, leading to models that are calibrated but not fairly-calibrated. An algorithm to achieve the latter for a binary classifier has been proposed in [21], but the work remains theoretical and no practical implementation is known.

Closely related to face verification is the problem of face identification, the goal of which is to determine the identity of a person in a test image from a set of known people. In this context, [10, 26] proposed different confidence measures based both on a classifier's output and on extracted face features. Recently, Eliades et al. [11] investigated the application of conformal prediction, although

their study, like the ones in [10, 26], did not tackle deep neural networks and did not include any analysis concerning bias. Finally, [44] proposed the Predictive Confidence Network (PCNet), which is used to reject low quality image pairs, thereby leading to reduction in false positives; this work likewise does not address bias.

Table 1: Comparison of desirable features of the different post-training bias mitigation methods.

Method	Improves accuracy over Baseline	Predictive equality (equal FPRs)	Fairly-calibrated	Does not require sensitive attribute during classifier training	Does not require sensitive attribute at test time
FTC [38]	✘	✔	✘	✘	✔
FSN [39]	✔	✔	✘	✔	✔
Oracle (Ours)	✔	✔	✔	✘	✘
BMC (ours)	✔	✔	✔	✔	✔

3 Face Verification and Fairness

In order to discuss bias mitigation strategies and their effectiveness, one must first agree on what constitutes a fair algorithm. We start by rigorously defining the face verification problem as a binary classification problem. Then we present the notion of a probabilistic classifier, which outputs calibrated probabilities that the classification is correct. Finally, since several different definitions of fairness have been proposed (see [40, 13] for a comparative analysis), we review the ones pertinent to our work.

3.1 Score-based and binary classifiers

Let f denote a trained neural network that encodes an image $\boldsymbol{x}$ into an embedding $f(\boldsymbol{x}) = \boldsymbol{z} \in \mathcal{Z} = \mathbb{R}^d$. Pairs of images of faces are drawn from a global pair distribution $(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{P}$. Given such a pair, let $Y(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1$ if the identities of the two images are the same and let $Y = 0$ otherwise. The face verification problem consists of learning a binary classifier for Y.

The baseline classifier for the face verification problem is based on the cosine similarity between the feature embeddings of the two images, $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{f(\boldsymbol{x}_1)^T f(\boldsymbol{x}_2)}{\|f(\boldsymbol{x}_1)\| \|f(\boldsymbol{x}_2)\|}$. The cosine similarity score is used to define a binary classifier $\widehat{Y} : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ by thresholding with a choice of $s_{\text{thr}} \in [-1, 1]$,

$$\begin{cases} \widehat{Y}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 & \text{if } s(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq s_{\text{thr}}, \\ \widehat{Y}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0 & \text{otherwise} \end{cases}$$

The accuracy of the algorithm, as with any binary classification problem, is measured by evaluating its false positive rate (FPR) and/or false negative rate (FNR). In the models we consider, the threshold s_{thr} is determined by a target FPR.

3.2 Calibration

Any score-based classifier can be converted to a calibrated probabilistic model using standard post-hoc calibration methods [47, 29, 33, 27], see Appendix D. A calibrated probabilistic model is *interpretable*: the output reflects the true probability of genuine match, in the sense of the following definition.

Definition 1. *The probabilistic model $\widehat{C} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is said to be **calibrated** if the probability of a match, conditioned on the output of the model being c, is equal to c,*

$$\mathbb{P}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{P}}(Y = 1 \mid \widehat{C} = c) = c.$$

The probabilistic model can be converted to a binary classifier by using the output probability as a score: thresholding at a given probability.

Applied to face verification, a calibrated probabilistic model outputs the likelihood/confidence that a pair of images $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are a match. Calibration is achieved by using a calibration set S^{cal} to learn a calibration map μ from the cosine similarity scores s to probabilities Our calibration methods define an increasing calibration map μ, which means that the binary classifier obtained by thresholding the score based model at s is equivalent to the classifier obtained by thresholding the probabilistic classifier at $c = \mu(s)$.

3.3 Fairness

In the context of face verification, fairness implies that the classifier has similar performance on the different population subgroups. For calibration, this means being calibrated conditional on subgroup membership.

Let $g(\boldsymbol{x}) \in G = \{g_i\}$ denote the sensitive/protected attribute of $\boldsymbol{x}$, such as ethnicity, gender, or age. Each sensitive attribute $g_i \in G$ induces a population subgroup, whose distribution on the intra-subgroup pairs we denote by $\mathcal{G}_i$.

Definition 2. *The probabilistic model $\widehat{C}$ is **fairly-calibrated**[1] for subgroups g_1 and g_2 if the classifier is calibrated when conditioned on each subgroup.*

$$\mathbb{P}_{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{G}_1}(Y = 1 \mid \widehat{C} = c) = \mathbb{P}_{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{G}_2}(Y = 1 \mid \widehat{C} = c) = c.$$

Intuitively, a model is considered biased if its accuracy alters when tested on different subgroups. In the case of face recognition applications, *predictive equality*, meaning equal FPRs across subgroups, is often of primary importance.

Definition 3. *A binary classifier $\widehat{Y}$ exhibits **predictive equality** for subgroups g_1 and g_2 if the classifier has equal FPRs for each subgroup,*

$$\mathbb{P}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim\mathcal{G}_1}(\widehat{Y} = 1 \mid Y = 0) = \mathbb{P}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim\mathcal{G}_2}(\widehat{Y} = 1 \mid Y = 0).$$

In certain applications of FR (such as office security, where high FNRs would cause disruption), equal opportunity, meaning equal FNRs across subgroups, is also important.

Definition 4. *A binary classifier $\widehat{Y}$ exhibits **equal opportunity** for subgroups g_1 and g_2 if the classifier has equal FNRs for each subgroup,*

$$\mathbb{P}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim\mathcal{G}_1}(\widehat{Y} = 0 \mid Y = 1) = \mathbb{P}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim\mathcal{G}_2}(\widehat{Y} = 0 \mid Y = 1).$$

Prior works [20, 6, 24, 34] have shown that it is impossible in practice to satisfy all three fairness properties at the same time: 1) fairness calibration (definition 2), 2) predictive equality (definition 3), 3) equal opportunity (definition 4). At most, two of these three desirable properties can be achieved simultaneously. In the context of our application, predictive equality is more critical than equal opportunity. In practice, models will not perfectly satisfy these definitions, since they require strict equality. However, these qualitative concepts are useful for discussing quantitative gaps in fairness or calibration.

4 Fairness Calibration Methods

In this section, we describe the Bias Mitigation Calibration (BMC) method, which constitutes the main contribution of this work.

Simply calibrating the method based on cosine similarity scores will not improve fairness: if the baseline score-based classifier is biased, the resulting probabilistic classifier remains equally biased. This is illustrated in Figure 2, where for both Baseline and Baseline Calibration any choice of global threshold will lead to different FPRs across sensitive subgroups; consequently, these models fail to achieve predictive equality.

Our proposed solution is to introduce *conditional* calibration methods, which involve partitioning the pairs into sets and calibrating each set. Given an image pair, we identify its set membership and apply the corresponding calibration map. For our BMC method, the sets are formed by clustering the images' feature vectors. In the case of our Oracle method, the sets are given by the sensitive attributes. By using more information, both methods are able to achieve fairness-calibration.

The methods we discuss below are as follows: (i) the Fair Score Normalization (FSN) method [39], which adjusts scores based on features but is not fairly-calibrated, (ii) our Bias Mitigation Calibration (BMC) method, which is designed to be fairly-calibrated, and (iii) our fairly-calibrated Oracle Calibration method, which requires knowledge of the images' sensitive attributes. Throughout this section, we let S^{cal} denote the cosine similarity scores of the pairs of images in our calibration dataset and $\mathcal{Z}^{\text{cal}} \subseteq \mathcal{Z}$ be the set of feature embeddings of those images, i.e. the outputs of a pretrained model f.

[1] This notion is also referred to as well-calibration [40] and calibration within subgroups [24, 3]

4.1 Fair Score Normalization (FSN) (prior method)

The Fair Score Normalization (FSN) method [39] aims to correct for the bias present in images' cosine similarity scores by performing feature-based score normalization via the following three steps:

(i) Cluster the image embeddings by applying the K-means algorithm to $\mathcal{Z}^{\text{cal}}$, partitioning the embedding space $\mathcal{Z}$ into K clusters $\mathcal{Z}_1, \ldots, \mathcal{Z}_K$. For each cluster $k \in \{1, \ldots, K\}$, associate a set S_k^{cal} formed by the pairs of images that contain at least one image belonging to $\mathcal{Z}_k$. Notice that an image pair may belong to one or two clusters.

(ii) For each $k \in \{1, \ldots, K\}$, using the scores of image pairs in S_k^{cal}, compute an optimal threshold s_{thr}^k for a pre-defined FPR of 0.1%. Additionally, calculate the threshold for the whole calibration set S^{cal}, i.e. s_{thr}^G.

(iii) Assuming that k_1 and k_2 denote the clusters for an image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (where k_1 may be equal to k_2), the normalized cosine similarity score $\hat{s}$ for this image pair is given by:

$$\hat{s}(\boldsymbol{x}_1, \boldsymbol{x}_2) = s(\boldsymbol{x}_1, \boldsymbol{x}_2) + s_{\text{thr}}^G - \frac{1}{2}\left(s_{\text{thr}}^{k_1} + s_{\text{thr}}^{k_2}\right), \tag{1}$$

where $s(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denotes the cosine similarity score between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Thus, by accounting for the differences in FPRs across the different cluster sets S^{cal}, FSN manages to equalize FPRs across different real subgroups in datasets such as RFW [42] and BFW [35].

4.2 Bias Mitigation Calibration (BMC, our method)

Our method performs score-based calibration on clusters of images. Like in the FSN method, we cluster the feature vectors into K unsupervised clusters (our step (i) is the same). The difference between our methods lies in the next two steps where we convert our model's cosine similarity scores into cluster-specific probabilities by performing calibration on each cluster:

(i) Apply the K-means algorithm to the image features, $\mathcal{Z}^{\text{cal}}$, partitioning the embedding space $\mathcal{Z}$ into K clusters $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$. These form the K calibration sets:

$$S_k^{\text{cal}} = \{s(\boldsymbol{x}_1, \boldsymbol{x}_2) : f(\boldsymbol{x}_1) \in \mathcal{Z}_k \text{ or } f(\boldsymbol{x}_2) \in \mathcal{Z}_k, \}, \quad k = 1, \ldots, K$$

(ii) For each calibration set S_k^{cal}, use a post-hoc calibration method (we use beta calibration [27], since it is recommended when dealing with bounded scores) to compute the calibration map μ^k, which maps scores $s(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to cluster-conditional probabilities $\mu^k(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$.

(iii) For an image pair ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$), if both images fall into the same cluster k, define the cluster calibrated score as:

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mu^k(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$$

Else, if the images are in different clusters, k_1 and k_2, respectively, define the pair's cluster calibrated score as the weighted average of the calibrated scores in each cluster:

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \theta\mu^{k_1}(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) + (1-\theta)\mu^{k_2}(s(\boldsymbol{x}_1, \boldsymbol{x}_2)),$$

where the weight θ is the relative population fraction of the two clusters,

$$\theta = \left|S_{k^1}^{\text{cal}}\right| / (\left|S_{k^1}^{\text{cal}}\right| + \left|S_{k^2}^{\text{cal}}\right|)$$

4.3 Oracle Calibration (our method : supervised BMC)

We include a second calibration method, Oracle, which performs BMC calibration as above but creates the calibration sets based on the sensitive attribute, instead of in an unsupervised fashion. If the images belong to different subgroups, $g(\boldsymbol{x}_1) \neq g(\boldsymbol{x}_2)$, then the classifier correctly outputs zero. This method is not feasible in practice, since the sensitive attribute may not be available, or because using the sensitive attribute may be not permitted for reasons of discrimination or privacy. However, the Oracle method represents a benchmark for our calibration if the sensitive attributes were known and could be used.

5 Experimental Details

Models: We used three distinct pretrained models: two Inception Resnet models obtained from [12] (MIT License), one trained on the VGGFace2 dataset [5] and another on the CASIA-Webface dataset [45], and an ArcFace model obtained from [36] (Apache 2.0 license) and trained on the refined version of MS-Celeb-1M [18]. We will refer to the models as Facenet (VGGFace2), Facenet (Webface), and ArcFace, respectively. As is standard for face recognition models, we pre-processed the images by cropping them using a Multi-Task Convolution Neural Network (MTCNN) algorithm [48]. If the algorithm failed to identify a face, the pair was removed from the analysis.

Datasets: We present experiments on two different datasets: Racial Faces in the Wild (RFW) [42] and Balanced Faces in the Wild (BFW) [35], both of which are available under licenses for non-commercial research purposes only. Both datasets already include predefined pairs separated into five folds. The results we present are the product of leave-one-out cross-validation. The RFW dataset contains a 1:1 ratio of genuine/imposter pairs and 23,541 pairs in total (after applying the MTCNN). The dataset's images are labeled by ethnicity (African, Asian, Caucasian, or Indian), with all pairs consisting of same-ethnicity images. The BFW dataset, which possesses a 1:3 ratio of genuine/imposter pairs, is comprised of 890,347 pairs (after applying the MTCNN). Its images are labeled by ethnicity (African, Asian, Caucasian, or Indian) and gender (Female or Male), and it includes mixed-gender and mixed-ethnicity pairs. The RFW and BFW datasets are made up of images taken from MS-Celeb-1M [18] and VGGFace [5], respectively. Thus, to expose the models to new images, only the two FaceNet models can be evaluated on the RFW dataset, while only the FaceNet (Webface) and ArcFace models can be evaluated on the BFW dataset.

Methods: For both the FSN and BMC method we used $K = 100$ clusters for the K-means algorithm, as recommended by [39]. For BMC we employed the recently proposed beta calibration method [27]. Our BMC method is robust to the choice of number of clusters and post-hoc calibration method (see Appendix E for more details). We discuss the parameters used in training FTC [38] in Appendix A. Our Oracle method is designed to be fairly-calibrated, but as it uses subgroup labels, it is not a feasible calibration method in practice. Thus, it can be used as an ideal baseline for fairness-calibration. Notice that the prior relevant methods (Baseline, FTC, and FSN) output scores that, even when rescaled to [0,1], do not result in calibrated probabilities and hence, by design, cannot be fairly-calibrated. Therefore, in order to fully demonstrate that our method is superior to those approaches, when measuring fairness-calibration we apply beta calibration (the same post-hoc calibration method used in our BMC method) to their final score outputs.

Implementation details: The embeddings from the pretrained models were obtained on a machine with one GeForce GTX 1080 Ti GPU. All methods were implemented in Python, and the code is provided in the supplemental material.

6 Results

In this section we report the performance of our models with respect to the three metrics: (i) accuracy, (ii) fairness calibration, and (iii) predictive equality. Our results show that among post hoc calibration methods,

1. BMC is best at global **accuracy**, see Table 2
2. BMC is best on **fairness calibration**, see Table 3.
3. BMC is best on **predictive equality**, i.e. equal FPRs, see Table 4.

We discuss these results in further detail below and provide additional detailed result in the Appendix. Overall, we are interested in a method that satisfies both fairness definitions without decreasing baseline model accuracy. For example, while the FTC method obtains slightly better predictive equality results than our method in one situation, this is achieved only at the expense of a significant decrease in accuracy.

6.1 Global accuracy

The receiver operating characteristic (ROC) curve plots the true positive rate (TPR, equivalent to $1-$FNR) against the FPR, obtained by thresholding the model at different values.

Table 2: Global **accuracy** measured by the AUROC and the TPR at different FPR thresholds.

	RFW						BFW					
	FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
(↑)	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR
Baseline	88.26	18.42	34.88	83.95	11.18	26.04	96.06	33.61	58.87	97.41	86.27	90.11
FTC	86.46	6.86	23.66	81.61	4.65	18.40	93.30	13.60	43.09	96.41	82.09	88.24
FSN [39]	90.05	23.01	40.21	85.84	17.33	32.80	96.77	**47.11**	68.92	97.35	86.19	90.06
BMC (Ours)	**90.58**	**23.55**	**41.88**	**86.71**	**20.64**	**33.13**	**96.9**	46.74	**69.21**	**97.44**	**86.28**	**90.14**
Oracle (Ours)	*89.74*	*21.4*	*41.83*	*85.23*	*16.71*	*31.6*	*97.28*	*45.13*	*67.56*	*98.91*	*86.41*	90.40

Table 3: **Fairness calibration:** measured by the mean KS across the sensitive subgroups. **Bias:** measured by the deviations of KS across subgroups in terms of three deviation measures: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) (lower is better).

	RFW								BFW							
	FaceNet (VGGFace2)				FaceNet (Webface)				FaceNet (Webface)				ArcFace			
(↓)	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD
Baseline	6.37	2.89	5.73	3.77	5.55	2.48	4.97	2.91	6.77	3.63	5.96	4.03	2.57	1.39	2.94	1.63
FTC [38]	5.16	2.31	4.44	2.87	4.11	1.87	3.74	2.20	6.60	2.42	5.19	2.93	3.70	1.50	3.07	1.78
FSN [39]	1.43	0.35	0.57	0.40	2.49	0.84	1.19	0.91	**2.76**	1.38	2.67	1.60	2.65	1.45	3.23	1.71
BMC (Ours)	**1.37**	**0.28**	**0.5**	**0.34**	**1.75**	**0.41**	**0.64**	**0.45**	3.09	**1.34**	**2.48**	**1.55**	**2.49**	**1.3**	**2.68**	**1.52**
Oracle (Ours)	*1.18*	*0.28*	*0.53*	*0.33*	*1.35*	*0.38*	*0.66*	*0.43*	*2.23*	*1.15*	*2.63*	*1.4*	*1.41*	*0.59*	*1.3*	0.69

Table 4: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, compare the deviations in subgroup FPRs in terms of three deviation measures: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) (lower is better).

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
		AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FPR	Baseline	0.10	0.15	**0.10**	0.14	0.26	0.16	0.29	1.00	0.40	0.12	0.30	0.15
	FTC	0.10	0.15	0.11	0.12	0.23	0.14	0.24	0.74	0.32	**0.09**	**0.20**	**0.11**
	FSN [39]	0.10	0.18	0.11	0.11	0.23	0.13	**0.09**	**0.20**	**0.11**	0.11	0.28	0.14
	BMC (Ours)	**0.09**	**0.14**	**0.10**	**0.09**	**0.16**	**0.1**	**0.09**	**0.20**	**0.11**	0.11	0.31	0.15
	Oracle (Ours)	*0.11*	*0.19*	*0.12*	*0.11*	*0.2*	*0.13*	*0.12*	*0.25*	*0.15*	*0.12*	*0.27*	*0.14*
1% FPR	Baseline	0.68	1.02	0.74	0.67	1.23	0.79	2.42	7.48	3.22	0.72	1.51	0.85
	FTC	0.60	0.91	0.66	0.54	1.05	0.66	1.94	5.74	2.57	**0.54**	**1.04**	**0.61**
	FSN [39]	0.37	0.68	0.46	0.35	0.61	0.40	0.87	2.19	1.05	0.55	1.27	0.68
	BMC (Ours)	**0.28**	**0.46**	**0.32**	**0.29**	**0.57**	**0.35**	**0.8**	**1.79**	**0.95**	0.63	1.46	0.78
	Oracle (Ours)	*0.4*	*0.69*	*0.45*	*0.41*	*0.74*	*0.48*	*0.77*	*1.71*	*0.91*	*0.83*	*2.08*	*1.07*

The area under the ROC curve (AUROC) is thus a holistic metric that summarizes the accuracy of the classifiers [7]. A higher AUROC is better, while an uninformative classifier has an AUROC of 50%.

Table 2 shows the AUROC for the different pre-trained models and different datasets, as well as the TPRs at 0.1% and 1% global FPR thresholds. Our BMC method achieves SOTA results, with the best values of AUROC in all cases, and with the highest TPR in seven of the eight cases. In addition, our BMC method surpasses our Oracle method (which uses subgroup information) on the RFW dataset.

This overall accuracy improvement of our BMC method compared to the baseline method can be explained as follows: the feature vectors contain more information than the score alone, so they can be used to identify pairs where the probability of an error (at a given similarity score) is higher or lower, allowing per-cluster thresholds to give better accuracy.

6.2 Fairness calibration

The prior relevant methods (Baseline, FTC, and FSN) do not output probabilities, hence by design they cannot be fairly-calibrated. However, in order to fully demonstrate that our method is superior to those approaches, we apply beta calibration (the same post-hoc calibration method used in our BMC method) to their score outputs.

For a model that outputs probabilities that image pairs are matches, the calibration error is a measure of the deviation in these predictions from the actual results. There are two quantities of interest when

considering subgroups: (i) the calibration error on each subgroup (smaller is better, because it means fewer errors) and (ii) the deviation in these quantities across subgroups (smaller is better, because it is more fair). There are several measures of calibration error: ECE, KS, and BS, which are discussed in Appendix B. We present results for KS, which has been established to be the best measure [19]. We report three measures of deviation: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD). Recall that, to obtain a fair comparison, we applied beta calibration (the same post-hoc calibration method used in our BMC method) to the final scores outputted by the Baseline, FTC, and FSN methods.

The results are summarized in Table 3. Our BMC method achieves the best results: the best mean in three of the four cases (smaller errors) and the best deviation in nine of the nine cases (more fair across subgroups). The improvement is most significant on the less accurate models. We discuss these results in more detail in the appendix.

6.3 Predictive equality

As predictive equality is achieved through equal FPRs between different subgroups, we can measure bias by quantifying the deviation of these FPRs. We report three measures of deviation (AAD, MAD, and STD) at two different global FPRs: 0.1% and at 1.0%.

The results are summarized in Table 4. Our BMC method achieves the best results: the best or tied measure of predictive equality in 18 of the 24 cases. In the cases where BMC is not best, i.e. when the ArcFace model evaluated on the BFW dataset, FTC provides the best results, but the differences between FTC's, FSN's, and BMC's deviation measures are within a fraction of 1%. Moreover, when applied to ArcFace, the FTC method reduces the model's accuracy: at FPRs of 0.1% and 1%, the TPRs are, respectively, 4% and 2% lower.

7 Conclusion

Societal Impact: Bolstered by dramatically improving accuracy, Facial Recognition (FR) systems have exploded in popularity in recent years and have been applied to innumerable settings and use cases. Undermining their success are severe biases typically against minority populations.

In this work, we address this issue by proposing a post-processing method that significantly reduces the false positive rates across demographic subgroups without requiring knowledge of the sensitive attributes in question.

Limitations: Only under very strict conditions (e.g. a perfect classifier) can all all three fairness notions be satisfied. While we have shown that our BMC method can achieve two out of the three definitions, which is the best we can hope for in practice, it requires the use of a calibration dataset. Deploying our method on a substantially different dataset would most likely require further calibration. Calibrating models for never-seen data is currently an open problem.

Advantages: Our method (BMC) achieves SOTA accuracy and is fairly-calibrated on challenging face verification datasets without the need for the sensitive attribute, i.e. it outputs true confidence estimates that a pair of images is indeed a genuine match independent of the subgroup identity. BMC can be readily applied to existing FR systems, as it is a statistical method that does not require retraining. It can aid human-users in using FR system to make better informed decisions based on interpretable, unbiased results.

References

[1] Bobby Allyn. 'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man. https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig, June 2020.

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[7] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. Association for Computing Machinery.

[8] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[9] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[10] S. Eickeler, M. Jabs, and G. Rigoll. Comparison of confidence measures for face recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE Comput. Soc, 2000.

[11] Charalambos Eliades, Ladislav Lenc, Pavel Král, and Harris Papadopoulos. Automatic face recognition with well-calibrated confidence measures. *Machine Learning*, 108(3):511–534, 2019.

[12] Tim Esler. Face recognition using pytorch. `https://github.com/timesler/facenet-pytorch`, 2021.

[13] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666, 2020.

[14] C. Garvie, Georgetown University. Center on Privacy, Technology, and Georgetown University. Law Center. Center on Privacy & Technology. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[15] Sixue Gong, Xiaoming Liu, and A Jain. Jointly de-biasing face recognition and demographic attribute estimation. ECCV, 2020.

[16] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test part 3:. Technical report, National Institute of Standards and Technology, December 2019.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1321–1330, 2017.

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[19] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

[20] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323. Curran Associates, Inc., 2016.

[21] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018.

[22] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2781–2794, 2020.

[23] Anil K. Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany*, 2017.

[25] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas 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)*, pages 2261–2268, 2019.

[26] Pavel Kral and Ladislav Lenc. Confidence measure for experimental automatic face recognition system. In *Lecture Notes in Computer Science*, pages 362–378. Springer International Publishing, 2015.

[27] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[28] Jian Liang, Yuren Cao, Chenbin Zhang, Shiyu Chang, Kun Bai, and Zenglin Xu. Additive adversarial learning for unbiased authentication. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11420–11429, 2019.

[29] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. Association for Computing Machinery.

[30] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

[31] M. Orcutt. Are face recognition systems accurate? depends on your race. In *MIT Technology Review*. 2016.

[32] Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. In *International Conference on Learning Representations*, 2021.

[33] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[34] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5680–5689. Curran Associates, Inc., 2017.

[35] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6 2020.

[36] Abhinav Sharma. ONNX Model Zoo. `https://github.com/onnx/models`, 2021.

[37] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King. Face recognition algorithm bias: Performance differences on images of children and adults. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2269–2277, 2019.

[38] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[39] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332 – 338, 2020.

[40] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[41] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in-the-wild: Reducing racial bias by information maximization adaptation network, 2019.

[43] Pingyu Wang, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 363:35–45, 2019.

[44] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face verification. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[45] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.

[46] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[47] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 10 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include instructions in the supplemental material. The datasets' licenses do not allow them to be shared.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See section 5 and Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We omit error bars for clarity, but we report in the Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Crowdsourcing was not used in this work.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Crowdsourcing was not used in this work.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Crowdsourcing was not used in this work.

Supplemental Material

A Fair Template Comparison (FTC) method

The Fair Template Comparison (FTC) method [37] learns a shallow network with the goal of outputting fairer decisions. We implemented the FTC method as follow. In order to keep the ratios between the dimensions of layers the same as in the original paper [37], we used a 512-dimensional input layer, followed by two 2048-dimensional intermediate layers. The final layer is a fully connected linear layer with 2-dimensional output with a softmax activation. All intermediate layers are followed by a ReLU activation function and dropout (with $p = 0.3$). The network was trained with a batchsize of $b = 200$ over 50 epochs, using an Adam optimizer with a learning rate of 10^{-3} and weight decay of 10^-4. Two losses, one based on subgroup fairness and the other on both subgroup and individual fairness, were proposed in [37]. Based on the paper's recommendations, we used the individual fairness loss with a trade-off parameter of $\lambda = 0.5$.

B Measuring Calibration Error

There are different metrics available to measure if a probabilistic classifier is calibrated or fairly-calibrated. Calibration error is the error between the true and estimated confidences and is typically measured by the Expected Calibration Error (ECE) [16]:

Despite being the most popular calibration error metric, the ECE has several weaknesses, chief among which is its dependence on the binning scheme [29]. Recently, Gupta et al. [18] introduced a simple, bin-free calibration measure. For calibrated scores $P(Y = 1|C = c) = c$ we have, by Bayes' rule:

$$P(Y = 1, C = c) = cP(C = c).$$

Inspired by the Kolmogorov-Smirnov (KS$\downarrow$) statistic test, Gupta et al proposed to measure the calibration error by comparing the cumulative distributions of $P(Y = 1, C = c)$ and $cP(C = c)$, which empirically correspond to computing the sequences

$$h_i = h_{i-1} + \mathbf{1}_{y_i=1}/N \quad \text{and} \quad \tilde{h}_i = \tilde{h}_{i-1} + c_i/N$$

with $h_0 = \tilde{h}_0 = 0$, and N is the total number of samples. Then the KS calibration error metric is given by

$$KS = \max_i \left| h_i - \tilde{h}_i \right|.$$

C Fairness calibration and Equal Opportunity (equal FNR)

C.1 Fairness calibration

Since the calibration map produced by beta calibration is monotone, the ordering of the images provided by the scores is the same as the ordering provided by the probabilities; therefore, the accuracy of the methods wheh thresholding remains unchanged. The calibration error (CE) measured with an adaptation of the Kolmogorov-Smirnov (KS) test (described in the Appendix) is computed for each subgroup of interest. Notice that for the BFW dataset we consider the eight subgroups that result from the intersection of the ethnicity and gender subgroups.

We first observe that all methods are equally globally calibrated (i.e. the calibration error is low) after the post-hoc calibration method is applied, except for the FTC on the RFW dataset (see the Global column in Table 5 and Table 6).

By inspecting Table 5 and Table 6, we notice that, after calibration, the Baseline method results in models that are not fairly-calibrated, though perhaps not in the way one would expect. Typically, bias is directed against minority groups, but in this case, it is the Caucasian subgroups that have the higher CEs. This is a consequence of the models' above average accuracy on this subgroup, which is underestimated and therefore not captured by the calibration procedure. It is important to point out that this is not a failure of the calibration procedure, since the global CE (i.e. the CE measured on all pairs) is low, as discussed above.

Table 5: KS on all the pairs (Global (Gl)) and on each ethnicity subgroup (African (Af), Asian (As), Caucasian (Ca), Indian (In)) using beta calibration on the RFW dataset.

	FaceNet (VGGFace2)					FaceNet (Webface)				
(↓)	Gl	Af	As	Ca	In	Gl	Af	As	Ca	In
Baseline	0.78	6.16	5.74	12.06	1.53	0.69	3.89	4.34	10.52	3.46
FTC [37]	1.12	5.13	5.41	10.19	2.02	1.25	3.19	3.81	8.59	3.35
FSN [38]	0.77	1.27	1.62	1.49	1.35	0.85	1.94	3.06	1.70	3.27
BMC (Ours)	0.81	1.11	1.41	1.29	1.70	0.70	1.48	1.62	1.68	2.21
Oracle (Ours)	*0.76*	*0.99*	*1.28*	*1.2*	*1.25*	*0.62*	*1.54*	*1.46*	*1.13*	1.25

Table 6: KS on all the pairs (Global (Gl)) and on each ethnicity and gender subgroup (African Females (AfF), African Males (AfM), Asian Females (AsF), Asian Males (AsM), Caucasian Females (CF), Caucasian Males (CM), Indian Females (IF), Indian Males (IM)) using beta calibration on the BFW dataset.

	FaceNet (Webface)									ArcFace								
(↓)	Gl	AfF	AfM	AsF	AsM	CF	CM	IF	IM	Gl	AfF	AfM	AsF	AsM	CF	CM	IF	IM
Baseline	0.48	5.00	2.17	11.19	2.93	12.06	10.41	5.58	4.80	0.37	1.52	3.17	5.30	4.28	1.31	1.10	2.09	1.81
FTC [37]	0.56	7.33	4.06	5.71	3.68	12.25	10.47	4.13	5.51	0.49	2.02	3.56	5.77	4.62	1.80	1.03	2.65	2.15
FSN [38]	0.39	2.35	3.12	4.16	4.40	1.50	0.99	3.54	2.02	0.38	1.74	3.01	5.70	4.30	1.02	1.15	2.45	1.81
BMC (Ours)	0.59	3.83	2.55	2.92	3.79	3.70	2.43	3.21	2.32	0.49	1.73	3.12	4.79	3.81	1.05	1.16	2.28	1.97
Oracle (Ours)	*0.43*	*1.67*	*2.3*	*2.83*	*2.49*	*0.67*	*1.24*	*4.6*	*2.02*	*0.32*	*1.26*	*0.99*	*1.93*	*1.72*	*0.86*	*1.15*	*1.64*	1.74

C.2 Equal Opportunity

While equal opportunity (equal FNRs between subgroups) is not prioritized for the FR systems when used by law enforcement, it may be prioritized in different contexts such as office building security. Empirically, our method also mitigates the equal opportunity bias at low global FNRs, as can be seen in Table 7.

D Standard Post-hoc Calibration Methods

For completeness, we provide a brief description of the post-hoc calibration methods used in this work. Beta calibration [26] was used to obtain our main results, but we show below that choosing another method (histogram binning [46], isotonic regression [46, 28]) does not impact the performance of our BMC method.

Table 7: Equal opportunity (equal FNR) at different global FNRs across across the sensitive subgroups.

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	(↓)	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FNR	Baseline	0.09	0.13	0.10	0.10	0.16	0.11	0.09	0.23	0.11	0.11	0.31	0.14
	FTC	0.09	**0.11**	**0.09**	**0.08**	0.14	**0.1**	**0.04**	**0.09**	**0.05**	**0.06**	**0.14**	**0.07**
	FSN [38]	**0.09**	0.13	0.09	0.09	**0.14**	0.10	0.07	0.22	0.10	0.12	0.33	0.15
	BMC (Ours)	0.10	0.14	0.10	0.11	0.17	0.12	0.10	0.27	0.13	0.09	0.17	0.10
	Oracle (Ours)	*0.11*	*0.18*	*0.12*	*0.12*	*0.21*	*0.13*	*0.09*	*0.24*	*0.11*	*0.11*	*0.32*	0.14
1% FNR	Baseline	0.60	0.96	0.67	0.45	0.81	0.53	0.39	0.84	0.47	0.75	1.85	0.93
	FTC	0.48	0.83	0.56	**0.32**	**0.58**	**0.38**	**0.3**	**0.62**	**0.34**	**0.49**	**1.12**	**0.6**
	FSN [38]	**0.28**	**0.47**	**0.32**	0.40	0.78	0.48	0.41	0.92	0.49	0.77	1.91	0.96
	BMC (Ours)	0.30	0.51	0.34	0.39	0.72	0.48	0.32	0.74	0.40	0.65	1.48	0.80
	Oracle (Ours)	*0.38*	*0.61*	*0.42*	*0.56*	*1.06*	*0.67*	*0.37*	*0.77*	*0.44*	*0.5*	*1.11*	0.60

D.1 Histogram Binning

In histogram binning [46], we partition S^{cal} into m bins B_i, where $i = 1, \dots, m$. Then, given a pair of images $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with score $s(\boldsymbol{x}_1, \boldsymbol{x}_2) \in B_i$, we define

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{|B_i|} \sum_{\substack{s(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \in B_i \\ (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{P}^{\text{cal}}}} \mathbf{1}_{I(\hat{\boldsymbol{x}}_1) = I(\hat{\boldsymbol{x}}_2)} \tag{2}$$

In other words, we simply count the number of scores in each bin that correspond to genuine pairs of images, i.e., images that belong to the same person. By construction, a confidence score c (Equation 2) satisfies the binned version of the standard calibration definition (1). As for the bins, they can be chosen so as to have equal mass or to be equally spaced, or else by maximizing mutual information, as recently proposed in [31]. In this work, we created bins with equal mass.

Despite being an extremely computationally efficient method and providing good calibration, histogram binning is not guaranteed to preserve the monotonicity between scores and confidences, which is typically a desired property. Monotonicity ensures that the accuracy of the classifier is the same when thresholding either the scores or the calibrated confidences.

D.2 Isotonic Regression

Isotonic Regression [46, 28] learns a monotonic function $\mu : \mathbb{R} \to \mathbb{R}$ by solving

$$\arg\min_{\mu} \frac{1}{|\mathcal{P}^{\text{cal}}|} \sum_{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{P}^{\text{cal}}} \left(\mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) - \mathbf{1}_{I(\hat{\boldsymbol{x}}_1) = I(\hat{\boldsymbol{x}}_2)}\right)^2$$

The confidence score is then given by $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$.

D.3 Beta Calibration

Beta calibration [31] is a parametric calibration method, which learns a calibration map $\mu : \mathbb{R} \to \mathbb{R}$ of the form

$$c_\theta(s) = \mu(s; \theta_1, \theta_2, \theta_3) = \frac{1}{1 + 1/\left(e^{\theta_3} \frac{s^{\theta_1}}{(1-s)^{\theta_2}}\right)}$$

where the parameters $\theta_1, \theta_2, \theta_3 \in \mathbb{R}$ are chosen by minimizing the log-loss function

$$LL(c, y) = y(-\log(c)) + (1 - y)(-\log(1 - c))$$

where $c = \mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$. By restricting, a and b to be positive, the calibration map is monotone.

E Robustness of BMC results to parameters

In this section we show that the results presented in the main paper still hold if we vary model hyperparameters, such as the number K of clusters used in BMC, and the calibration method.

Choice of post-hoc calibration: The implementation of the BMC method requires choosing a post-hoc calibration method and the number of clusters K in the K-means algorithm. Our method is robust to the choice of both with respect to fairness-calibration (the metric of interest when it comes to calibration) and its bias as depicted in Figure 3, Figure 4, Figure 5 and Figure 6. We compare with binning and isotonic regression. For the former, we chose 10 and 25 bins for the RFW and BFW datasets, respectively, given the different number of pairs in each dataset.

Comparison to FSN [38]: The improved performance of BMC over FSN is consistent across different choices of K. Fixing the choice of the post-hoc calibration method as beta calibration as in the results in the paper, we compare the two, together with Baseline and Oracle for additional baselines. Results are displayed in Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13

F Results presented with Standard Deviations

Recall that the results presented in the main text were computed by taking the mean of a 5-fold leave-one-out cross-validation. Below, we report the corresponding standard deviations of the five folds. The standard deviations for the results on accuracy reported in Table 2 can be found in Table 8, Table 9, Table 10. For fairness-calibration in Table 3, they can be found in Table 11, Table 12 Table 13, Table 14. Finally, for predictive equality and equal opportunity in Table 4 and Table 7, they can be found in Table 15, Table 16, Table 17 and Table 18, Table 19, Table 20.

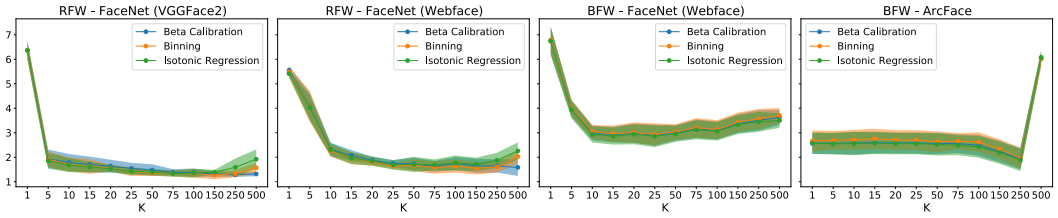

Figure 3: Comparison of fairness-calibration as measured by the subgroup mean of the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

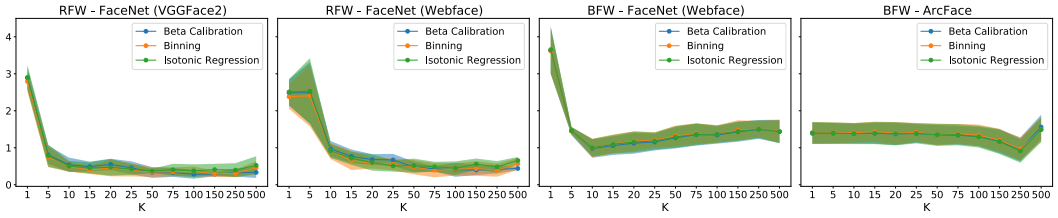

Figure 4: Bias in fairness-calibration as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

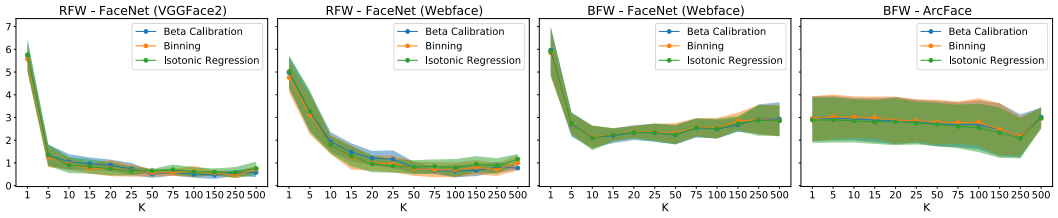

Figure 5: Bias in fairness-calibration as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

Table 8: Global **accuracy** measured by the AUROC.

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	88.26± 0.19	83.95± 0.22	96.06± 0.16	97.41± 0.34
FTC	86.46± 0.17	81.61± 0.57	93.30± 0.70	96.41± 0.53
FSN [38]	90.05± 0.26	85.84± 0.34	96.77± 0.20	97.35± 0.33
BMC (Ours)	**90.58± 0.29**	**86.71± 0.25**	**96.90± 0.17**	**97.44± 0.34**
Oracle (Ours)	*89.74± 0.31*	*85.23± 0.18*	*97.28± 0.13*	98.91± 0.12

Table 9: Global **accuracy** measured by the TPR at 0.1% FPR threshold.

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	18.42± 1.28	11.18± 3.45	33.61± 2.10	86.27± 1.09
FTC	6.86± 5.24	4.65± 2.10	13.60± 4.92	82.09± 1.11
FSN [38]	23.01± 2.00	17.33± 3.01	**47.11± 1.23**	86.19± 1.13
BMC (Ours)	**23.55± 1.82**	**20.64± 3.09**	46.74± 1.49	**86.28± 1.24**
Oracle (Ours)	*21.40± 3.54*	*16.71± 1.98*	*45.13± 1.45*	86.41± 1.19

Table 10: Global **accuracy** measured by the TPR at 1% FPR threshold.

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	34.88± 3.27	26.04± 2.11	58.87± 0.92	90.11± 0.87
FTC	23.66± 6.58	18.40± 4.02	43.09± 5.70	88.24± 0.63
FSN [38]	40.21± 2.09	32.80± 1.03	68.92± 1.01	90.06± 0.84
BMC (Ours)	**41.88± 1.99**	**33.13± 1.67**	**69.21± 1.19**	**90.14± 0.86**
Oracle (Ours)	*41.83± 2.98*	*31.60± 1.08*	*67.56± 1.05*	90.40± 0.91

Table 11: Fairness-calibration as measured by mean KS across sensitive subgroups.

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	6.37± 0.35	5.55± 0.14	6.77± 0.57	2.57± 0.43
FTC	5.69± 0.14	4.73± 0.53	6.64± 0.41	2.95± 0.45
FSN [38]	1.43± 0.28	2.49± 0.46	**2.76± 0.21**	2.65± 0.43
BMC (Ours)	**1.37± 0.17**	**1.75± 0.26**	3.09± 0.37	**2.49± 0.43**
Oracle (Ours)	*1.18± 0.05*	*1.35± 0.09*	*2.23± 0.14*	1.41± 0.33

Table 12: Bias in fairness-calibration as measured by the deviations of KS across subgroups in terms of AAD (Average Absolute Deviation).

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	2.89± 0.29	2.48± 0.36	3.63± 0.63	1.39± 0.28
FTC	2.32± 0.28	1.93± 0.35	2.80± 0.55	1.48± 0.31
FSN [38]	0.35± 0.15	0.84± 0.38	1.38± 0.27	1.45± 0.31
BMC (Ours)	**0.28± 0.12**	**0.41± 0.19**	**1.34± 0.24**	**1.30± 0.26**
Oracle (Ours)	*0.28± 0.08*	*0.38± 0.20*	*1.15± 0.24*	0.59± 0.18

Table 13: Bias in fairness-calibration as measured by the deviations of KS across subgroups in terms of MAD (Maximum Absolute Deviation).

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	5.73± 0.63	4.97± 0.72	5.96± 1.05	2.94± 0.99
FTC	4.51± 0.64	3.86± 0.70	5.61± 0.66	3.03± 0.88
FSN [38]	0.57± 0.21	1.19± 0.38	2.67± 0.32	3.23± 0.99
BMC (Ours)	**0.50± 0.15**	**0.64± 0.28**	**2.48± 0.41**	**2.68± 1.07**
Oracle (Ours)	*0.53± 0.18*	*0.66± 0.28*	*2.63± 0.60*	1.30± 0.29

Table 14: Bias in fairness-calibration as measured by the deviations of KS across subgroups in terms of STD (Standard Deviation).

	RFW		BFW	
	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	3.77± 0.33	2.91± 0.41	4.03± 0.70	1.63± 0.40
FTC	2.95± 0.32	2.28± 0.43	3.27± 0.46	1.74± 0.42
FSN [38]	0.40± 0.15	0.91± 0.36	1.60± 0.23	1.71± 0.41
BMC (Ours)	**0.34± 0.12**	**0.45± 0.20**	**1.55± 0.24**	**1.52± 0.37**
Oracle (Ours)	*0.33± 0.10*	*0.43± 0.20*	*1.40± 0.27*	0.69± 0.18

Table 15: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, we compare the deviations in subgroup FPRs in terms of AAD (Average Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.02	0.14± 0.03	0.29± 0.04	0.12± 0.03
	FTC	0.10± 0.02	0.12± 0.04	0.24± 0.02	**0.09± 0.02**
	FSN [38]	0.10± 0.05	0.11± 0.04	**0.09± 0.03**	0.11± 0.02
	BMC (Ours)	**0.09± 0.03**	**0.09± 0.03**	**0.09± 0.02**	0.11± 0.03
	Oracle (Ours)	*0.11± 0.05*	*0.11± 0.03*	*0.12± 0.03*	0.12± 0.04
1% FPR	Baseline	0.68± 0.06	0.67± 0.15	2.42± 0.14	0.72± 0.19
	FTC	0.60± 0.11	0.54± 0.12	1.94± 0.22	**0.54± 0.09**
	FSN [38]	0.37± 0.12	0.35± 0.16	0.87± 0.11	0.55± 0.11
	BMC (Ours)	**0.28± 0.11**	**0.29± 0.10**	**0.80± 0.10**	0.63± 0.15
	Oracle (Ours)	*0.40± 0.09*	*0.41± 0.10*	*0.77± 0.17*	0.83± 0.15

Table 16: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, we compare the deviations in subgroup FPRs in terms of MAD (Maximum Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.15± 0.05	0.26± 0.09	1.00± 0.28	0.30± 0.08
	FTC	0.15± 0.03	0.23± 0.08	0.74± 0.22	**0.20± 0.03**
	FSN [38]	0.18± 0.10	0.23± 0.07	0.20± 0.06	0.28± 0.08
	BMC (Ours)	**0.14± 0.04**	**0.16± 0.06**	**0.20± 0.04**	0.31± 0.10
	Oracle (Ours)	*0.19± 0.10*	*0.20± 0.07*	*0.25± 0.06*	0.27± 0.09
1% FPR	Baseline	1.02± 0.01	1.23± 0.30	7.48± 1.75	1.51± 0.44
	FTC	0.91± 0.08	1.05± 0.17	5.74± 1.73	**1.04± 0.15**
	FSN [38]	0.68± 0.23	0.61± 0.25	2.19± 0.58	1.27± 0.35
	BMC (Ours)	**0.46± 0.16**	**0.57± 0.23**	**1.79± 0.54**	1.46± 0.29
	Oracle (Ours)	*0.69± 0.19*	*0.74± 0.23*	*1.71± 0.59*	2.08± 0.57

Table 17: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, we compare the deviations in subgroup FPRs in terms of STD (Standard Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.03	0.16± 0.04	0.40± 0.09	0.15± 0.04
	FTC	0.11± 0.02	0.14± 0.05	0.32± 0.05	**0.11± 0.02**
	FSN [38]	0.11± 0.06	0.13± 0.04	**0.11± 0.03**	0.14± 0.03
	BMC (Ours)	**0.10± 0.03**	**0.10± 0.03**	**0.11± 0.03**	0.15± 0.03
	Oracle (Ours)	*0.12± 0.05*	*0.13± 0.03*	*0.15± 0.03*	0.14± 0.04
1% FPR	Baseline	0.74± 0.04	0.79± 0.18	3.22± 0.44	0.85± 0.20
	FTC	0.66± 0.09	0.66± 0.12	2.57± 0.45	**0.61± 0.08**
	FSN [38]	0.46± 0.14	0.40± 0.17	1.05± 0.18	0.68± 0.14
	BMC (Ours)	**0.32± 0.12**	**0.35± 0.13**	**0.95± 0.16**	0.78± 0.15
	Oracle (Ours)	*0.45± 0.11*	*0.48± 0.12*	*0.91± 0.22*	1.07± 0.18

Table 18: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, we compare the deviations in subgroup FNRs in terms of AAD (Average Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.09± 0.01	0.10± 0.02	0.09± 0.03	0.11± 0.02
	FTC	0.09± 0.01	**0.08± 0.03**	**0.04± 0.02**	**0.06± 0.01**
	FSN [38]	**0.09± 0.02**	0.09± 0.02	0.07± 0.02	0.12± 0.01
	BMC (Ours)	0.10± 0.02	0.11± 0.02	0.10± 0.02	0.09± 0.02
	Oracle (Ours)	*0.11± 0.02*	*0.12± 0.02*	*0.09± 0.02*	0.11± 0.02
1% FPR	Baseline	0.60± 0.17	0.45± 0.09	0.39± 0.05	0.75± 0.16
	FTC	0.48± 0.06	**0.32± 0.12**	**0.30± 0.07**	**0.49± 0.14**
	FSN [38]	**0.28± 0.06**	0.40± 0.19	0.41± 0.10	0.77± 0.17
	BMC (Ours)	0.30± 0.14	0.39± 0.12	0.32± 0.10	0.65± 0.11
	Oracle (Ours)	*0.38± 0.15*	*0.56± 0.11*	*0.37± 0.09*	0.50± 0.10

Table 19: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, we compare the deviations in subgroup FNRs in terms of MAD (Maximum Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.13± 0.02	0.16± 0.08	0.23± 0.11	0.31± 0.07
	FTC	**0.11± 0.02**	0.14± 0.07	**0.09± 0.03**	**0.14± 0.04**
	FSN [38]	0.13± 0.06	**0.14± 0.06**	0.22± 0.11	0.33± 0.06
	BMC (Ours)	0.14± 0.06	0.17± 0.09	0.27± 0.09	0.17± 0.05
	Oracle (Ours)	*0.18± 0.07*	*0.21± 0.08*	*0.24± 0.08*	0.32± 0.15
1% FPR	Baseline	0.96± 0.21	0.81± 0.14	0.84± 0.14	1.85± 0.66
	FTC	0.83± 0.21	**0.58± 0.24**	**0.62± 0.11**	**1.12± 0.29**
	FSN [38]	**0.47± 0.15**	0.78± 0.38	0.92± 0.28	1.91± 0.67
	BMC (Ours)	0.51± 0.25	0.72± 0.20	0.74± 0.18	1.48± 0.36
	Oracle (Ours)	*0.61± 0.15*	*1.06± 0.18*	*0.77± 0.19*	1.11± 0.27

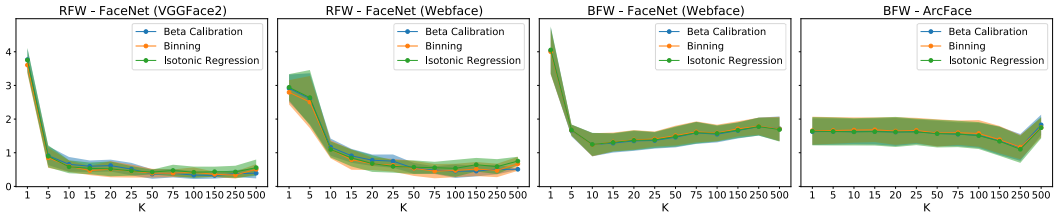

Figure 6: Bias in fairness-calibration as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

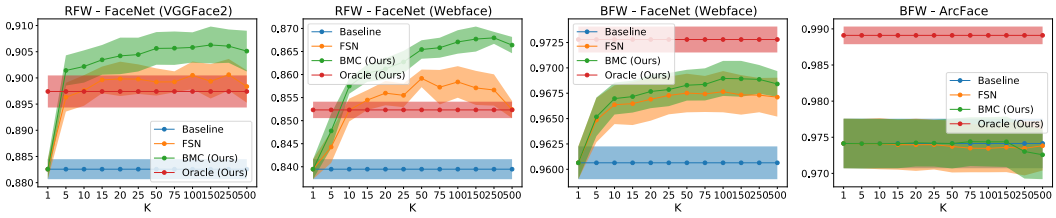

Figure 7: Global accuracy measured by the AUROC for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

Table 20: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, we compare the deviations in subgroup FNRs in terms of STD (Standard Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
		FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.01	0.11± 0.03	0.11± 0.04	0.14± 0.02
	FTC	**0.09± 0.01**	**0.10± 0.03**	**0.05± 0.02**	**0.07± 0.02**
	FSN [38]	0.09± 0.03	0.10± 0.03	0.10± 0.04	0.15± 0.02
	BMC (Ours)	0.10± 0.02	0.12± 0.03	0.13± 0.03	0.10± 0.02
	Oracle (Ours)	*0.12± 0.03*	*0.13± 0.03*	*0.11± 0.03*	0.14± 0.04
1% FPR	Baseline	0.67± 0.15	0.53± 0.09	0.47± 0.06	0.93± 0.23
	FTC	0.56± 0.10	**0.38± 0.15**	**0.34± 0.06**	**0.60± 0.16**
	FSN [38]	**0.32± 0.09**	0.48± 0.23	0.49± 0.12	0.96± 0.23
	BMC (Ours)	0.34± 0.17	0.48± 0.14	0.40± 0.10	0.80± 0.13
	Oracle (Ours)	*0.42± 0.14*	*0.67± 0.11*	*0.44± 0.10*	0.60± 0.12

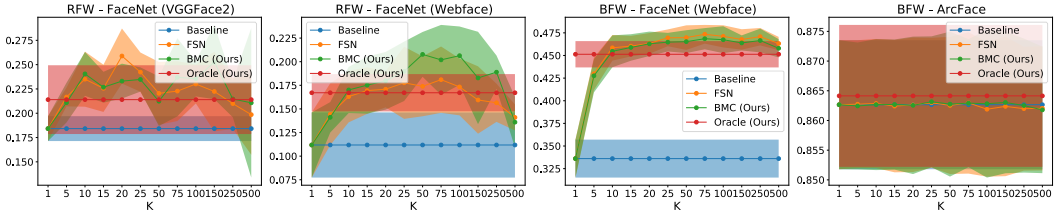

Figure 8: Global accuracy measure by the TPR at different a global 0.1% FPR for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

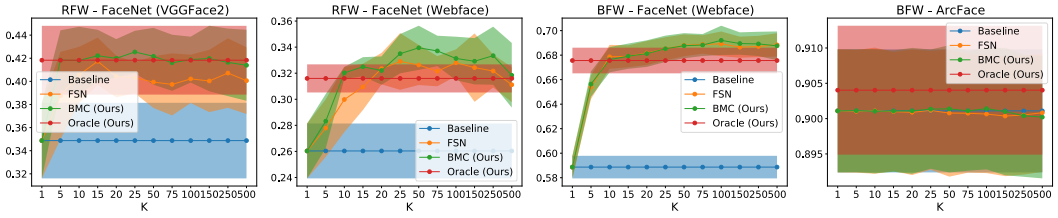

Figure 9: Global accuracy measure by the TPR at different a global 1% FPR for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

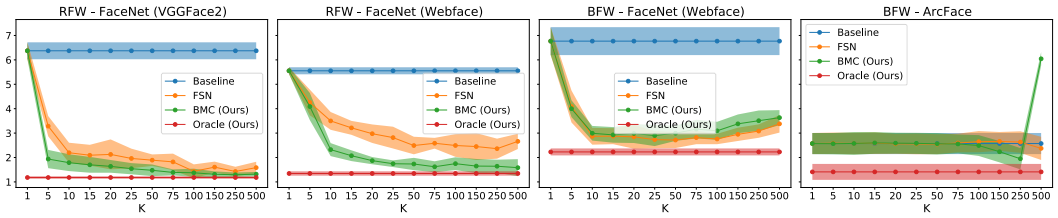

Figure 10: Comparison of fairness-calibration as measured by the subgroup mean of the KS across the sensitive subgroups for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

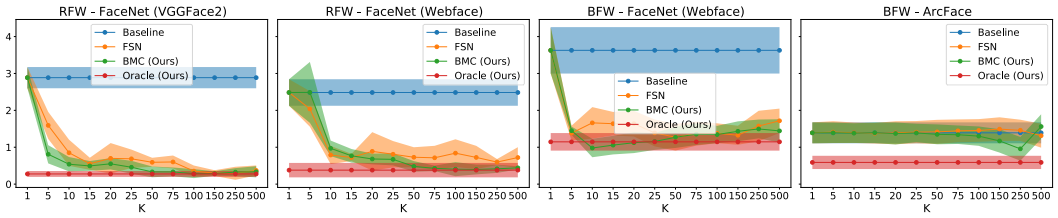

Figure 11: Bias in fairness-calibration as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

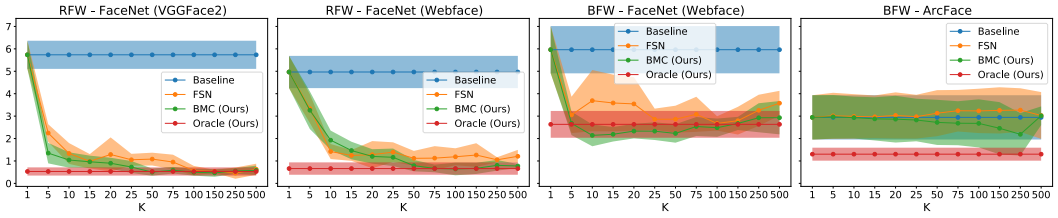

Figure 12: Bias in fairness-calibration as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

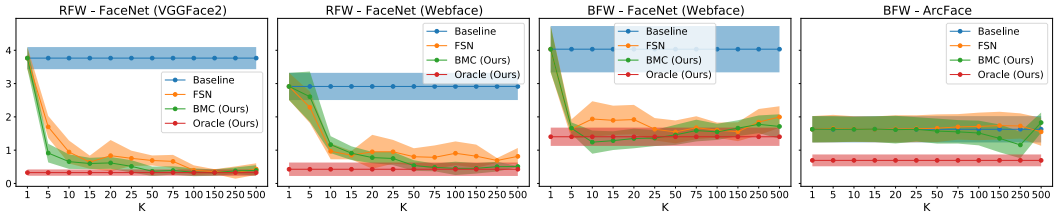

Figure 13: Bias in fairness-calibration as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN [38], BMC, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.