Fairness in Oculomotoric Biometric Identification

Paul Prasse*
University of Potsdam
Potsdam, Germany
paul.prasse@uni-potsdam.de

David R. Reich*
University of Potsdam
Potsdam, Germany
david.reich@uni-potsdam.de

Silvia Makowski
University of Potsdam
Potsdam, Germany
silvia.makowski@uni-potsdam.de

Lena A. Jäger
University of Zurich
Zurich, Switzerland
University of Potsdam
Potsdam, Germany
jaeger@cl.uzh.ch

Tobias Scheffer
University of Potsdam
Potsdam, Germany
tobias.scheffer@uni-potsdam.de

ABSTRACT

Gaze patterns are known to be highly individual, and therefore eye movements can serve as a biometric characteristic. We explore aspects of the *fairness* of biometric identification based on gaze patterns. We find that while oculomotoric identification does not favor any particular gender and does not significantly favor by age range, it is unfair with respect to ethnicity. Moreover, fairness concerning ethnicity cannot be achieved by balancing the training data for the best-performing model.

CCS CONCEPTS

• **Security and privacy** → *Biometrics*; • **Computer systems organization** → *Neural networks.*

KEYWORDS

biometrics, neural networks, fairness

ACM Reference Format:
Paul Prasse, David R. Reich, Silvia Makowski, Lena A. Jäger, and Tobias Scheffer. 2022. Fairness in Oculomotoric Biometric Identification. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22), June 8–11, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3517031.3529633

1 INTRODUCTION

It has long been known that gaze patterns have strong individual characteristics [Noton and Stark 1971] that are persistent over time [Bargary et al. 2017]. Motivated by these findings, eye movements have been explored as a biometric characteristic [Bednarik et al. 2005; Kasprowski and Ober 2004]. Earlier approaches rely on engineered features of fixations and saccadic movements [Bednarik et al. 2005; Holland and Komogortsev 2013; Kasprowski and Ober 2004; Rigas et al. 2016]. Since these macro-movements occur at a low frequency, such methods require minutes' worth of gaze data to identify a user [Makowski et al. 2019]. Recent work uses the sequence of raw angular velocities as input to a deep neural network [Jäger et al. 2020; Makowski et al. 2020, 2021]. In addition to information about macro-movements, the angular velocities contain information of high-frequency micro-movements of the eyes; this increases identification accuracy and reduces the time to identification to mere seconds.

Machine-learning systems are prone to offering disparate benefit to differing subgroups of users [Barocas and Selbst 2016; Feuerriegel et al. 2020]. Factors that contribute to such systematic unfairness can include biases and prejudices in the ground-truth labels as well as the distribution underlying the training data; the data selection process may reflect existing patterns of exclusion and inequality. Often-cited examples of unfairness include recidivism-prediction instruments that pick up racial biases from the data [Chouldechova 2017], and application-screening systems that pick up gender, racial, and other biases [Barocas et al. 2017].

Bias in face recognition has attracted public attention in 2017 when a simple test of Amazon's face recognition system showed that a disproportionate amount of false matches occur between people of color [Snow 2018]. In a comparison of four face-recognition systems, all systems have been observed to be unfair with respect to ethnicity [de Freitas Pereira and Marcel 2022]. The Association for Computing Machinery has even issued a statement urging the suspension of face-recognition technology due to a clear bias based on ethnicity, gender, and other characteristics [Committee 2020]. However, the issue of demographic bias is not limited to face recognition, but also affects other biometric modalities such as fingerprints, iris scan, and voice recognition. In a review paper on open questions in biometrics, fairness has been identified as one of the major challenges biometric research is currently facing [Ross et al. 2019] and the social impact of biased biometric systems has been highlighted [Drozdowski et al. 2020].

Motivated by the above findings, this paper explores the fairness of oculomotoric identification. We evaluate the performance of two state-of-the art models within and across genders, ethnicities, and age groups. We hypothesize that a major reason of any disparate performance between specific groups may be owed to imbalanced training data. We therefore investigate whether systematic unfairness can be mitigated by rebalancing the training data along the affected dimension.

*Both authors contributed equally to this research.

ETRA '22, June 8–11, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9252-5/22/06.
https://doi.org/10.1145/3517031.3529633

2 RELATED WORK

Fairness in biometrics has attracted increasing attention in recent years. The European Association for Biometrics (EAB) is announcing a conference specifically dedicated to the demographic fairness of biometric systems[1]. In a recent review paper on demographic bias in face, iris and fingerprint recognition systems, Drozdowski et al. [2020] conclude that all of these are biased with respect to gender, age and ethnicity. Buolamwini [2017] show that benchmark face databases are biased towards lighter skin tones and that gender classifiers fail more often on users with darker skin. Karkkainen and Joo [2021] acquire a face data set that is balanced with respect to ethnicity, gender and race and demonstrate that balanced training data improve the results on all ethnicities. Along the same lines, Fenu et al. [2021, 2020] show that in voice recognition systems, balancing the training data also reduces demographic biases. These findings are in contrast to a more recent investigation of gender bias in a range of voice recognition algorithms by Kathiresan [2022] who find that the gender bias is unaffected by balancing the training data, but is rather explained in terms of physiological differences between males and females and the resulting differences in the acoustic signal.

The precise definition of fairness and the corresponding metric used to evaluate a biometric system has social, ethical and legal implications and has been debated by different stakeholders [Rathgeb et al. 2021]. Previous research has used a range of different fairness metrics for the evaluation of biometric systems. The Disparity in Demographic Parity (DP) [Fenu et al. 2021] measures to which extent the system's decision is dependent of the user's protected group membership. The Disparity in Equal Opportunity (EOpp) [Fenu et al. 2021], measures whether true-positive events are independent of group membership. Finally, the Disparity in Equalized Odds (EOdd) [Fenu et al. 2021] measures whether both true- and false-positive events are independent of group membership.

de Freitas Pereira and Marcel [2022] introduce the Fairness Discrepancy Rate (FDR) which measures the difference in terms of the verification performance between demographic groups. In this paper, we use the FDR metrics to evaluate systems of eye movement-based biometric systems. For a formal definition of the FDR, see Section 3.

Two major approaches have been proposed to achieve fairness in algorithmic decision making. The first class of methods debiases the training data by balancing the classes or by adding adversarial examples [Fenu et al. 2021; Karkkainen and Joo 2021; Zhang et al. 2018]. The second class of methods aims at achieving fairness through the training procedure. Zemel et al. [2013] propose to learn fair representations by applying a fairness metric as optimization criterion. Subsequent work by Hardt et al. [2016] tries to achieve fairness by constructing randomized decision rules that ensure equalized odds of true- and false-positive events across protected groups.

3 PROBLEM STATEMENT

We study fairness in an *identity verification* setting. In oculomotoric biometric identity verification, each user is enrolled using at least one enrollment gaze sequence. At application time, the enrollment sequences are compared to a probe sequence using a similarity metric. If a similarity threshold τ is exceeded for an enrollment sequence, the presumed identity of the user is verified; otherwise, the user is exposed as an impostor. The performance can be characterized by a *false-match rate* (FMR, fraction of impostors among all accepted users) and a *false non-match rate* (FNMR, fraction of falsely rejected users among all rejected users). By changing the decision threshold, one can observe a *detection error trade-off curve (DET curve)*. The *equal error rate* (EER) is the point on this curve for which FMR equals FNMR.

To measure the fairness of the biometric system, we use the *fairness discrepancy rate* (FDR) [de Freitas Pereira and Marcel 2022]. The fairness discrepancy rate measures the difference in terms of FMR and FNMR between demographic groups, for a given decision threshold τ. Formally, given a set of demographic groups $D = \{d_1, \ldots, d_n\}$, FDR at threshold τ is defined as follows: when $A(\tau)$ (defined in Equation 2) is the greatest discrepancy in false-match rates and $B(\tau)$ (Equation 3) is the greatest discrepancy in false non-match rates between any two demographic groups, then the fairness discrepancy rate at threshold τ, as defined by Equation 1, is one minus the weighted sum of $A(\tau)$ and $B(\tau)$. Weight α controls the trade-off between false-match and false non-match rates; in our experiments, we use a value of $\alpha = 0.5$.

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)), \text{ with} \tag{1}$$

$$A(\tau) = \max_{d_i, d_j \in D} (|FMR^{d_i}(\tau) - FMR^{d_j}(\tau)|), \text{ and} \tag{2}$$

$$B(\tau) = \max_{d_i, d_j \in D} (|FNMR^{d_i}(\tau) - FNMR^{d_j}(\tau)|). \tag{3}$$

The fairness discrepancy rate $FDR(\tau)$ is a function of the decision threshold τ that controls the trade-off between FMR and FNMR. The notion of $\tau = FMR_x$ refers to the threshold value that achieves a FMR of at most x. The *area under the FDR curve* (FDR_{AUC}) integrates the area under $FDR(\tau)$ over different thresholds τ. By scaling the values of different thresholds into the range of zero to one, the FDR_{AUC} lies in the range of between zero and one. In this manuscript, we use the thresholds for $FMR \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$. de Freitas Pereira and Marcel used synthetic data to determine a threshold for the FDR_{AUC}, for which models can be characterized as unfair. For $\alpha = 0.5$, de Freitas Pereira and Marcel [2022] consider FDR_{AUC} values of below 0.9 to indicate unfair models.

4 METHODS

In this section we describe the biometric models and data.

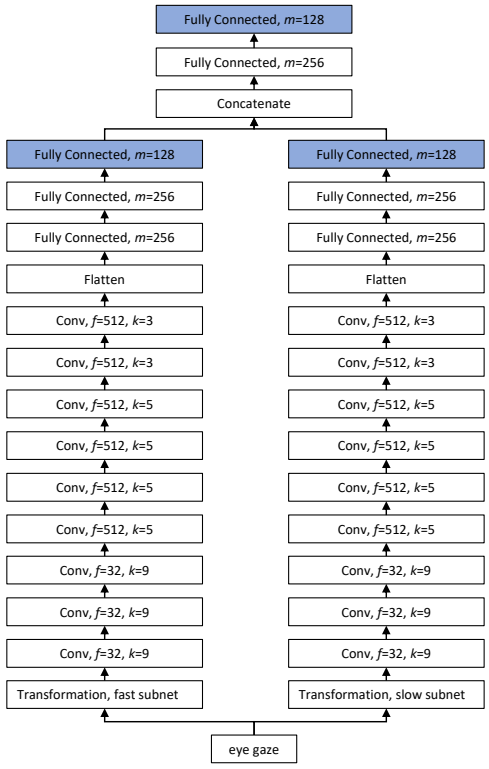
4.1 Biometric Models

We investigate two state-of-the-art methods of eye-tracking-based biometric identification, one feature-based method proposed by Lohr et al. [2020] (referred to as *Lohr et al.*) that processes a range of engineered fixational, saccadic, and post-saccadic oscillation features, and *DeepEyedentification*, an end-to-end-trained neural network that processes the raw eye-tracking signal [Jäger et al. 2020]; we use a version of the network that processes binocular input signals [Makowski et al. 2021] which we will refer to as *DeepEye*.

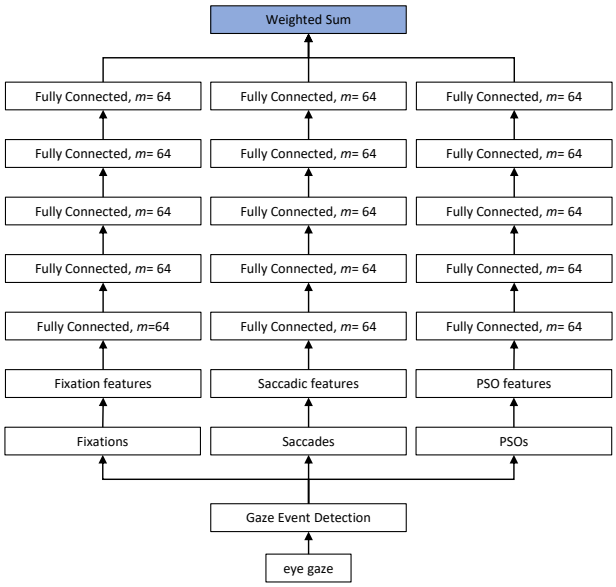
[1] EAB virtual events series—Demographic Fairness in Biometric Systems https://eab.org/events/program/237.

Table 1: Demographics of the GazeBase data [Griffith et al. 2021]. Numbers in brackets show the number of female (F) and male (M) subjects.

	Self-reported Ethnicity					
Age	Asian	Black	Caucasian	Hispanic	Mixed	Sum
≤ 20	3 (2 F, 1 M)	15 (4 F, 11 M)	73 (36 F, 37 M)	44 (20 F, 24 M)	15 (8 F, 7 M)	150 (70 F, 80 M)
21-29	6 (3 F, 3 M)	14 (6 F, 8 M)	88 (44 F, 44 M)	31 (15 F, 16 M)	11 (5 F, 6 M)	150 (73 F, 77 M)
≥ 30	1 (0 F, 1 M)	0 (0 F, 0 M)	16 (7 F, 9 M)	4 (1 F, 3 M)	1 (0 F, 1 M)	22 (8 F, 14 M)
Sum	10 (5 F, 5 M)	29 (10 F, 19 M)	177 (87 F, 90 M)	79 (36 F, 43 M)	27 (13 F, 14 M)	322 (151 F, 171 M)



(a) DeepEye model architecture. For more details see [Jäger et al. 2020; Makowski et al. 2021; Prasse et al. 2020].



(b) Lohr *et al.* model architecture. For more details see [Lohr et al. 2020].

Figure 1: Overview of the two models used for biometric identification. The blue boxes highlight the layers used to extract the neural embedding of an eye-gaze sequence.

4.1.1 DeepEyedentification. The DeepEyedentifaction network proposed by [Makowski et al. 2021] is an end-to-end trained convolutional neural network that processes the binocular yaw and pitch gaze angles recorded by an eye tracker. After transforming the angular coordinates into gaze velocities, the signal is duplicated and processed by two separate convolutional subnets. One subnet is optimised to process fast saccadic eye movements whereas the other subnet is optimised to process slow fixational micro-movements such as drift. The output of the two subnets is concatenated and further processed by a dense network. The model is trained in a multi-class classification setting with known user identities as labels. At application time, the softmax output layer is removed and the embedding of a gaze sequence is computed by concatenating the activation of the outputs of the two subnets and the joint dense network. An overview of the architecture is shown in Figure 1(a). A probe sequence is confirmed to match a presumed identity if the cosine similarity of the embeddings of probe sequence and any enrollment sequence exceeds a threshold τ.

4.1.2 Lohr et al. The model of Lohr et al. [2020] segments gaze sequences into fixations, saccades, and post-saccadic oscillations (PSOs). Various statistics introduced by Rigas et al. [2018] and George and Routray [2016] of these oculomotoric events serve as input features to three multilayer perceptrons (MLPs), one for each type of

event. The MLPs are trained using the triplet loss function [Schroff et al. 2015] to learn an embedding that distinguishes different users from each other. The embedding for each user at application time is the weighted sum of the three MLP output embeddings [Lohr et al. 2020]. An overview of the architecture is provided in Figure 1(b). At application time, the cosine similarity of probe and enrollment gaze sequences are compared to threshold τ.

4.2 Training Data and Environment

We use the *GazeBase* database [Griffith et al. 2021]; Table 1 shows its underlying demographic information. The GazeBase meta-data distinguishes between five different self-reported ethnicities; we split participants into three age groups. GazeBase contains gaze recordings of 322 college-aged participants who are recorded monocularly with an EyeLink 1000 eye tracker at a sampling frequency of 1,000 Hz. Participants are recorded over two identical, consecutive recording sessions with a break of at most 5 minutes and in which they perform seven tasks: A random saccade (RAN), a reading task (TEX), two video viewing tasks (VD1 and VD2), a fixation task (FXS), a horizontal saccade task (HSS) and a gaze-driven game called Balura (BLG). This experiment is repeated over a time period of 37 months, resulting in 9 recording rounds. Participants of subsequent round are recruited only from the pool of participants of the previous round. In each recording round, the subjects participated in two recording sessions separated by a short break in which they were presented with the same type of stimuli. In this study, we use the data collected in both recording sessions from the first four rounds which are temporally relatively close (within 8 month). This ensures that any individual or group-level effects of growing older do not confound our comparisons, and, moreover, accounts for the participants' high dropout rate (only 14 out of 322 subjects are participating in the last round). The GazeBase subset that we use in this study was collected over a period of eight months.

We evaluate every setting on a NVIDIA DGX A100. We train all neural networks using the Keras [Chollet et al. 2015] and Tensorflow [Dean et al. 2015] libraries utilizing the NVIDIA CUDA platform. We implement the evaluation framework using the scikit-learn [Pedregosa et al. 2011] machine learning package. The code can be found online.[2]

5 RESULTS

The two models under investigation differ in terms of the input they consume. DeepEye uses a fixed length of one second as input. In this work, we investigate the setting of a probe window of ten seconds for DeepEye by averaging the similarity metric over 10 embeddings of one-second probe subsequences. The Lohr *et al.* model uses engineered features for extracted eye movement events and processes a complete trial. For that reason, we evaluate the models with respect to their performance within 10 seconds for DeepEye and within one trial for the model by Lohr *et al.*, respectively. Our experiments therefore (unfairly) favor Lohr *et al.* over DeepEye by supplying this model with more input data.

[2]https://www.github.com/aeye-lab/etra-fairness

5.1 Fairness Experiments in the Wild

In this section, both models are evaluated by resampling 10 times from the GazeBase data set. In each iteration, a training population of 100 subjects is selected and the remaining 222 subjects are used for evaluation. We use a large test set to ensure that we can evaluate our model on all demographic groups. Moreover, previous research has shown that 100 persons in the training data are sufficient for training the two models under investigation [Lohr et al. 2020; Makowski et al. 2021]. At application time, one user is enrolled by calculating and storing the mean embedding of input data (i.e., 80 randomly selected one-second windows for DeepEyedentification and complete trials for the Lohr et al. model), taken from the first recording sessions of all available rounds of a task. At test time, each user in turn serves as test user: a probe sequence (i.e., 10 consecutive one-second windows for DeepEyedentification and one complete trial for the Lohr et al. model), taken from the second recording session, is then compared against the enrollment embedding. We count the verification of an enrolled user as true match if the cosine similarity between any of the probe embeddings and the enrollment embedding exceeds a detection threshold and else as false non-match.

For each demographic (gender, ethnicity, age), we compare the verification performance (EER, FMR, FNMR) of the models for the different demographic groups and values of FDR and FDR_{AUC}.

5.1.1 Gender. For gender, Table 2 shows fairness discrepancy rates for thresholds that result in false non-match rates of 10^{-1} through 10^{-3} across the different tasks of the GazeBase data for both DeepEye and Lohr *et al.*. All FDR_{AUC}-values are significantly higher than 0.9 ($p < 0.05$, based on a one-tailed t-test), indicating fairness across genders for all tasks. There is no significant difference between the verification performance of different genders either. Tables 1-6 and Figure 1 in the Appendix give a detailed account of FMR-, FNMR-, and EER-values between and within genders.

5.1.2 Age. For the evaluation of different age groups, we define a group of younger (age below 20) and older (age above 30) subjects; this particular partitioning is owed to the age distribution in the data. Table 3 shows that in this setting, nearly all the FDR_{AUC}-values are significantly higher ($p < 0.05$, one-tailed t-test) than 0.9, indicating that both models are fair across age groups. For the FXS, and Video task the FDR_{AUC}-values for the DeepEye model are not significantly higher than 0.9. There is no significant difference between the verification performance of different age groups for the DeepEye model, but we observe that for the Lohr *et al.* model and the TEX task, the EER for a comparison between older subjects is slightly but significantly higher than between younger subjects (see Table 7-12 and Figure 1 in the Appendix).

5.1.3 Ethnicity. Table 4 shows that both models are unfair with respect to ethnicities. Both models reach the best EER-values for the comparison between Caucasians and the lowest EER-values for the comparison between Asians and Black subjects (see Table 13-18 in the Appendix). For DeepEye, all EER-values for the comparison of Black, Asian, and Hispanic subjects are significantly worse than the comparison between Caucasians for the TEX, HSS, and Video tasks. For the RAN and the BLG task, the EER-values for DeepEye are not significantly different between ethnic groups. The EER-values

Table 2: FDR for different thresholds and different genders. All values are averaged over 10 iterations and the standard error is reported. For results marked "*", the FDR_{AUC} is significantly greater ($p < 0.05$) than 0.9.

	$\tau = FMR_x$	$x = 10^{-1}$	$x = 10^{-2}$	$x = 10^{-3}$	
	Task		$FDR(\tau)$		FDR_{AUC}
DeepEye	TEX	0.99 ± 0.0	0.98 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	BLG	0.98 ± 0.01	0.98 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	FXS	0.97 ± 0.0	0.97 ± 0.01	0.99 ± 0.0	0.97 ± 0.01*
	HSS	0.98 ± 0.0	0.97 ± 0.01	0.97 ± 0.01	0.98 ± 0.0*
	RAN	0.98 ± 0.0	0.98 ± 0.01	0.98 ± 0.0	0.98 ± 0.0*
	Video	0.98 ± 0.0	0.99 ± 0.0	0.98 ± 0.0	0.98 ± 0.0*
Lohr *et al.*	TEX	0.97 ± 0.01	0.98 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	BLG	0.98 ± 0.01	0.99 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	FXS	0.98 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	HSS	0.97 ± 0.01	0.98 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	RAN	0.97 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.98 ± 0.0*
	Video	0.98 ± 0.0	0.98 ± 0.0	1.0 ± 0.0	0.98 ± 0.0*

Table 3: FDR for different thresholds and different age groups. All values are averaged over 10 iterations and the standard error is reported. For results marked "*", the FDR_{AUC} is significantly greater ($p < 0.05$) than 0.9.

	$\tau = FMR_x$	$x = 10^{-1}$	$x = 10^{-2}$	$x = 10^{-3}$	
	Task		$FDR(\tau)$		FDR_{AUC}
DeepEye	TEX	0.97 ± 0.0	0.93 ± 0.01	0.91 ± 0.01	0.95 ± 0.0*
	BLG	0.94 ± 0.01	0.91 ± 0.01	0.96 ± 0.01	0.92 ± 0.01*
	FXS	0.91 ± 0.01	0.93 ± 0.02	0.98 ± 0.01	0.91 ± 0.01
	HSS	0.96 ± 0.01	0.91 ± 0.02	0.87 ± 0.02	0.93 ± 0.01*
	RAN	0.95 ± 0.01	0.94 ± 0.01	0.93 ± 0.02	0.94 ± 0.01*
	Video	0.94 ± 0.01	0.91 ± 0.02	0.94 ± 0.01	0.92 ± 0.01
Lohr *et al.*	TEX	0.95 ± 0.01	0.98 ± 0.01	0.99 ± 0.0	0.96 ± 0.01*
	BLG	0.96 ± 0.01	0.96 ± 0.01	0.99 ± 0.0	0.96 ± 0.01*
	FXS	0.9 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	0.92 ± 0.01*
	HSS	0.92 ± 0.01	0.95 ± 0.01	0.98 ± 0.0	0.93 ± 0.01*
	RAN	0.93 ± 0.02	0.96 ± 0.01	0.98 ± 0.01	0.94 ± 0.01*
	Video	0.96 ± 0.01	0.97 ± 0.01	0.99 ± 0.0	0.96 ± 0.0*

for the comparison between Hispanic and Black subjects for the DeepEye model is significantly worse than the comparison between Caucasians for the FXS task.

For the Lohr *et al.* model there is no significant difference between all EER-values for the comparison of Asians and Caucasians across all tasks. The EER-values are significantly worse for the comparison between Hispanics and Caucasians for the TEX, BLG, FXS, and Video task, and the comparison between Black or Hispanic and Caucasian users for the HSS, and RAN task.

Figure 2 shows a t-SNE visualization of the internal embedding of a model trained on the TEX task. Here, we see that the embeddings of different ethnicities only partially overlap, and that Caucasians appear to have the largest inter-subject variability. This is consistent with the models' favoring of Caucasian users.

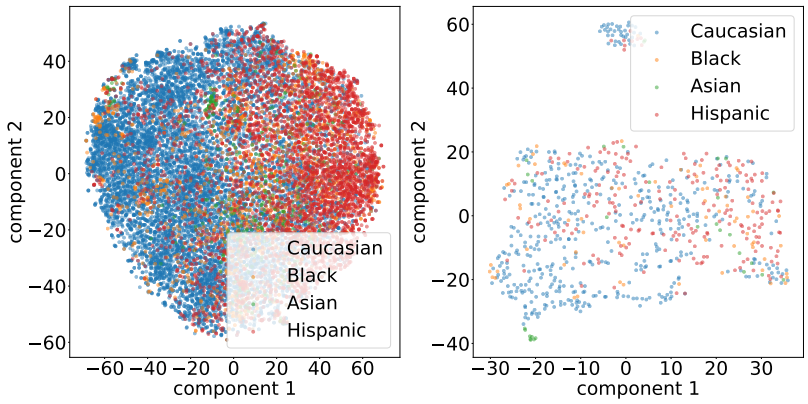
5.2 Impact of Training Data Distribution on Fairness

As Table 1 shows, the distribution of the demographic groups in the GazeBase data set is unbalanced, especially with respect to ethnicity. In this section, we investigate how varying the training data distribution affects the fairness of both models. To this end, we use the same evaluation protocol as in Section 5.1, but only use Caucasian and Hispanic subjects and a training population of only 50 users to ensure to have at least 29 Hispanics in the test set. We then study the performance of the models as a function of the proportion of Caucasian subjects in the training data, ranging from zero to one.

Figure 3 shows the effect of varying the proportion of Caucasian versus Hispanic subjects in the training data on the FDR_{AUC} (detailed results can be found in Figure 2 and 3 in the Appendix). For the Lohr *et al.* model, we obtain models with FDR_{AUC}-values greater than 0.9 for the TEX, BLG, FXS, and Video setting; for the BLG task,

Table 4: Ethnicity. FDR for different thresholds. All values are averaged over 10 iterations and the standard error is reported. For results marked "*", the FDR_{AUC} is significantly greater ($p < 0.05$) than 0.9.

	$\tau = FMR_x$	$x = 10^{-1}$	$x = 10^{-2}$	$x = 10^{-3}$	
	Task	$FDR(\tau)$			FDR_{AUC}
DeepEye	TEX	0.89 ± 0.02	0.86 ± 0.01	0.9 ± 0.01	0.87 ± 0.01
	BLG	0.84 ± 0.02	0.87 ± 0.02	0.95 ± 0.01	0.85 ± 0.02
	FXS	0.78 ± 0.02	0.86 ± 0.02	0.97 ± 0.0	0.81 ± 0.02
	HSS	0.83 ± 0.02	0.84 ± 0.02	0.9 ± 0.01	0.84 ± 0.01
	RAN	0.86 ± 0.01	0.8 ± 0.01	0.93 ± 0.0	0.81 ± 0.01
	Video	0.82 ± 0.02	0.86 ± 0.01	0.92 ± 0.01	0.83 ± 0.01
Lohr *et al.*	TEX	0.84 ± 0.02	0.88 ± 0.01	0.96 ± 0.01	0.85 ± 0.01
	BLG	0.88 ± 0.01	0.93 ± 0.01	0.96 ± 0.0	0.9 ± 0.01
	FXS	0.85 ± 0.02	0.92 ± 0.01	0.96 ± 0.0	0.88 ± 0.01
	HSS	0.82 ± 0.01	0.9 ± 0.01	0.92 ± 0.01	0.86 ± 0.01
	RAN	0.8 ± 0.04	0.91 ± 0.01	0.96 ± 0.01	0.86 ± 0.02
	Video	0.84 ± 0.02	0.92 ± 0.01	0.96 ± 0.01	0.88 ± 0.01



(a) DeepEyedentification (b) Ethnicity Lohr et al.

Figure 2: Embedding space for a trained model on the TEX task for different ethnicities.

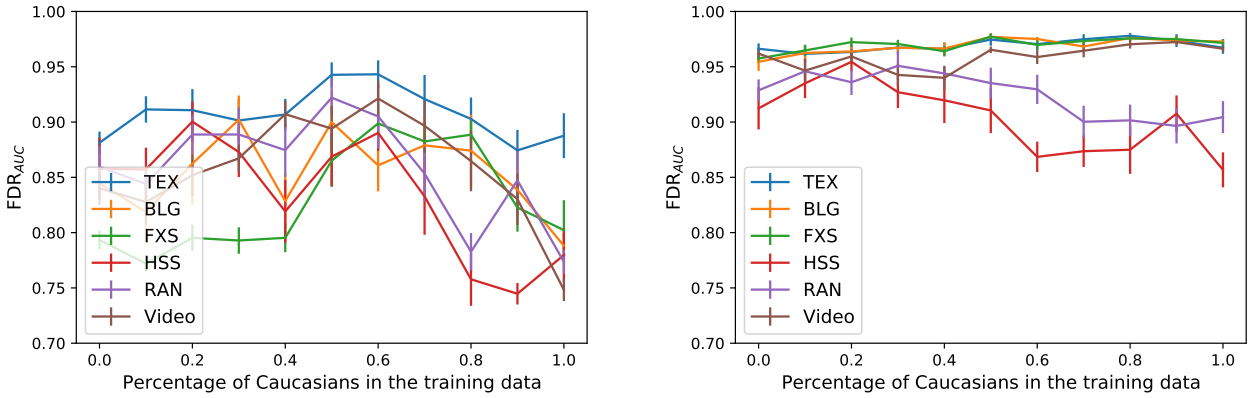
proportions of Caucasians below 0.3 lead to a significantly lower FDR_{AUC} than balanced data sets. For the HSS setting, a balanced dataset does not result in an FDR_{AUC} significantly greater than 0.9. For the RAN setting, a balanced data set reaches FDR_{AUC}-values greater than 0.9, wheras for proportions of Caucasians greater than 0.6 the FDR_{AUC}-values are not significantly greater than 0.9.

DeepEye only reaches an FDR_{AUC}-values significantly greater than 0.9 for TEX and using balanced data (proportion of Caucasians of 0.5 or 0.6). For the TEX setting, the FDR_{AUC} for only using Caucasians or only using Hispanics is significantly lower than the FDR_{AUC} for a balanced data set. We conclude that at least for DeepEyedentification, for most viewing tasks fairness cannot be achieved by any balancing of ethnicities in the training data.

6 DISCUSSION

We investigated the fairness of two different oculomotoric biometric systems: a deep neural network that has been trained to extract features from the raw eye-tracking signal and a model that uses engineered features which are extracted in a preprocessing step. We conclude that both these state-of-the-art systems are unfair with respect to different ethnicities. With respect to age and gender, we do not see any evidence for bias: the verification performance is nearly the same for different genders and different age groups. This is a major difference from other biometric systems like face [de Freitas Pereira and Marcel 2022] or voice recognition systems [Kathiresan 2022]. However, it should be noted that the available age-range in the *GazeBase* data set is relatively limited. Hence, further investigations into bias of oculomotoric biometric identification with participants of a more diverse age-range is warranted [Lee et al. 2019].

We further investigated if we are able to train a fair model by controlling the distribution of the input data. To this end, we trained a model on controlled proportions of Caucasian and Hispanic subjects. From this experiment we conclude, that balancing the training data is not enough to get a fair model. Unfortunately, the *GazeBase* database incorporates too few individuals of other ethnicities



(a) FDR_{AUC} **over varying percentage of Caucasian subjects in the training population of DeepEye.**

(b) FDR_{AUC} **over varying percentage of Caucasian subjects in the training population of Lohr *et al.*.**

Figure 3: Ethnicity. FDR over the proportion of ethnic groups in the training distribution.

for it to be possible to balance more than Caucasian and Hispanic subjects. Considering only two ethnicities, and in contrast to experiments in e. g. facial recognition, the demographic distribution in the training data does not explain the bias we observe in our experiments. This raises the question what is the reason for the observed bias? One possible explanation could be a measurement bias within eye-tracking devices.

Our study underscores the need to collect ethnically diverse and balanced data and to develop new methods to investigate the reasons for unfair model behavior, and ultimately achieve fairness in eye movement-based biometrics.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research under grant 01|S20043.

REFERENCES

Gary Bargary, Jenny M. Bosten, Patrick T. Goodbourn, Adam J. Lawrance-Owen, Ruth E. Hogg, and J.D. Mollon. 2017. Individual differences in human eye movements: An oculomotor signature? *Vision Research* 141 (Dec. 2017), 157–169.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS tutorial* 1 (2017), 2017.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 3 (June 2016), 671–732.

Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. 2005. Eye-movements as a biometric. In *Scandinavian Conference on Image Analysis*. 780–789.

Joy Adowaa Buolamwini. 2017. *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers.* Ph. D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA, USA.

François Chollet et al. 2015. *Keras.* https://keras.io

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

ACM U.S. Public Policy Committee. 2020. *Statement on principles and prerequisites for the development, evalutation and use of unbiased facial recognition technologies.* Retrieved Dec. 20, 2021 from https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

Tiago de Freitas Pereira and Sébastien Marcel. 2022. Fairness in biometrics: A Figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 1 (Jan. 2022), 19–29.

Jeffrey Dean, Rajat Monga, et al. 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems.* https://www.tensorflow.org/

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

Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. 2021. Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition. In *Proceedings of Interspeech 2021*. 1892–1896.

Gianni Fenu, Giacomo Medda, Mirko Marras, and Giacomo Meloni. 2020. Improving Fairness in Speaker Recognition. In *Proceedings of the 2020 European Symposium on Software Engineering*. 129–136.

Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair AI: Challenges and opportunities. *Business & Information Systems Engineering* 62, 4 (May 2020), 379–384.

Anjith George and Aurobinda Routray. 2016. A score level fusion method for eye movement biometrics. *Pattern Recognition Letters* 82 (Oct. 2016), 207–215.

Henry Griffith, Dillon Lohr, Evgeny Abdulin, and Oleg Komogortsev. 2021. GazeBase, a large-scale, multi-stimulus, longitudinal eye movement dataset. *Scientific Data* 8, Article 184 (July 2021), 9 pages.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. 3323–3331.

Corey D. Holland and Oleg V. Komogortsev. 2013. Complex eye movement pattern biometrics: Analyzing fixations and saccades. In *2013 International Conference on Biometrics (ICB '13)*. IEEE, 1–8.

Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2020. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science*. Springer, Cham, 299–314.

Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1548–1558.

Pawel Kasprowski and Jozef Ober. 2004. Eye movements in biometrics. In *International Workshop on Biometric Authentication*. 248–258.

Thayabaran Kathiresan. 2022. Gender bias in voice recognition: An i- and x-vector-based gender-specific automatic speaker recognition study. *ASIV* (2022).

Won June Lee, Ji Hong Kim, Yong Un Shin, Sunjin Hwang, and Han Woong Lim. 2019. Differences in eye movement range based on age and gaze direction. *Eye* 33, 7 (2019), 1145–1151.

Dillon Lohr, Henry Griffith, Samantha Aziz, and Oleg Komogortsev. 2020. A metric learning approach to eye movement biometrics. In *2020 IEEE International Joint Conference on Biometrics (IJCB '20)*. 1–7.

Silvia Makowski, Lena A. Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2019. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*. Springer, Cham, 209–225.

Silvia Makowski, Lena A. Jäger, Paul Prasse, and Tobias Scheffer. 2020. Biometric identification and presentation-attack detection using micro-movements of the

eyes. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB '20)*. 1–10.

Silvia Makowski, Paul Prasse, David R Reich, Daniel Krakowczyk, Lena A Jäger, and Tobias Scheffer. 2021. DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 4 (2021), 506–518.

David Noton and Lawrence Stark. 1971. Scanpaths in eye movements during pattern perception. *Science* 171, 3968 (1971), 308–311.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Paul Prasse, Lena A. Jäger, Silvia Makowski, Moritz Feuerpfeil, and Tobias Scheffer. 2020. On the relationship between eye tracking resolution and performance of oculomotoric biometric identification. *Procedia Computer Science* 176 (2020), 2088–2097.

Christian Rathgeb, Pawel Drozdowski, Naser Damer, Dinusha C. Frings, and Christoph Busch. 2021. Demographic fairness in biometric systems: What do the experts say? arXiv:2105.14844 [cs.CV]

Ioannis Rigas, Lee Friedman, and Oleg Komogortsev. 2018. Study of an extensive set of eye movement features: Extraction methods and statistical analysis. *Journal of Eye Movement Research* 11, 1 (2018).

Ioannis Rigas, Oleg Komogortsev, and Reza Shadmehr. 2016. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transactions on Applied Perception* 13, 2 (2016), 6.

Arun Ross, Sudipta Banerjee, Cunjian Chen, Anurag Chowdhury, Vahid Mirjalili, Renu Sharma, Thomas Swearingen, and Shivangi Yadav. 2019. Some research problems in biometrics: The future beckons. In *2019 International Conference on Biometrics (ICB)*. IEEE, 1–8.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

Jacob Snow. 2018. *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. 335–340.