Operational fairness when coding facial authentication

Mélanie Gornet, Claude Kirchner, Catherine Tessier

Operational fairness when coding facial authentication

Mélanie Gornet
melanie.gornet@telecom-paris.fr
Institut Polytechnique de Paris, Télécom Paris, i3 *

Claude Kirchner
claude.kirchner@inria.fr
CNPEN, Inria

Catherine Tessier
catherine.tessier@onera.fr
ONERA / DTIS, Université de Toulouse

Abstract

When dealing with machine learning, engineers tend to focus on improving certain aspects of performance of their system, such as efficiency, possibly dismissing other important criteria, like fairness. This mindset can have dreadful consequences for companies as well as for end users and may yield discrimination, for instance when resulting in automated facial recognition systems that work better for white men than for women of color (Buolamwini & Gebru, 2018). Researchers have long reduced fairness to a data issue: if the learning data is unbalanced, the system is quite likely to be biased. But this belief overlooks other parameters or coding choices that are also likely to affect fairness. Which coding choices really affect fairness and what are the trade-offs with efficiency? In this paper, focusing on facial recognition, various choices are considered regarding data sampling, normalization and augmentation, neural network depth, loss function margin, learning rate, and the authentication threshold. All of these choices have been tested on different metrics for efficiency and fairness. The results show that all of them have an impact on fairness at various scales. The best choice for fairness is not always the best for efficiency and trade-offs are sometimes necessary. Ethical discussions should therefore come with the design of machine learning systems, making such conflicts explicit and guiding the decisions at coding and software maintenance times.

*When this work begun, the author was affiliated to ISAE-SUPAERO, Toulouse, France and CNPEN

1 Introduction

With the growing popularity of machine learning systems, it is essential to recognise their role in the transformation of societies and to encourage researchers and industrial stakeholders on the one hand, and decision-makers and public authorities on the other, to take into account the ethical stakes of these technologies.

Numerous recommendations have been issued over the last five years, listing values, principles and criteria[1] to be considered during the development and, more generally, the life cycle of a machine learning system. In all these texts, fairness appears to be a crucial principle. Jobin et al. (2019), studied 84 of them in 2019 and fairness appears to be the second most cited principle after transparency (in 68 texts out of 84). Since then, more initiatives have emerged[2]. For instance, UNESCO (2021) cites *"fairness and non-discrimination"* as a key principle to safeguard social justice, while the OECD (2019) underlines the need for *"human-centred values and fairness"*.

At the European level, the High Level Expert Group on Artificial Intelligence, lists four *"ethical principles"*, including fairness that is broken down into a set of commitments: a "fair" system should *"ensur[e] equal and just distribution of both benefits and costs, [...][ensure] that individuals and groups are free from unfair bias, discrimination and stigmatisation. [...] Equal opportunity [...] should also be fostered [...][and] practitioners should respect the principle of proportionality. [Fairness also] entails the ability to contest [...] decisions made by AI systems."* (HLEG, 2019)

Nevertheless, the recommendations, although paving the way for standardized methods to design algorithms complying with certain values, do not explain how to actually implement these criteria in real-life contexts: what should researchers and engineers do to design "fair" machine-learning based systems? How can the "fairness" of a machine-learning based system be assessed or even proved? What are the concrete criteria?

Machine learning researchers usually think that *"ML systems are biased when data is biased"*[3]. This sentence is representative of a mindset in the machine learning community, sometimes expressed as "garbage in, garbage out". As such, it should be easy to mitigate biases by cleaning the data, rebalancing it or even collecting more of it. This led Google in 2019 to target specifically black people faces to feed their facial recognition network, hoping the biases will simply fade away as a result (Wong, 2019). Scholars have then warned about this phenomenon of *"ex-*

[1] To differentiate ethical values from principles, UNESCO (2021) gives the following definition: *"Values play a powerful role as motivating ideals in shaping policy measures and legal norms. While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily made operational in policy statements and actions."*

[2] See the "National AI policies & strategies" database created by the OECD and the European Commission: `https://oecd.ai/en/dashboards/overview`

[3] Yann LeCun on Twitter in June 2020: `https://twitter.com/ylecun/status/1274782757907030016`

ploiting marginalized groups in the blind pursuit of increasing representation" (Raji et al., 2020).

At the same time, research has been flourishing to allow for the constant improvement of systems' efficiency. In this work, we have chosen to use the term *efficiency* rather than the term *performance* usually used in the literature, because we believe that *performance* should be an umbrella term to refer to various desirable criteria, such as efficiency, fairness, or explainability. *Efficiency* then denotes *"a situation in which a system [...] works well and quickly"*[4]. This definition of efficiency encompasses several notions, notably speed and prediction exactness[5]. The efficiency criterion is often the only one that system designers and developers consider, sometimes not even being aware that another approach is even possible. Indeed, when designing an algorithm, one has to make technical choices as all the desirable criteria, cannot be met at the same time. This leads to trade-offs between values, for instance between efficiency and fairness (Corbett-Davies et al., 2017; Zliobaite, 2015): there is *"no free lunch theorem"* (Wolpert & Macready, 1997). The designer and the developer make these choices according to their own values and habits, which are not neutral and can have a great impact on the individuals that will interact with the resulting system. This implies that algorithm designers and code developers must be held accountable for their design choices (Dignum et al., 2018). We ought to state that a new approach to algorithms and program developments is possible, one that does not rely only on building the most accurate and performative system, but promotes other values, like fairness.

This work focuses on how to design a facial authentication system taking into account both efficiency and fairness when coding, highlighting where trade-offs are necessary. While many definitions of fairness exist (see Section 2.3), it is defined in this work as having the same chances of being recognised by the system, in similar conditions, whoever you are. This implies checking whether people from different subgroups of population (like men or women or groups based on skin color for instance) have the same rate of false positive and false negatives, same accuracy or even if the system has correctly learned on the specific subgroup. We consider the different choices in the design process and how they affect the overall fairness of the system. After detailing in Section 2 the issue of fairness in facial recognition systems and how to define and measure it, we describe the investigated parameters choices as well as the investigated face authentication system in Section 3. Finally, in Section 4, we discuss the overall results. Most importantly, in Section 5, we consider the lessons learned. We notably argue that "performance" should not be a synonym for efficiency but rather a vector of multiple and equally important criteria, including fairness.

[4] `https://dictionary.cambridge.org/us/dictionary/english/efficiency`

[5] "Exactness" should not be confused with "accuracy" which is one of the metrics used for prediction exactness. The accuracy metric is defined in equation 8. For a complete view of our nomenclature, including "performance", "efficiency", "prediction exactness" and "accuracy", see figure 6.

2 Fairness and facial recognition

2.1 Facial authentication systems

2.1.1 How authentication works

Digital facial authentication is defined as the comparison of recorded biometric data with those presented by a person. It is a "one-to-one matching" system: a human face must be recognized as that of a given portrait. Its output is binary: if the output is "yes", the system validates the authentication, otherwise, it refuses it. The most common examples are face recognition systems to unlock a smartphone, enter a public or private place or crossing the border between two countries. In the research community, it can also be called "biometric verification" or "identity match".

A related but different process is *identification* consisting in a "one-to-many" facial recognition process where, for example, a person whose face appear on a photograph must be recognized in a crowd. In this work, we focus on authentication only but identification is also an important process raising complex ethical issues.

For authentication, the decision is based on a scoring function (or similarity measure) s assessing the closeness between two face embeddings[6] z_i and z_j. A decision threshold τ is chosen to classify the pair as genuine (same identity) if $s(z_i, z_j) < \tau$ or impostor (distinct identities) if $s(z_i, z_j) \geq \tau$[7].

FMR and FNMR in the case of facial authentication correspond to respectively the False Positive and False Negative Rate used for binary classification tasks:

$$\begin{aligned} FMR &= \frac{\#\{(z_i, z_j) \in I | s(z_i, z_j) < \tau\}}{\#\{(z_i, z_j) \in I\}} \\ FNMR &= \frac{\#\{(z_i, z_j) \in G | s(z_i, z_j) \geq \tau\}}{\#\{(z_i, z_j) \in G\}} \end{aligned} \tag{1}$$

where G is the set of genuine pairs and I the set of impostor pairs in a given test set. It should be noted that these rates are called respectively False Acceptance Rate (FAR) and False Rejection Rate (FRR) in some studies.

The choice of the threshold τ is crucial because it strongly affects the accuracy and error rates of the system. Indeed, a high threshold would increase the number of matches, but generate more false matches. On the contrary, a too low threshold would prevent some people to be correctly identified. The value of the threshold should thus depend on the use case. As such, some industrial stakeholders choose to let their clients decide on the threshold according to their needs. It is the case for

[6] In a neural network for automated facial recognition, the image of the face is represented at the input by a tensor with the values of each pixel. As the image passes through the network, mathematical processes change its shape. At the output, the image is represented by a vector of dimension n, modelling the image in a n-dimensional space. This vector is called the face embedding.

[7] Note that here, the higher the scoring function s, the less similar the images. This is not always the case in the literature.

instance of facial recognition systems used by police departments that can usually choose this parameter in a range of figures recommended by the provider. While looking for different suspects in a criminal investigation, a high threshold would permit not to miss anyone, but if the system is used to send someone to prison for instance, a lower threshold might help avoid sentencing an innocent person. Without any specific instructions, developers usually try to choose a compromise between FMR and FNMR.

2.1.2 How to train a neural network for facial authentication

Nearly all facial recognition systems use deep neural networks. State-of-the-art networks for image data are called convolutional neural networks (CNNs)[8]. For facial recognition systems, CNNs are used jointly with specific training methods. The two main families of models are siamese networks and triplet loss (Wang & Deng, 2021). Nevertheless, this field of research is constantly evolving and architectures are regularly developed and deserve to be tested. Yet, the triplet loss method remains a quite recent approach and is commonly studied in the literature.

To train a triplet loss model (Schroff et al., 2015), triplets of face images are formed with an anchor image (A), a positive image (P) of the same person, and a negative image (N) of a different person. The loss of the triplet, through the network (f), with an α margin, is given by the function:

$$L : (A, P, N) \rightarrow L(A, P, N) = max(d(f(A) - f(P)) - d(f(A) - f(N)) + \alpha, 0) \quad (2)$$

with $(A, P) \in G$ and $(A, N) \in I$, and d the distance between two embeddings.

It should be noted that d is a norm function[9] that can be similar or different from the s scoring function. Many formula of losses exist and are constantly tested (Wang & Deng, 2021). Among others, the euclidean-distance-based loss, the angular/cosine-margin-based loss or the softmax loss are the most common.

After the loss is computed, the network weights are updated for the distance between A and P to be smaller than the distance between A and N. Thus, the neural network learns to distinguish different clusters representing individuals in a n-dimensional space - with n the output dimension of the network and also the dimension of the embedding vectors.

[8] A CNN is a deep neural network especially used when the input is an image. It is made of an alternation of convolutional and pooling layers. For more information on how CNNs work, see the Stanford course *CS231n convolutional neural networks for visual recognition* at `https://cs231n.github.io/convolutional-networks/`

[9] For the mathematical definition of a norm, see `https://en-academic.com/dic.nsf/enwiki/496296`

2.2 The need for fairness

Automated facial recognition has been strongly criticized for reinforcing overall racial discrimination in a society contaminated by racism and racial prejudice (Bacchini & Lorusso, 2019).

In computer vision, and especially in face recognition, fairness plays an important role as the data used by the algorithm is strongly dependent on protected attributes[10] such as gender or skin color. The first outrage in this field dates back to a video that went viral on the internet, where a black man was not recognized by a face detection camera (CNN, 2009). Since then, many researchers have found that facial recognition algorithms had disparities in prediction exactness when considering different demographics (Phillips et al., 2011; O'Toole et al., 2012; Klare et al., 2012; Cook et al., 2019; Howard et al., 2019; Krishnapriya et al., 2020; Cavazos et al., 2020). This phenomenon, sometimes called *"the other race effect"* (Phillips et al., 2011), was found in different types of algorithms, from older generation systems to new convolutional neural networks systems. Despite these warnings, biased systems were still developed, until MIT researcher Joy Buolamwini and MSR researcher Timnit Gebru, discovered that face analysis systems from big tech companies were misclassifying dark-skinned women much more often than their light-skinned male counterparts (Buolamwini & Gebru, 2018). This was then confirmed by the NIST[11] that conducted a study to quantify the exactness of face recognition algorithms for demographic groups defined by sex, age, and race or country of birth and found significant discrepancies (Grother et al., 2019).

Other companies have experienced similar fairness issues in computer vision applications. For instance, when two people of different skin colors were present in the same picture, the Twitter cropping algorithm would keep the white person more often[12] (Yee et al., 2021). Instagram filters that were supposed to embellish faces were also criticized for whitening the skin of their users (Jerkins, 2015). Google had to apologize for their photo service that was classifying photos of black people with the label "gorillas" (Simonite, 2018). More recently, Robert Williams, an Afro-American man was arrested at his house by the police after being recognized by a surveillance camera equipped with a face recognition software. Not only the man on the footage was seemingly not him, but he still had to struggle to prove it (Hill, 2020b). Despite being the most controversial mistake of that kind, it is far from being the only one (Anderson, 2020; Hill, 2020a).

These issues led many big tech companies such as IBM, Amazon and Microsoft

[10]Protected attributes are features that should not be used by the model for decisions. They are often features protected by law, such as skin color, gender, sexual orientation, religion, disability, and so on.

[11]The National Institute of Standards and Technology is a US organization, in charge of measurements across different technologies. See their website: `https://www.nist.gov/`

[12]This problem was first discovered by a Twitter user: `https://twitter.com/bascule/status/1308147385202167808` and then investigated further in a research paper (Yee et al., 2021).

to announce they were renouncing, for some time, to facial recognition software[13] in particular when used for facial identification. Some other companies, like IDEMIA in France, proudly advertise on their fairness test results (IDEMIA, 2021). Nonetheless, fairness should not be reduced to a false positive and false negative table, nor can biases be reduced to a data issue.

Some have argued that it is a historical issue, deep-rooted in society and linked to diversity issues (Leslie, 2020). Fairness issues in facial recognition can be compared with the one of the Shirley card for color films in the 40s that famously advantaged white people on photographs. The problem was not fixed until the 70s, when advertisers realized the rendering was horrendous for chocolate and wood, reflecting the priorities of the time (Roth, 2009; Caswell, 2015).

2.3 Defining fairness

To design a facial authentication system taking fairness into account, one need to overcome a definition problem: what is a "fair" system? To begin with, each scientific discipline has its own way of defining fairness and one can look at it from a philosophical, legal or mathematical perspective and get diverse definitions (Mulligan et al., 2019). The common meaning of fairness is *"the quality of treating people equally or in a way that is right or reasonable"*[14]. From a legal standpoint, fairness is strongly linked to non-discrimination, a principle enshrined in several texts on human rights. Even within one discipline, several visions of what fairness is can coexist. In machine learning, fairness can be considered as equal treatment for different demographic groups (group fairness) or equal treatment for similar individuals (individual fairness) (Narayanan, 2019). Moreover, there are several ways of considering that an algorithm meets group fairness. The first one is not training the system on the variables that lead to discrimination (like gender or skin color), but even then, discrimination can persist due to correlated variables (Corbett-Davies & Goel, 2018). The second, and undoubtedly the most common one, is to achieve the same prediction exactness on each subgroups, which ensures that the results and risks of errors are independent of the given subgroups (Corbett-Davies & Goel, 2018).

A counterfactual definition of fairness could also be considered, where the results should not change if the same individual is taken but with one attribute change, like their gender or skin color (Kusner et al., 2017). Such a definition can be easily applied on tabular data but it is much more difficult in the case of facial recognition. One way would be to use Generative Adversarial Networks (GANs). GANs are used to create synthetic data from scratch but they can also be used to transfer characteristics from one image to another, for instance capturing the stripes in an

[13] *"In his letter Monday to members of Congress, CEO Arvind Krishna cited the potential for police to use the technology to violate "basic human rights and freedoms" in its decision to end all research, development and production"* (Denham, 2020).

[14] `https://dictionary.cambridge.org/dictionary/english/fairness`

image of a zebra and pasting them on an image of a horse (Zhu et al., 2017). For facial recognition, it can be used to change the apparent gender or skin color of a person (Dash et al., 2022). Yet, GANs are still very complex to process and even more challenging to evaluate.

2.4 Measuring fairness

Several metrics have been developed to try and measure fairness across groups: some authors have identified more than twenty of them (Verma & Rubin, 2018; Narayanan, 2019). They have a broad range of application and are used in various statistical studies, far beyond facial recognition. They broadly fall into three major categories: Independence, Separation, and Sufficiency (Barocas et al., 2021). Independence requires the acceptance rate to be the same in all groups, separation requires that all groups experience the same false negative rate and the same false positive rate, and sufficiency requires a parity of positive/negative predictive values across all groups.

One of the most famous group fairness metric is called statistical parity, demographic parity, or disparate impact (Barocas et al., 2021; Narayanan, 2019). A perfectly fair system under this metric should satisfy the following conditions on probabilities:

$$P(Y = 1|D_i) = P(Y = 1|D_j), \forall i \neq j \tag{3}$$

with Y the output of the system and D the demographic group. In the case of authentication, $Y = 1$ would mean that the authentication is successful.

A model is considered to be biased if significant differences can be observed for different demographic groups of individuals (Drozdowski et al., 2020). This results in a strong gap in the measures between two groups that pushes the probabilities further apart. The extent of a bias is then defined as:

$$|P(Y = 1|D_i) - P(Y = 1|D_j)|, \forall i \neq j \tag{4}$$

These metrics, although widely used, do not account for the contextual dimension of fairness (Selbst et al., 2019; Wachter et al., 2021a; Cheng et al., 2021) and are not always sufficient to guarantee non-discrimination before the law (Wachter et al., 2021b). Additionally, different definitions of fairness may not be satisfied simultaneously as they are mutually incompatible (Kleinberg et al., 2016; Friedler et al., 2016; Zhao & Gordon, 2019). In the end, developers have to choose one metric, or a set of metrics, adapted to their use case. Anahideh, Nezami, and Asudeh (2021) even proposes to help selecting the appropriate fairness notion by detecting correlations between different fairness metrics to reduce the size of this set of metrics. But selecting the right set is a highly sensitive issue, because a wrongly chosen metric could prevent biases detection and lead to rationalization (Aivodji et al., 2019; Weinberg, 2022).

Moreover, when the machine learning community talks about "fairness metrics", they mean it in the restricted sense of a bias measure. A bias is often used to refer to demographic disparities in algorithmic systems, as it is the case in this paper. Note that a statistical bias in the larger sense is a systematic error, meaning that an estimator is biased if its expected or average value differs from the true value that it aims to estimate (Barocas et al., 2021).

Even if these general fairness metrics can still be used on biometrics applications such as automated facial recognition, their independence from the context makes them less relevant, especially because they are not designed to deal with *"multiple failures"*, where different users are affected in different ways (Howard et al., 2022). They often deal with the impact of False Match Rate (FMR) and False Non-Match Rate (FNMR) separately (Pereira & Marcel, 2022), hence the trade-off between the two is not taken into account. The biometrics community has then proposed new metrics, specifically adapted to authentication.

Most methods to measure fairness in biometrics fall into two categories: *differential performance*, considering the difference in distributions between specific demographic groups, or *differential outcome*, considering the difference in FMR or FNMR (Howard et al., 2019). Differential outcome can be seen as of way of achieving statistical parity[15] between groups.

The first metric developed specifically for biometrics is the Fairness Discrepancy Rate (FDR), that calculates the maximum difference in FMR and FNMR between two demographic groups (Pereira & Marcel, 2022). This measure can be integrated to see if FDR is stable across all thresholds. The Area Under FDR can be used as a new metric, closer to 1 when the system is considered more "fair". The second metric, the Inequity Rate (IR) (Grother, 2021) uses the ratio between maximum and minimum FMR and FNMR. Finally the Gini Aggregation Rate for Biometric Equitability (GARBE) (Howard et al., 2022) uses the Gini coefficient[16] instead of the max difference. Variants of these metrics exist, where the ratio is normalized by the mean, or relative to a nominal error rate (Duewer, 2022). Weights can also be applied to the different demographic groups (Duewer, 2022).

But these metrics, despite being better suited for facial recognition, require several non-trivial properties of the data (Howard et al., 2022). To reach their higher potential, several algorithms with the same threshold should be compared on different demographic groups. Only the NIST has the capacities to gather so much data.

Other methods consist in the statistical study of error rates through a bootstrapped-based hypothesis test (Schuckers et al., 2022), ROC curve[17] analysis (Gong et al.,

[15] See Section 2.4

[16] The Gini coefficient is a measure of statistical dispersion often used as a measure of inequality in economics or decision theory.

[17] The Receiver Operating Characteristic (ROC) curve is used to illustrate the predictive capacity of a binary classifier by plotting the true positive rate against the false positive rate at various threshold settings.

2020), or the separation difficulty of different demographic groups (Kotwal & Marcel, 2022).

2.5 Mitigating biases and coding fairness

As seen in Section 2.3, when the machine learning community talks about ensuring fairness in their systems, they usually think about reducing biases. There are various forms of biases (Suresh & Guttag, 2021; Mehrabi et al., 2021; Danks & London, 2017) and they can of course emerge in the data through generation, sampling or measurement, but they can also be embedded in the model and its implementation and appear during the learning process, the evaluation or the deployment (Suresh & Guttag, 2021). Bias reduction has been central to the research initiatives about fairness in machine learning.

Once a bias is noticed in a system, mitigation techniques to correct it can be implemented. There are often classified into three categories (Friedler et al., 2019). Pre-processing methods consist in fixing the imbalances before training the model (e.g. reweighting the data), in-processing ones fix them during training (e.g. adversarial debiasing) and post-processing methods fix them after training by changing the outputs (e.g. the equalized odds algorithm). But these categories can be even more refined (Caton & Haas, 2020). Benchmarks of the said techniques have been realized by other researchers to help a developer choose the most appropriate one for their use case (Friedler et al., 2019).

Yet, model-related biases are often concealed and are dismissed by considering that the quality of the system is strictly restricted to the quality of the data used for it. Because of this belief that unbalanced data alone is responsible for biases, the research community in facial recognition has focused on creating demographically balanced datasets (Hupont & Fernández, 2019; Buolamwini & Gebru, 2018). However, they are rarely used by software companies in practice, but can be useful for academic research. Which dataset to choose is a critical question that should be justified by a thorough evaluation benchmark (Raji & Fried, 2021).

Moreover, facial recognition differs from general machine learning as it is a spatial problem: for the system to work correctly on each subgroup of the population, they should be evenly distributed in the latent space[18]. Researchers are now working on more advanced methods to reach higher scores on the NIST benchmark (Despiegel, 2021) as the improvements due to working only on data have reached a plateau. Recent techniques notably focus on using different loss functions to separate clusters of demographic groups (Conti et al., 2022). Efforts to mitigate biases have also focused on training different systems for each demographics, adapting training parameters (Wang et al., 2022) or changing the network completely for different groups (Vera-Rodriguez et al., 2019).

[18]The latent space, or embedding space, is a representation of data in which similar data points are close together.

The question of whether biases are mainly due to data or if other technical parameters also play a role has not been yet concretely answered by researchers. For facial recognition applications, some works have recently started to tackle the issue: Michalsky (2019) studies the evolution of fairness metrics when changing the test/train split and Atzori et al. (2022) does the same for the authentication threshold. The impact of the architecture of neural networks is also questioned: Krishnan et al. (2020) focuses on the discrepancies in prediction exactness of gender classification systems while Sukthanker et al. (2022) compares rank disparity for different well-known architectures of neural networks.

Sukthanker et al. (2022) may be the closest to our work in that it also reviews the impact of various hyperparameters: the architecture head type, the optimizer type and the learning rate. Out of this list, only the architecture and learning rate are also addressed in our work. But Sukthanker et al. (2022) does not investigate slight changes in the architecture but rather compares very different networks. Similarly, its study on the learning rate does not include modifications of the scheduler.

In the end, these studies usually investigate the impact of a single parameter, not giving a whole overview of the quantitative impact of design and parameters choices. Furthermore, they only consider the impact on fairness metrics, rarely addressing the trade-offs with efficiency. This work aims to contribute to filling this gap, by answering the question of what exactly impacts fairness in the code.

3 Investigated face authentication system and parameters

3.1 General approach

Let us consider a machine learning model for facial authentication. The code used for this system is available[19] and a full report[20] can be found in the same repository.

We consider the choices that a developer makes when coding such a system, evaluate their impacts on the global fairness of the model and highlight the trade-offs with efficiency. To do so, a model called `BaseModel` is built with the parameters that yield the best efficiency[21], as it is usually the case in machine learning development. Then, some of the technical choices that have been made during the development of the model are investigated further. Starting from `BaseModel`, alternative models with some changes are considered. Fairness is then measured on each of these models as well as alternative efficiency measures. We conclude by giving leads on which choice is the best for fairness and for efficiency.

[19] `https://github.com/mgornet/CNPEN`

[20] In French

[21] efficiency metrics are described in Section 3.4

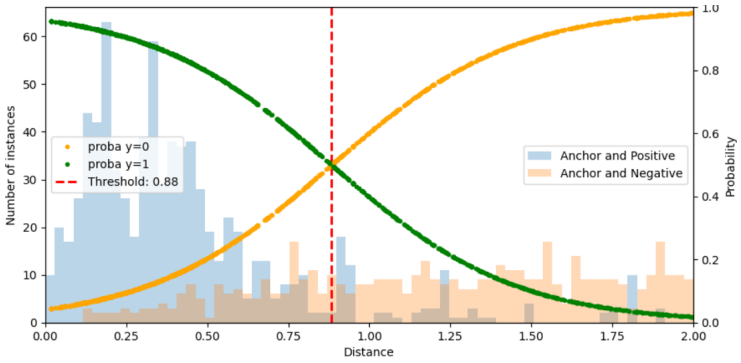

Figure 1: Logistic regression classifier and its intercept

3.2 The system

The system is a CNN, trained by triplet loss for facial authentication (Schroff et al., 2015)[22]. The detailed architecture of the network is given in Appendix A.

For the triplet loss function, an euclidean-distance-based is used, mainly because of its simplicity. The loss formula of equation 2 becomes:

$$L : (A, P, N) \rightarrow L(A, P, N) = max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0) \quad (5)$$

with the same notations as in Section 2.1.2 where distance d corresponds to the squared euclidean norm: $d = \|.\|^2$.

To score the similarity between two face embeddings when the authentication is realized[23], the squared euclidean norm is also used: $s = d = \|.\|^2$.

To determine the authentication threshold, a logistic regression method is trained to find the optimal cut-off value automatically[24]. The intercept is taken as the threshold for the system (see figure 1).

The database used for this work is called Labelled Faces in the Wild (LFW) (Huang et al., 2007). It contains images of famous people, taken in an everyday environment and not having specially posed for the occasion. The database and corresponding paper can be found on the authors' website[25]. The database is known to be unbalanced[26] but we have kept it in particular to see whether relevant choices of

[22] See Section 2.1.2 to learn how this type of model is generally trained.

[23] See Section 2.1.1

[24] A logistic regression method is a classifier, looking for correlations between variables y and X, with y the dependent variable (here it is binary and corresponds to the match or non-match classification), and X the independent variables, which in our case boils down to a single variable: the opposite of the distance between two images. The binary variable y is then transformed into a probability. Finally, logistic regression looks for the maximum likelihood, i.e. finds the minimum point of difference between what the classifier predicts and the reality. This search for the minimum is carried out by gradient descent. The intercept corresponds to the maximum likelihood.

[25] `http://vis-www.cs.umass.edu/lfw/`

[26] The LWF dataset is estimated to be 77.5% male and 83.5% white (Han & Jain, 2014)

parameters could mitigate this issue.

To measure fairness, as laid out in Section 2.4, data needs to be separated in different groups of people, based on protected attributes. For these attributes, the labels provided with the dataset are kept. These labels were computed automatically as described in the authors' paper (Kumar et al., 2009). The results are thus approximated. For instance, a visual check shows that some labels do not always correspond to the image: for instance many male-presenting individuals remain in the non-male category, possibly affecting the prediction exactness of the training phase.

The choice of which demographic groups to compare is dependent on which labels are provided. In LFW, they include, for ethnicities[27]: "white", "black", "asian", "indian", "none". We have kept the first two categories to refer to white and dark skins. LFW also includes for gender: "male" and "non-male"[28].

Three sets of analyses are carried out as followed: (i) gender: male vs female; (ii) skin color: white vs non-white vs black; (iii) intersectional: white male vs non-white female vs black female.

Here the non-white group includes the three other categories so called in the dataset: black, asian, indian. These categories are not separated in order to get enough data for statistical analysis.

3.3 Investigated parameters

The parameters choices that we have investigated are the following ones:

1. Data processing
   (a) Data sampling
   (b) Data normalization
   (c) Data augmentation
2. Neural network
   (a) Depth of the network
3. Training
   (a) Margin for the loss function

[27]Note that some countries prohibit ethnicity statistics. In France, Decision Nr. 2007-557 DC of 15 November 2007 of the Constitutional Council prohibits the implementation of processing operations necessary for the conduct of studies on the measurement of diversity. These processes are deemed contrary to the first article of the French Constitution that states the *"equality before the law of all citizens without distinction of origin, race or religion"*.

[28]In this work, "non male" is considered as an equivalent to female. Even if that is not necessary true, the label was intended this way: male if the variable "male" is equal to one, female if it is equal to zero.

(b) Base learning rate and scheduler

4. Evaluation

   (a) Distance threshold

3.3.1 Data sampling

For the sampling of `BaseModel`, epochs are constructed so that every identity[29] is represented only once. Because the database is unbalanced, this is a way of forcing even under-represented identities to show up at least once during the training. An alternative method is to use random sampling, that is to say, to pick two images of the same identity at random a certain number of times. But as there are more white men in the database, random sampling will most likely pick them more often than not. This alternative is explored further in `RandomSampling`.

3.3.2 Data normalization

Regarding normalization, the model would train more easily if the value of the pixel were in [0,255][30]: we have kept these values for `BaseModel`. Yet, a common practice is to normalize the data in [0,1]: this possibility is explored in `Normalized`.

3.3.3 Data augmentation

Various data augmentations are also tested. `BaseModel` consists of an 50% chance random horizontal flip, an up to 20% zoom with an up to 0.5% deformation at a probability of 90%, an up to 10% random color jitters with a probability of 70% and an up to 5 degrees random rotation.

For the derivative models, one is computed with no augmentation at all: `NoAugment`. Other derivative models consist in different transformations: `HighJitter` has double the color jitters percentage and rate of occurrence of `BaseModel`, `HighDeformation` doubles the deformation, `HighRotation` doubles the rotation and `HighZoom` has a zoom of 10 to 30%.

Intuitively, data augmentation should have an impact on fairness as strong data augmentation could destroy some identities, making the faces unrecognizable. For instance, darkening a picture could worsen the image quality of a black person, while a higher brightness could have the same effect for white person.

[29] An identity is understood here as a representation of a specific person. Each person is represented in the database by one or several facial images. At each epoch, one of these images is drawn randomly for each person.

[30] The numerical representation of an image is usually a tensor of size l*h*c, with l and h being the dimensions of the image in pixels and c the number of channels. In LFW database, the number of channels is 3 for the red, green and blue channels, and the images where preprocessed to get a 60*60 size. Each pixel in this tensor is an integer representing the intensity of the color at the specific location. Initially in the database, these values span from 0 to 255 but other databases normalize them to [0,1] or [-1,1] depending on the intended use.

3.3.4 Depth of the network

An analysis of choices that are not directly linked to the data but rather to the model itself are also included. Instead of comparing various well-known architectures, a simple CNN is used. Derivative models are computed with different sizes for the hidden layers, to measure the impact of the depth of the network. In the end, when the input goes through all of the layers, the length of the base channel is multiplied by 16, meaning that for a base channel of 32, as in `BaseModel`, the embedding will have a length of 512. This architecture provides a sufficient depth to fully learn on the data without being too complex and time consuming during training. Two alternative models have been computed for base channels of 16 (`Depth16`) and 64 (`Depth64`).

3.3.5 Margin for the loss function

Fairness constraints are not added to the training, but instead, common training parameters are studied. The first parameter is the margin for the loss function. For `BaseModel`, the 0.2 value given by the initial paper (Schroff et al., 2015) is kept. The fairness of the model is then studied for derivative models with a lower or higher margin: 0.1 in `Margin01`, 0.5 in `Margin05` and 1. in `Margin1`.

3.3.6 Learning rate and scheduler

The second training parameter investigated is the learning rate and its step scheduler. `BaseModel` network trains with a base learning rate of 0.0005 until epoch 200 and then each 100 epochs, this base value is divided by two.

`BaseModel` is compared with $\texttt{Lr10}^{-3}$ trained with a base learning rate of 0.001, $\texttt{Lr10}^{-4}$ trained with a base of 0.0001, `Scheduler300` that keeps the base rate but starts to divide by two at the 300th epoch and `Scheduler100` that starts to divide at the 100th.

3.3.7 Distance threshold

Finally, the choice of the distance threshold, which determines who gets rejected or accepted, is investigated. The accuracy as a function of the threshold can be computed and it can be verified that the logistic regression finds the threshold corresponding to the maximum of accuracy. In Figure 2, a maximum of accuracy of 86% is reached for a threshold of 0.81: it is the value used for `BaseModel`. But if we only want to maintain a certain level of accuracy, like 95% of the maximum value, very different thresholds can be used. An accuracy of 82%, i.e. 95% of the maximum accuracy, is reached for a threshold of 0.48 or a threshold of 1.15. These two alternatives are tested in two models respectively called `LowThreshold` and `HighThreshold`.

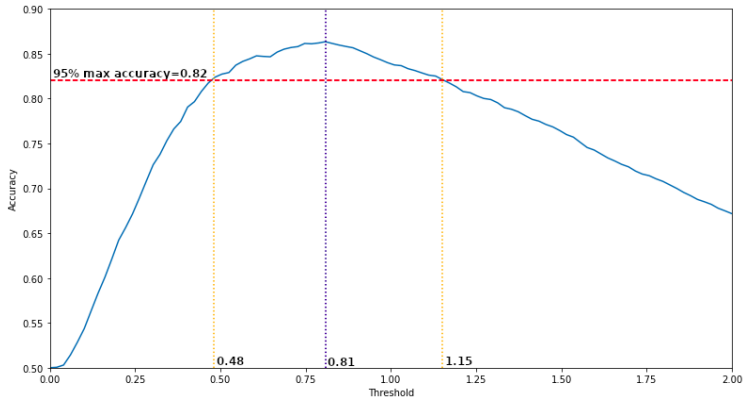

Figure 2: Model accuracy as a function of the distance threshold

Investigated choice	**List of models**
Data sampling	`BaseModel`; `RandomSampling`
Data normalization	`BaseModel`; `Normalized`
Data augmentation	`BaseModel`; `NoAugment`; `HighJitter`; `HighDeformation`; `HighRotation`; `HighZoom`
Depth of the network	`BaseModel`; `Depth16`; `Depth64`
Margin	`BaseModel`; `Margin01`; `Margin05`; `Margin1`
Learning rate	`BaseModel`; $\texttt{Lr10}^{-3}$; $\texttt{Lr10}^{-4}$; `Scheduler300`; `Scheduler100`
Threshold	`BaseModel`; `LowThreshold`; `HighThreshold`

Figure 3: Table of the different models ordered by investigated choices

All derivative models and their respective coding choices are summarized in Table 3.

3.4 Metrics

To quantify the impact of the choices, measures of efficiency and fairness need to be defined.

3.4.1 Choosing `BaseModel` parameters and comparing models on efficiency

The model selection is based on the final validation loss. The best model according to this criterion is the one that, at the end of the training phase, gets the lowest validation function value. In industrial settings, other metrics can be used

for this measure of efficiency such as accuracy or the time taken by the model to train. What metric to use is in the end a policy choice made by the developer or the company itself. Appendix C shows the evolution of the validation loss during training for `BaseModel` and for derivative models.

Once the `BaseModel` parameters are selected on this validation loss metrics, all the different models are also tested on various efficiency metrics: accuracy, FMR, FNMR, and a new coined metric, the Triplet Learned Rate (TLR). These results are shown in Appendix B. Note that this is a study on the overall dataset, the results on different subgroups are considered as part of the fairness study.

To define the accuracy, FMR and FNMR, we first define the True Matches (TM), True Non Matches (TNM), False Matches (FM) and False Non Matches (FNM) as followed:

$$\begin{aligned} TM &= \#\{(z_i, z_j) \in G | s(z_i, z_j) < \tau\} \\ TNM &= \#\{(z_i, z_j) \in I | s(z_i, z_j) \geq \tau\} \\ FM &= \#\{(z_i, z_j) \in I | s(z_i, z_j) < \tau\} \\ FNM &= \#\{(z_i, z_j) \in G | s(z_i, z_j) \geq \tau\} \end{aligned} \tag{6}$$

with the same notations as Section 2.4. These metrics correspond respectively to the True Positives, True Negatives, False Positives and False Negatives in binary classification tasks. The FMR and FNMR of equation 1 can then be redefined as:

$$\begin{aligned} FMR &= \frac{FM}{FM + TNM} \\ FNMR &= \frac{FNM}{FNM + TM} \end{aligned} \tag{7}$$

Moreover, the accuracy is defined as:

$$accuracy = \frac{TM + TNM}{TM + TNM + FM + FNM} \tag{8}$$

Note that other metrics that measure prediction exactness could be investigated, like the precision, recall or f1-score[31] but the information they contain is redundant with the accuracy, FMR and FNM.

Finally, we define the TLR, a new metric specific to the use case of this work. The TLR counts the ratio of triplets for which the distance between the anchor and positive images is shorter than the distance between the anchor and negative images. This allows to approximate how well the system has learned:

$$TLR = \frac{\#\{(A, P, N) | s(A, P) < s(A, N)\}}{\#\{(A, P, N)\}} \tag{9}$$

[31] $precision = \frac{TM}{TM+FM}$, $recall = \frac{TM}{TM+FNM}$, $f1 - score = \frac{2 * precision * recall}{precision + recall}$

3.4.2 Comparing models on fairness

To measure fairness, the chosen definition is similar to that of group fairness and equalized odds in statistics, also called "differential outcome" in biometrics[32], i.e. to have the same probability of being accepted or rejected whatever the group of the person is. As proxies for this probability, several metrics are used and should yield the same results for any group if fairness is respected.

We use for these metrics the same as the ones used to assess efficiency: accuracy, FMR, FNMR and TLR. Yet, this time, they are considered on subgroups of the dataset.

Fairness is respected when the metrics are equal for any group:

$$\begin{aligned} accuracy_{D_i} &= accuracy_{D_j}, \forall i \neq j \\ FMR_{D_i} &= FMR_{D_j}, \forall i \neq j \\ FNMR_{D_i} &= FNMR_{D_j}, \forall i \neq j \\ TLR_{D_i} &= TLR_{D_j}, \forall i \neq j \end{aligned} \tag{10}$$

Yet, this condition is almost impossible to reach. Instead of defining an arbitrary ϵ that would act as a maximum difference between groups, a 90% confidence interval for each metrics is computed thanks to a bootstrap[33] on the test set.

A high prediction exactness within a group is then characterized by a confidence interval that will be small and close to 1 for accuracy and TLR, and close to 0 for FMR and FNMR. A high fairness will be characterized by small confidence intervals that overlap between groups or for which mean values are close. The discrepancy between two groups is considered significant, and thus there is a bias in the model, if the confidence intervals do not overlap.

4 Results

4.1 Analysis of `BaseModel`

4.1.1 Efficiency

`BaseModel`'s overall efficiency can be found in Appendix B. It was inherently built to yield the best possible results in terms of validation loss. Other metrics of efficiency can be verified such as accuracy, error rates, TLR or the time the model has taken to train. The mean accuracy is high enough to say that the system works well; without being at the state-of-the-art, it is sufficient for the statistical analysis. Moreover, FMR and FNMR are quite balanced (0.157 and 0.100 respectively).

[32] See Section 2.4

[33] Bootstrapping is a common method in machine learning and statistics which consists in using a collection of random samples from a population to infer results on this population.

$TLR = 0.93$ signifies that for 93% of the triplets in the test data, the distance between the anchor and positive is indeed lower than the distance between the anchor and negative; therefore the training went well and the system is able to generalize.

4.1.2 Fairness

Regarding fairness, as expected, the system displays some biases. But accuracy is not as expected: for instance, the mean accuracy is significantly better for females than for males, and it is slightly better for white people than for non-white or black people. This result is even more exacerbated for the intersectional study.

This is mostly due to the fact that accuracy is a poorly chosen metric for this problem and that FMR and FNMR metrics need to be investigated. For example, for women or black people, FMR is extremely high but this is offset by a very low FNMR, which results in a high accuracy overall. This high FNMR for minority groups[34] is perhaps the most problematic bias in the system. The rest of our parameters analysis will thus focus on this metric.

As for TLR, it is significantly better for the male, white and white male groups with respect to the female, non-white and non-white female groups, showing that the learning process is, as expected, more efficient for over-represented groups in the database.

Moreover, some subgroups are too small to compute statistical results. This is the case for the FNMR of black females. Similarly, the confidence interval for the TLR is often quite large for black people and for black female as the bootstrap captures too little data.

4.2 Impact of parameters choices on fairness and efficiency

The complete quantified results for each parameter choice can be found in Appendix B for efficiency and Appendix C for fairness. Additionally, Figure 4 displays the FNMR confidence intervals of different groups for `BaseModel` and one alternative model for each parameter choice. Figure 5 displays the mean differences between majority groups, respectively Male, White and White Male, and minority groups, respectively Female, Non-White and Non-White Female, for these models.

4.2.1 Data sampling

The gender analysis reveals that interval gaps between minority and majority groups stay significant for both `BaseModel` and `RandomSampling`: biases are present for the same metrics. On the other hand, `RandomSampling` introduces a bias on

[34]Here we use the term "minority" as there is less data for those groups. Another definition can be that minority groups are the ones that usually experience discrimination. Here those two definitions coincide for the following groups: "female" (as opposed to "male"), "non-white" and "black" (as opposed to "white"), "non-white female" and "black female" (as opposed to "white male").

accuracy in the skin color analysis but to the benefit of minority groups: accuracy is significantly better for non-white than for white people. Yet, because accuracy is a poorly chosen metric, this bias is not of great importance if it is not followed by changes in the FMR or FNMR metrics. The intersectional study reveals that the bias on FNMR was corrected with `RandomSampling`. Looking at Figure 5, even if the biases were not corrected in the gender study and in the skin color study, the intervals between means of majority and minority groups are reduced, showing that the system is less biased.

In terms of efficiency, `RandomSampling` has a significantly higher accuracy and TLR for the male and white groups, as well as a lower FMR. For the white male group, only FMR is improved. But this gain in efficiency is not restricted to majority groups. For the female group, all four metrics seems to be slightly better than for `BaseModel`. This phenomenon is also present in the skin color analysis where `RandomSampling` yields better results in accuracy and FMR than `BaseModel` for the non-white group. For the intersectional study, the accuracy, FMR and FNMR are better for the non-white female group.

A non-random sampling was chosen during model selection[35] for its lower validation loss, but in terms of prediction exactness metrics, `RandomSampling` seems to surpass `BaseModel`. This might be due to the fact that random sampling allows for more variate pictures to show. On the contrary, `BaseModel`, by forcing the sampling of all identities, yields at every epoch the same picture for some identities that are underrepresented in the dataset, causing overfitting. Yet, `RandomSampling` has a much higher training time, which could also undermine being chosen in an industrial setting. If, according to the efficiency metrics, both `BaseModel` and `RandomSampling` could be elected best efficient model, the best model for fairness is undoubtedly `RandomSampling`, suppressing a key bias and improving the average efficiency of minority groups.

4.2.2 Data normalization

Normalization does not seem to have a huge impact on fairness as there is neither biases created nor avoided with `Normalized`. Figure 5 shows that the gap between means is even greater for `Normalized`.

Moreover, the efficiency measures are worse for `Normalized`, especially for majority groups. For minority groups, the drop in efficiency is not as significant as confidence intervals overlap, as seen on Figure 4 for instance, but metrics are still slighly worse.

It is not far fetched to say that `BaseModel` is the best regarding efficiency. Regarding fairness, the two models are not that different, but at least regarding FNMR, `BaseModel` seems less biased.

[35] See Section 3.3 to understand `BaseModel` sampling

4.2.3 Data augmentation

For the most part, the models tested here display the same biases as `BaseModel`, with a few notable exceptions: `HighDeformation` removes the bias on accuracy for the gender study, `HighZoom` creates a bias in accuracy for the skin color one, and `NoAugment` removes the bias on TLR for the intersectionary study. Consequently, the latter is the most valuable change as it affects another metric than accuracy. Yet, Figure 5 reveals that the gaps in FNMR are larger for `NoAugment` than for `BaseModel`.

`NoAugment` has lower efficiency for the three analyses. And once again, the minority groups are somewhat affected by this decrease in efficiency, particularly they have a lower accuracy and higher FNMR, but the most affected are the majority groups that suffer from retrogression on every metrics.

If `BaseModel` is, here, the best for efficiency, each derivative seems to behave differently. `NoAugment` gets rid of one bias but at the cost of a greater gap in FNMR. One could then considered that `BaseModel` remains the better choice in this case.

4.2.4 Depth of the network

If `Depth16` and `BaseModel` behave similarly, there is a drop in efficiency for `Depth64`. It does not have a concrete impact on biases, except for creating a small one on accuracy for the skin color analysis. Yet, as for data normalization and augmentation, it seems to decrease efficiency for the white group, reducing the gap with the non-white group. But for FNMR in the intersectional study, `Depth64` actually reduces the gap between groups, as seen on Figure 5.

There is no direct answer for which model is the fairer. If a fair model is defined as having the best results possible for minority groups, `BaseModel` is better. But if a fair model is the one displaying the smaller gap between groups, as we have previously stated[36], then `Depth64` is fairer. Both decisions are legitimate and cause a dilemma. `BaseModel` improves the efficiency of everyone overall, even minority groups which will then experience less wrong algorithmic decisions this way but to the expense of majority groups experiencing them even less. `Depth64` reduces the gap so that everyone would experience wrong algorithmic decisions but to the expense of these wrong decisions being even more frequent.

4.2.5 Loss function margin

Regarding the margin choice, the most significant bias suppression is in the intersectional study on FNMR for `Margin1` and `Margin05`. `Margin05` decreases the efficiency on majority groups for all studies: this does not result in other bias suppression but helps bridging the gap slightly. For the skin color and intersectional

[36] See Section 3.4

studies, the FNMR gap is actually reduced by choosing `Margin05` as displayed on Figure 5.

The best model for efficiency is again `BaseModel`. But as for the choice of depth, a derivative model for the choice of margin, such as `Margin05`, helps reduce the gaps between groups and can be considered a fairer model. It is in this case closely followed by `Margin1`, a more balanced choice if a trade-off were to be made.

4.2.6 Learning rate

A surprising result is that derivative models like $\texttt{Lr10}^{-3}$, `Scheduler100` and `Scheduler300` seem to suppress biases on accuracy for the gender analysis but $\texttt{Lr10}^{-3}$, `Scheduler100` also create biases on the skin color analysis. The intersectional analysis is more balanced, with some models like $\texttt{Lr10}^{-3}$ and `Scheduler100` suppressing biases, and the other models behaving just as `BaseModel`. No other metric than accuracy is affected by these biases changes, making it hard to decide which one could be the best model. It can be noted that the other efficiency metrics of some derivative models are often worse than for `BaseModel`, notably for $\texttt{Lr10}^{-3}$. Drops in efficiency can be observed for both majority and minority groups. In the case of $\texttt{Lr10}^{-3}$, the gap in FNMR is larger, as seen on Figure 5.

It can be observed that each derivative model deteriorates a efficiency metric other than accuracy for a minority group in at least one of the studies. It can be argued that the fairest model is in this case the most efficient, that is to say, `BaseModel`.

4.2.7 Authentication threshold

As mentioned in Section 3.3, the choice of threshold, even before realizing a group study, is already a fairness choice as it directly impacts FMR and FNMR. Choosing to improve one metric over the other is a decision that should be taken only after cautious considerations.

This property of the threshold influences the error rates across groups as expected. But these changes are not homogeneous between groups. If for most groups, a high threshold means a higher FMR and lower FNMR, the non-white group does not have a significant improvement of FNMR. Similarly, if a low threshold usually means a lower FMR and higher FNMR, FMR is not significantly improved for the female group. We can also note that both error rates are slightly deteriorated for the black group when we increase or decrease this threshold. This can be an issue if we consider a system for police forces used to search for possible suspects in a database. If it is shown that minorities are showing up more often, then a solution could be to reduce the threshold. Yet, this would likely reduce majority groups FMR and not minorities FMR. On the contrary, if a facial authentication system for border crossing was criticized for regularly rejecting people from minority groups

and customs authorities decided to raise the threshold, it could still result in wrong rejections for those people while improving the majority error rate.

For this parameter, an adequate answer on the fairest choice is not possible as it is strongly linked with the context of deployment. In our specific case, we do not have this context so we can only reason on the empiric results. In `BaseModel`, the system has a higher FNMR than FMR for minority groups. Even if, as we have seen, increasing the threshold will not totally solve this issue for minority groups, it can still help balancing the results. The results on Figure 5 show that choosing `HighThreshold` will not particularly reduce FNMR gaps between groups but it will drop the mean FNMR for all groups.

4.3 Results summary

We have chosen to analyse four metrics, accuracy, FMR, FNMR and TLR, but with a focus on FNMR, which can have great impact on individuals. The results displayed in Figure 4 show that FNMR is clearly higher for minority groups, with a greater uncertainty as the 90% interval is larger. Some models, like `HighThreshold` or `RandomSampling` yield lower FNMR but to the cost of a higher FMR.

This work shows that under specific conditions, the coding choices that are made can impact fairness and that data is not solely responsible for the presence of biases in a model. Indeed, different models have been computed that only differ by one parameter's value. These models all yield different results in terms of fairness and efficiency. For some models, the best model for efficiency is also the best for fairness. But many times, there is a trade-off with efficiency, or a different interpretation of the fairest choice to made, depending on your definition of fairness. Even if system developers decide to make their system the fairest possible, there is not always a straight answer. The decision to make depends strongly on the context of deployment and on the values of the developer.

The results of this work need to be recomputed for each system and one cannot conclude the superiority of some parameter value over another in a general case. Indeed, neural networks are inherently stochastic, which prevents the generalization of the results besides the exact experimental conditions. This work can thus be reproduced if trained on the same network, same seed and same processor[37]. To ensure the reproducibility of our work, we provide access to our algorithm for training and testing facial authentication systems, as well as to the different models that have been studied in this project[38].

Our contribution is to provide a new way of considering the design of machine learning systems, questioning each and every decisions and showing how this mindset

[37]The Pytorch documentation reads: *"Completely reproducible results are not guaranteed across PyTorch releases, individual commits, or different platforms. Furthermore, results may not be reproducible between CPU and GPU executions, even when using identical seeds."* (`https://pytorch.org/docs/stable/notes/randomness.html`)

[38]`https://github.com/mgornet/CNPEN`

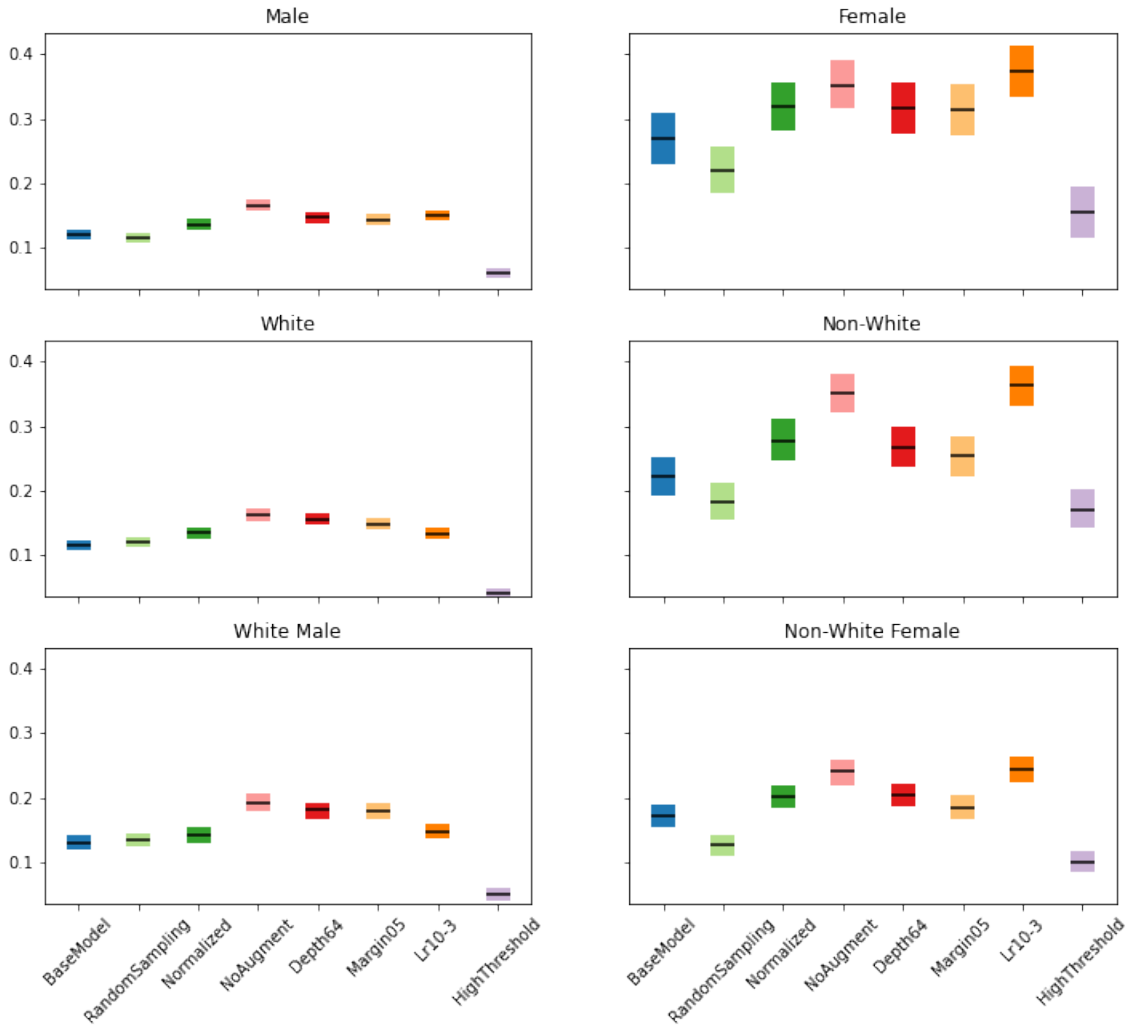

Figure 4: FNMR confidence intervals of the different models for different subgroups

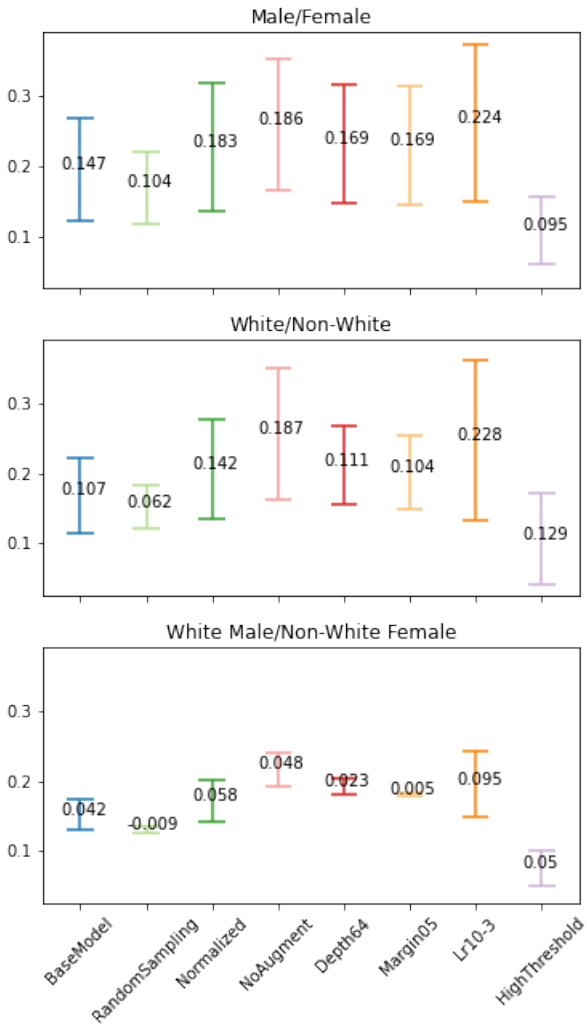

Figure 5: Discrepancy between means in FNMR for each study

can be applied on a use case. Developers need to experiment by themselves the effects of design choices on the fairness of their own systems.

The facial authentication system `BaseModel` is itself a combination of several design preferences: it is a CNN, trained with triplet loss, more specifically, with an euclidean-distance-based loss, using the unbalanced LFW dataset and its attributes. These design preferences have not been investigated further and this work rather focuses on seven of more technical choices, ranging from data processing to training parameters and network architecture. Some other technical choices are discussed in our report[39]: the loss function, regularization, optimizer, dropout, earlystopping or the negative sampling. These choices as well as the design preferences could be investigated in future work.

Another extension of this work could focus on multiple choice analysis. Indeed, we have stated for instance that `RandomSampling` and `Depth64` were both better for fairness than `BaseModel`, but what about a model with both random sampling and the architecture of `Depth64`? One could perform a set of tests, in the manner of a grid search[40]. But of course, as every model needs to be trained for each combination of choices, the cost in time and computational energy of comparing all of them is extremely high.

To counterbalance the performative approach, this work focuses on the value of fairness, but the impact of technical choices on other values like explainability or human autonomy could also be investigated in future work.

5 Lessons learned

5.1 Most design choices are not neutral

The main take away of this work might be to reaffirm that design choices are not neutral and thus, that developers are accountable for each of their decisions. There is a will to improve fairness in the development phase, yet often datascientists still see it as an additional constraint, the main objective remaining efficiency. Moreover, they are not always aware of the impact of the design choices they make, openly advertising for fairness and ethics but blaming the lack of precise regulation. However, this mindset is not sufficient to hold a proper ethical reasoning and each stakeholders needs to take accountability for their contribution to the system's behavior.

5.2 There is no perfectly fair system

If it was already known that fairness mathematical criteria cannot be met all at once (Kleinberg et al., 2016; Friedler et al., 2016; Zhao & Gordon, 2019), this work

[39] `https://github.com/mgornet/CNPEN/tree/main/tech\%20report`

[40] A grid search is a technique used for optimizing the parameters of a machine learning model by looking at all the combinations possible.

additionally shows that different systems can be considered more fair, depending on what "fair" means. Different stakeholders might not agree then on which system to choose.

Some hard choices might include dilemma as whether to decrease the efficiency of majority groups to bridge the gap with minority groups when improving efficiency on the latter is too difficult or impossible. Of course a perfect solution would be to improve the efficiency of the system on minority groups to be as high as for majority groups, and this should be the main goal of algorithmic fairness research. Yet, we should be aware that this result is hard to achieve and that in the mean time, one must choose between several, less perfect solutions.

5.3 Fairness as a dimension of performance

There is no single measure of efficiency, just as there is no single measure of fairness. Some models studied in this work, such as `RandomSampling`, are not the best ranking while evaluated on the specific metric of validation loss. Yet, it surpasses `BaseModel` when it comes to mean accuracy or error rates.

This phenomenon also exists for fairness as some models can be considered fairer according to one definition but not to all. For instance, `Depth64` is fairer than `BaseModel` if fairness is regarded as a discrepancy between groups as it is often the case. Yet, if fairness is regarded as the reduction of harmful outcomes, then the fairest model may be the one that performs best for minority groups, independently of majority groups and discrepancies between group, leading to preferring `BaseModel`.

A possible solution is to consider fairness and efficiency not as antagonistic criteria, but as parts of the same vector of performance that would include various desirable criteria. Each of these criteria could then be broken down into several metrics to measure it. For instance, efficiency can include measures of prediction exactness such as mean accuracy or error rates, measures of how well the model has learned, like validation loss or TLR, measures of speed such as effective training time or computational complexity. Fairness includes measures of group fairness and individual fairness. A more complete vector could also include measures for every "ethical principle" that can be found in international texts, with measures of explainability, transparency, human autonomy, respect for the environment, and so on. Figure 6 summarizes our view of what a performance vector should be.

5.4 The need to address trade-offs

As this work shows, trade-offs are an important part of designing a system. Yet, international recommendations about "the ethics of AI" hardly mention that all the proposed criteria cannot be met at the same time and that trade-offs are often necessary.

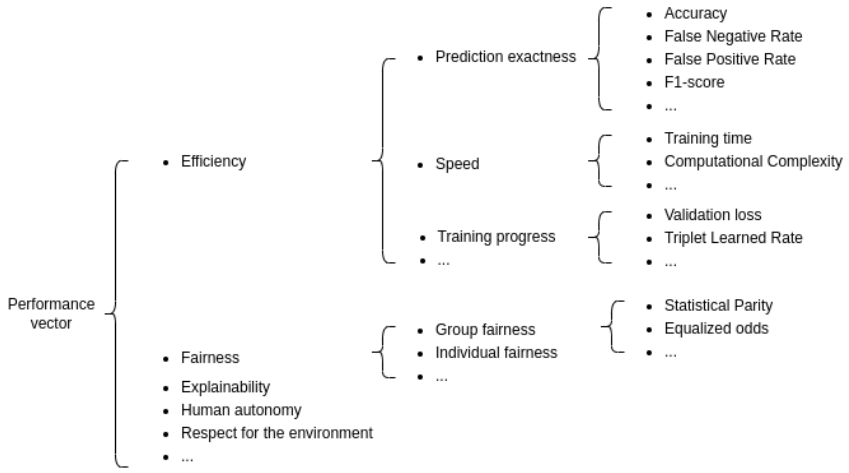

Figure 6: Performance vector

There is a need for more work focusing on choices and trade-offs, as well as ethical thoughts involving all the stakeholders. These initiatives should be set before the design of machine leaning systems, making the conflicts explicit and guiding the decisions concerning the code implementation as well as the main decision to deploy or not such digital processes. Engineering ethics issues should be addressed at all stages of machine learning systems design and implementation. Identifying, analysing and managing trade-offs are important tasks when designing and implementing digital systems. This should be made fully explicit and methodological processes as well as digital tools to facilitate these fundamental tasks are yet missing and should be considered in the near future.

Acknowledgements

Many thanks to Olivier Grisel for his help in understanding the recent developments of machine learning techniques applied to image recognition and to our colleagues of the national pilot committee for digital ethics (CNPEN) for the many discussions when setting up our opinion on the ethical issues of facial recognition, to be published later in 2023.

References

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 161–170.

Anahideh, H., Nezami, N., & Asudeh, A. (2021). On the choice of fairness: Finding representative fairness metrics for a given context. *arXiv:2109.05697*.

Anderson, E. (2020). Controversial detroit facial recognition got him arrested for a crime he didn't commit. *Detroit Free Press*.

Atzori, A., Fenu, G., & Marras, M. (2022). The more secure, the less equally usable: Gender and ethnicity (un)fairness of deep face recognition along security thresholds. In *The 2nd International Workshop on Artificial Intelligence Methods for Smart Cities (AISC 2022)*, Leuven, Belgium.

Bacchini, F., & Lorusso, L. (2019). Race, again: how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society*, *17*(3), 321–335.

Barocas, S., Hardt, M., & Narayanan, A. (2021). *Fairness and Machine Learning*. fairmlbook.org.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Caswell, E. (2015). Color film was built for white people. here's what it did to dark skin.. *Vox*.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv:2010.04053*.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *arXiv:1912.07398*.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.

CNN, B. M. S. (2009). Hp looking into claim webcams can't see black people - cnn.com..

Conti, J.-R., Noiry, N., Clemencon, S., Despiegel, V., & Gentric, S. (2022). Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 4344–4369.

Cook, C. M., Howard, J. J., Sirotin, Y. B., Tipton, J. L., & Vemury, A. R. (2019). Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *1*(1), 32–41.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, Halifax NS Canada.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4691–4697, Melbourne, Australia.

Dash, S., Balasubramanian, V. N., & Sharma, A. (2022). Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924.

Denham, H. (2020). Ibm's decision to abandon facial recognition technology fueled by years of debate. *Washington Post*.

Despiegel, V. (2021). Fairness for face recognition. European Association for Biometrics (EAB) virtual events series.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M. S., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Toree, L., Villata, S., & de Wildt, T. (2018). Ethics by design: Necessity or curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 60–66, New York, NY, USA.

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, *1*(2), 89–103.

Duewer, D. L. (2022). Face recognition vendor test (frvt) part 8: Summarizing demographic differentials. Tech. rep. NIST IR 8429, National Institute of Standards and Technology.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv:1609.07236*.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338, New York, NY, USA.

Gong, S., Liu, X., & Jain, A. K. (2020). Jointly de-biasing face recognition and demographic attribute estimation. In Vedaldi, A., Bischof, H., Brox, T., & Frahm, J.-M. (Eds.), *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science. Springer International Publishing.

Grother, P. (2021). Demographic differentials in face recognition algorithms. European Association for Biometrics (EAB) virtual events series.

Grother, P. J., Ngan, M. L., & Hanaoka, K. K. (2019). Face recognition vendor test part 3: Demographic effects. Tech. rep..

Han, H., & Jain, A. K. (2014). Age, gender and race estimation from unconstrained face images. Msu technical report MSU-CSE-14-5, Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA.

Hill, K. (2020a). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*.

Hill, K. (2020b). Wrongfully accused by an algorithm. *The New York Times*.

HLEG (2019). Ethics guidelines for trustworthy ai. Tech. rep., European Commission.

Howard, J. J., Laird, E. J., Sirotin, Y. B., Rubin, R. E., Tipton, J. L., & Vemury, A. R. (2022). Evaluating proposed fairness models for face recognition algorithms. *arXiv:2203.05051*.

Howard, J. J., Sirotin, Y. B., & Vemury, A. R. (2019). The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep. 07-49, University of Massachusetts, Amherst.

Hupont, I., & Fernández, C. (2019). Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*.

IDEMIA (2021). Idemia's facial recognition ranked #1 in nist's latest frvt test..

Jerkins, M. (2015). The quiet racism of instagram filters. *Racked*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807*.

Kotwal, K., & Marcel, S. (2022). Fairness index measures to evaluate bias in biometric recognition. In *International Conference on Pattern Recognition (ICPR)*.

Krishnan, A., Almadan, A., & Rattani, A. (2020). Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1028–1035.

Krishnapriya, K. S., Albiero, V., Vangara, K., King, M. C., & Bowyer, K. W. (2020). Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, *1*(1), 8–20.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372, Kyoto.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

Leslie, D. (2020). Understanding bias in facial recognition technologies. Tech. rep., The Alan Turing Institute.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Michalsky, F. (2019). Fairness criteria for face recognition applications. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 527–528, Honolulu HI USA.

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–36.

Narayanan, A. (2019). 21 fairness definitions and their politics. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).

OECD (2019). Recommendation of the council on artificial intelligence. Tech. rep..

O'Toole, A. J., Phillips, P. J., An, X., & Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, *30*(3), 169–176.

Pereira, T. d. F., & Marcel, S. (2022). Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(1), 19–29.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception*, *8*(2), 14:1–14:11.

Raji, I. D., & Fried, G. (2021). About face: A survey of facial recognition evaluation.. AAAI 2020 Workshop on AI Evaluation.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151, New York, NY, USA.

Roth, L. (2009). Looking at shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, *34*(1).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Schuckers, M., Purnapatra, S., Fatima, K., Hou, D., & Schuckers, S. (2022). Statistical methods for assessing differences in false non-match rates across demographic groups. In *Understanding and Mitigating Demographic Bias in Biometric Systems(UMDBB) Workshop part of 2022 International Conference on Pattern Recognition (ICPR)*.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, New York, NY, USA.

Simonite, T. (2018). When it comes to gorillas, google photos remains blind. *Wired*.

Sukthanker, R., Dooley, S., Dickerson, J. P., White, C., Hutter, F., & Goldblum, M. (2022). On the importance of architectures and hyperparameters for fairness in face recognition. *arXiv:2210.09943*.

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

UNESCO (2021). Recommendation on the ethics of artificial intelligence. Tech. rep..

Vera-Rodriguez, R., Blazquez, M., Morales, A., Gonzalez-Sosa, E., Neves, J. C., & Proença, H. (2019). Facegenderid: Exploiting gender information in dcnns face recognition systems. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2254–2260.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pp. 1–7, Gothenburg Sweden.

Wachter, S., Mittelstadt, B., & Russel, C. (2021a). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, *41*.

Wachter, S., Mittelstadt, B., & Russell, C. (2021b). Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *Social Science Research Network*.

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244.

Wang, M., Zhang, Y., & Deng, W. (2022). Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 8433–8448.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109.

Wolpert, D., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Wong, J. C. (2019). Google reportedly targeted people with 'dark skin' to improve facial recognition. *The Guardian*.

Yee, K., Tantipongpipat, U., & Mishra, S. (2021). Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–24.

Zhao, H., & Gordon, G. (2019). Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv:1505.05723*.

Appendix A Model architecture

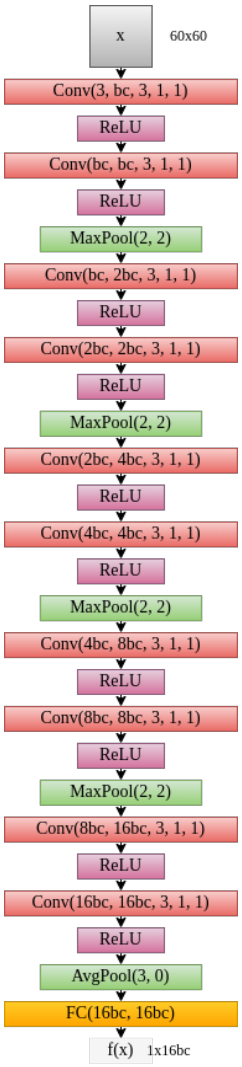

Figure 7: Architecture of the neural network. Conv(a, b, c, d, e) denotes a convolutional layer with number of input channels a, number of output channels b, kernel size c x c, stride size d and padding e. MaxPool(a, b) and AvgPool(a, b) denote a max pooling layer and average pooling layer respectively with kernel size a and stride b. FC(a, b) denotes a fully connected layer with number of in features a, number of out features b. `bc` denotes the `base channel` parameter, which is set at 32 for `BaseModel` and 16 and 64 for `Depth16` and `Depth64` respectively.

Appendix B Efficiency results

B.1 Validation loss

Figure 9 shows the evolution of the validation loss during training for the different models used in this work and compiled in Table 3.

After 600 epochs, the validation loss is the lowest for the `BaseModel`, showing why it was elected to serve as a basis for the comparison with other models.

The only doubt could be for `RandomSampling` as figure 9 shows that its validation loss function decreases much quicker in the early epochs. Yet, it plateaus immediately. Observations on the training loss show that the model seems to overfit quite a lot, yet the validation loss is similar after 600 epochs.

Note that for the choice of margin, this measure could not be taken as the scales of the functions are very different. Yet, it can be observed that the ratio between the starting and ending values is higher, which means that the model has learned more for a margin of 0.2. In the original paper, the choice of a 0.2 margin is not explained (Schroff et al., 2015).

B.2 Other efficiency measures

Model	Final Validation Loss	Time to train	Accuracy	FMR	FNMR	TLR
BaseModel	0.03001	69 min	0.869	0.157	0.100	0.935
RandomSampling	0.03122	345 min	0.879	0.140	0.099	0.951
Normalized	0.06323	81 min	0.830	0.204	0.128	0.907
NoAugment	0.07485	74 min	0.831	0.193	0.142	0.907
HighJitter	0.04562	84 min	0.839	0.190	0.126	0.920
HighDeformation	0.04464	89 min	0.847	0.176	0.126	0.926
HighRotation	0.03985	69 min	0848	0.178	0.121	0.927
HighZoom	0.04682	68 min	0.840	0.180	0.133	0.921
Depth16	0.04915	85 min	0.853	0.173	0.115	0.930
Depth64	0.04556	95 min	0.843	0.184	0.126	0.924
Margin01	0.01945	74 min	0.854	0.171	0.117	0.932
Margin05	0.1112	80 min	0.848	0.180	0.119	0.926
Margin1	0.2272	70 min	0.852	0.176	0.114	0.926
$Lr10^{-3}$	0.04441	71 min	0.842	0.180	0.132	0.918
$Lr10^{-4}$	0.04515	76 min	0.853	0.173	0.117	0.927
Scheduler300	0.0347	78 min	0.865	0.157	0.111	0.939
Scheduler100	0.03972	80 min	0.855	0.172	0.112	0.933
LowThreshold	0.03001	69 min	0.816	0.081	0.246	0.935
HighThreshold	0.03001	69 min	0.826	0.245	0.049	0.935

Figure 8: Table of the different models with their relative efficiency

Appendix C Fairness results

italic: gap between groups for a specific model to the benefit of minority groups

italic and bold: gap between groups for a specific model to the benefit of majority groups

writing in green: metric improvement from base model

writing in red: metric deterioration from base model

highlighting in green: bias suppression from base model

highlighting in yellow: bias creating from base model to the benefit of minority groups

highlighting in red: bias creating from base model to the benefit of majority groups

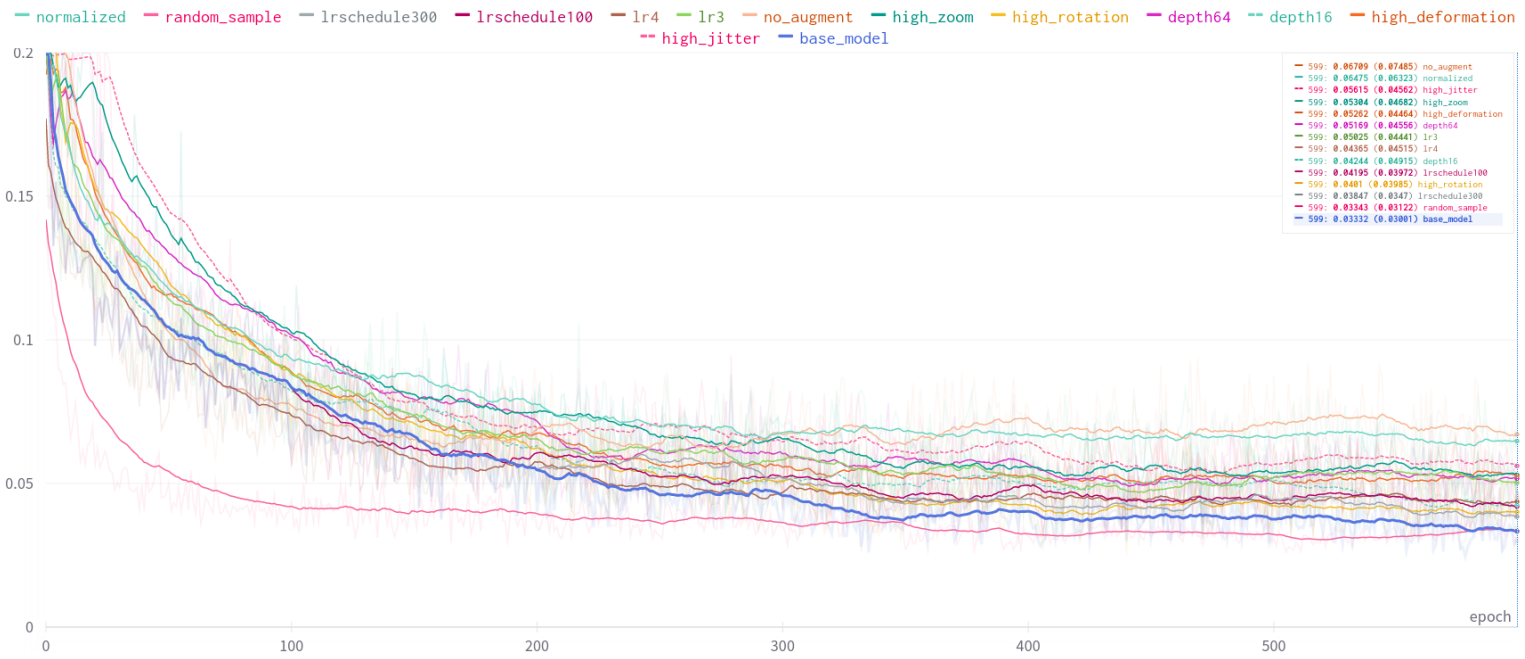

Figure 9: Validation loss function for the base model and all its studied variations (except for the margin choice)

C.1 Sampling choice

Sampling	Base	Random
	Male	
Accuracy	*0.847 (0.841-0.852)*	*0.867 (0.861-0.872)*
TLR	***0.889 (0.883-0.896)***	***0.912 (0.906-0.917)***
FMR	*0.176 (0.169-0.184)*	*0.145 (0.138-0.153)*
FNMR	***0.122 (0.114-0.129)***	***0.117 (0.109-0.124)***
	Female	
Accuracy	*0.894 (0.882-0.906)*	*0.909 (0.898-0.92)*
TLR	***0.651 (0.599-0.706)***	***0.68 (0.624-0.729)***
FMR	*0.064 (0.054-0.075)*	*0.054 (0.044-0.063)*
FNMR	***0.269 (0.231-0.31)***	***0.221 (0.186-0.257)***

Sampling	Base	Random
	White	
Accuracy	0.866 (0.86-0.872)	*0.878 (0.873-0.884)*
TLR	***0.886 (0.879-0.892)***	***0.898 (0.892-0.904)***
FMR	*0.147 (0.14-0.155)*	*0.122 (0.115-0.129)*
FNMR	***0.115 (0.107-0.124)***	***0.121 (0.113-0.129)***
	Non-White	
Accuracy	0.879 (0.867-0.891)	*0.913 (0.901-0.924)*
TLR	***0.736 (0.693-0.781)***	***0.731 (0.681-0.778)***
FMR	*0.082 (0.07-0.093)*	*0.053 (0.044-0.063)*
FNMR	***0.222 (0.193-0.252)***	***0.183 (0.156-0.212)***
	Black	
Accuracy	0.899 (0.869-0.93)	*0.873 (0.84-0.903)*
TLR	0.5 (0.143-1.0)	0.5 (0.0-1.0)
FMR	*0.011 (0.0-0.023)*	*0.034 (0.019-0.054)*
FNMR	***0.695 (0.564-0.816)***	***0.725 (0.605-0.829)***

Sampling	Base	Random
White Male		
Accuracy	*0.853 (0.846-0.859)*	*0.865 (0.858-0.872)*
TLR	***0.844 (0.835-0.855)***	***0.858 (0.848-0.868)***
FMR	*0.156 (0.147-0.164)*	*0.134 (0.126-0.143)*
FNMR	***0.131 (0.121-0.142)***	0.136 (0.124-0.146)
Non-White Female		
Accuracy	*0.878 (0.869-0.887)*	*0.913 (0.906-0.92)*
TLR	***0.794 (0.772-0.815)***	***0.807 (0.785-0.829)***
FMR	*0.096 (0.087-0.105)*	*0.066 (0.059-0.075)*
FNMR	***0.173 (0.155-0.19)***	0.127 (0.111-0.143)
Black Female		
Accuracy	*0.962 (0.924-0.987)*	*0.96 (0.92-0.987)*
TLR	0.0 (0.0-1.0)	***0.0 (0.0-0.5)***
FMR	*0.025 (0.0-0.064)*	*0.041 (0.014-0.083)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]

C.2 Normalization choice

Normalization	Base	Normalized
Male		
Accuracy	*0.847 (0.841-0.852)*	*0.818 (0.812-0.824)*
TLR	***0.889 (0.883-0.896)***	***0.861 (0.854-0.868)***
FMR	*0.176 (0.169-0.184)*	*0.211 (0.203-0.219)*
FNMR	***0.122 (0.114-0.129)***	***0.136 (0.128-0.145)***
Female		
Accuracy	*0.894 (0.882-0.906)*	*0.865 (0.85-0.878)*
TLR	***0.651 (0.599-0.706)***	***0.652 (0.597-0.707)***
FMR	*0.064 (0.054-0.075)*	*0.084 (0.073-0.096)*
FNMR	***0.269 (0.231-0.31)***	***0.319 (0.281-0.357)***

Normalization	Base	Normalized
White		
Accuracy	0.866 (0.86-0.872)	*0.843 (0.837-0.849)*
TLR	***0.886 (0.879-0.892)***	***0.862 (0.855-0.868)***
FMR	*0.147 (0.14-0.155)*	*0.172 (0.164-0.18)*
FNMR	***0.115 (0.107-0.124)***	***0.135 (0.126-0.143)***
Non-White		
Accuracy	0.879 (0.867-0.891)	*0.864 (0.851-0.877)*
TLR	***0.736 (0.693-0.781)***	***0.7 (0.645-0.748)***
FMR	*0.082 (0.07-0.093)*	*0.083 (0.072-0.095)*
FNMR	***0.222 (0.193-0.252)***	***0.277 (0.248-0.312)***
Black		
Accuracy	0.899 (0.869-0.93)	0.84 (0.803-0.87)
TLR	0.5 (0.143-1.0)	0.5 (0.0-0.881)
FMR	*0.011 (0.0-0.023)*	*0.032 (0.016-0.052)*
FNMR	***0.695 (0.564-0.816)***	***0.776 (0.679-0.863)***

Normalization	Base	Normalized
White Male		
Accuracy	*0.853 (0.846-0.859)*	*0.835 (0.828-0.842)*
TLR	***0.844 (0.835-0.855)***	***0.821 (0.811-0.832)***
FMR	*0.156 (0.147-0.164)*	*0.176 (0.167-0.186)*
FNMR	***0.131 (0.121-0.142)***	***0.143 (0.131-0.155)***
Non-White Female		
Accuracy	*0.878 (0.869-0.887)*	*0.865 (0.856-0.873)*
TLR	***0.794 (0.772-0.815)***	***0.78 (0.758-0.801)***
FMR	*0.096 (0.087-0.105)*	*0.102 (0.093-0.112)*
FNMR	***0.173 (0.155-0.19)***	***0.201 (0.184-0.22)***
Black Female		
Accuracy	*0.962 (0.924-0.987)*	*0.92 (0.867-0.96)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.025 (0.0-0.064)*	*0.043 (0.014-0.087)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]

C.3 Augmentation choice

Augmentation	Base	No augmentation	High zoom
Male			
Accuracy	*0.847 (0.841-0.852)*	*0.815 (0.809-0.821)*	*0.83 (0.823-0.836)*
TLR	***0.889 (0.883-0.896)***	***0.865 (0.858-0.872)***	***0.88 (0.874-0.886)***
FMR	*0.176 (0.169-0.184)*	*0.198 (0.19-0.207)*	*0.192 (0.184-0.2)*
FNMR	***0.122 (0.114-0.129)***	***0.166 (0.157-0.175)***	***0.139 (0.13-0.147)***
Female			
Accuracy	*0.894 (0.882-0.906)*	*0.856 (0.842-0.87)*	*0.852 (0.838-0.866)*
TLR	***0.651 (0.599-0.706)***	***0.65 (0.597-0.706)***	***0.655 (0.598-0.706)***
FMR	*0.064 (0.054-0.075)*	*0.078 (0.066-0.09)*	*0.087 (0.075-0.099)*
FNMR	***0.269 (0.231-0.31)***	***0.352 (0.317-0.391)***	***0.364 (0.326-0.405)***
Augmentation	**High deformation**	**High color jitter**	**High rotation**
Male			
Accuracy	0.835 (0.829-0.841)	*0.819 (0.813-0.825)*	*0.83 (0.824-0.836)*
TLR	***0.884 (0.878-0.89)***	***0.873 (0.866-0.88)***	***0.881 (0.875-0.888)***
FMR	*0.183 (0.175-0.191)*	*0.205 (0.197-0.213)*	*0.19 (0.182-0.198)*
FNMR	***0.14 (0.131-0.148)***	***0.146 (0.137-0.155)***	***0.14 (0.132-0.149)***
Female			
Accuracy	0.852 (0.838-0.864)	*0.842 (0.827-0.856)*	*0.88 (0.867-0.892)*
TLR	***0.664 (0.605-0.717)***	***0.652 (0.594-0.704)***	***0.665 (0.612-0.72)***
FMR	*0.08 (0.068-0.093)*	*0.094 (0.082-0.107)*	*0.078 (0.066-0.089)*
FNMR	***0.377 (0.339-0.416)***	***0.391 (0.353-0.43)***	***0.297 (0.254-0.336)***

Augmentation	Base	No augmentation	High zoom
White			
Accuracy	0.866 (0.86-0.872)	0.834 (0.828-0.84)	*0.841 (0.835-0.846)*
TLR	***0.886 (0.879-0.892)***	***0.859 (0.852-0.865)***	***0.871 (0.864-0.877)***
FMR	*0.147 (0.14-0.155)*	*0.168 (0.16-0.175)*	*0.161 (0.154-0.169)*
FNMR	***0.115 (0.107-0.124)***	***0.163 (0.153-0.173)***	***0.155 (0.146-0.165)***
Non-White			
Accuracy	0.879 (0.867-0.891)	0.835 (0.821-0.847)	*0.866 (0.854-0.879)*
TLR	***0.736 (0.693-0.781)***	***0.707 (0.654-0.757)***	***0.73 (0.684-0.773)***
FMR	*0.082 (0.07-0.093)*	*0.084 (0.072-0.097)*	*0.079 (0.067-0.091)*
FNMR	***0.222 (0.193-0.252)***	***0.35 (0.32-0.381)***	***0.263 (0.235-0.294)***
Black			
Accuracy	0.899 (0.869-0.93)	0.85 (0.816-0.883)	0.835 (0.799-0.868)
TLR	0.5 (0.143-1.0)	0.5 (0.0-1.0)	0.5 (0.143-1.0)
FMR	*0.011 (0.0-0.023)*	*0.028 (0.012-0.045)*	*0.024 (0.008-0.041)*
FNMR	***0.695 (0.564-0.816)***	***0.745 (0.641-0.841)***	***0.746 (0.644-0.837)***
Augmentation	**High deformation**	**High color jitter**	**High rotation**
White			
Accuracy	0.846 (0.84-0.852)	0.841 (0.835-0.847)	0.853 (0.847-0.859)
TLR	***0.875 (0.869-0.882)***	***0.866 (0.859-0.873)***	***0.871 (0.864-0.878)***
FMR	*0.154 (0.147-0.162)*	*0.173 (0.165-0.181)*	*0.159 (0.152-0.167)*
FNMR	***0.154 (0.144-0.162)***	***0.136 (0.127-0.145)***	***0.128 (0.12-0.137)***
Non-White			
Accuracy	0.862 (0.85-0.875)	0.85 (0.837-0.864)	0.843 (0.83-0.857)
TLR	***0.741 (0.695-0.784)***	***0.717 (0.669-0.766)***	***0.731 (0.681-0.77)***
FMR	*0.09 (0.078-0.103)*	*0.082 (0.07-0.095)*	*0.091 (0.077-0.104)*
FNMR	***0.25 (0.224-0.28)***	***0.307 (0.274-0.341)***	***0.299 (0.27-0.328)***
Black			
Accuracy	0.822 (0.783-0.859)	0.882 (0.853-0.912)	0.864 (0.832-0.893)
TLR	0.5 (0.167-0.875)	0.556 (0.2-1.0)	0.5 (0.143-0.857)
FMR	*0.033 (0.016-0.053)*	*0.027 (0.011-0.045)*	*0.034 (0.016-0.054)*
FNMR	***0.754 (0.656-0.844)***	***0.667 (0.545-0.778)***	***0.646 (0.543-0.764)***

Augmentation	Base	No augmentation	High zoom
White Male			
Accuracy	*0.853 (0.846-0.859)*	*0.816 (0.808-0.823)*	*0.827 (0.82-0.835)*
TLR	***0.844 (0.835-0.855)***	0.813 (0.802-0.823)	***0.832 (0.822-0.843)***
FMR	*0.156 (0.147-0.164)*	*0.179 (0.169-0.188)*	*0.172 (0.162-0.181)*
FNMR	***0.131 (0.121-0.142)***	***0.193 (0.18-0.206)***	***0.175 (0.163-0.187)***
Non-White Female			
Accuracy	*0.878 (0.869-0.887)*	*0.851 (0.841-0.86)*	*0.865 (0.855-0.874)*
TLR	***0.794 (0.772-0.815)***	0.78 (0.757-0.803)	***0.79 (0.768-0.81)***
FMR	*0.096 (0.087-0.105)*	*0.097 (0.088-0.107)*	*0.099 (0.089-0.109)*
FNMR	***0.173 (0.155-0.19)***	***0.241 (0.22-0.26)***	***0.206 (0.189-0.224)***
Black Female			
Accuracy	*0.962 (0.924-0.987)*	*0.96 (0.92-0.987)*	*0.987 (0.962-1.0)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.025 (0.0-0.064)*	*0.028 (0.0-0.068)*	*0.0 (0.0-0.0)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]	xxx [xxx - xxx]

Augmentation	High deformation	High color jitter	High rotation
White Male			
Accuracy	*0.835 (0.829-0.843)*	*0.826 (0.819-0.833)*	*0.841 (0.833-0.848)*
TLR	***0.834 (0.824-0.844)***	***0.823 (0.812-0.833)***	***0.833 (0.821-0.844)***
FMR	*0.162 (0.153-0.172)*	*0.184 (0.175-0.193)*	*0.168 (0.159-0.177)*
FNMR	***0.167 (0.155-0.179)***	***0.154 (0.143-0.166)***	***0.145 (0.133-0.158)***
Non-White Female			
Accuracy	*0.866 (0.856-0.875)*	*0.845 (0.835-0.854)*	*0.858 (0.848-0.867)*
TLR	***0.799 (0.777-0.82)***	***0.777 (0.754-0.8)***	***0.799 (0.779-0.82)***
FMR	*0.1 (0.091-0.11)*	*0.107 (0.097-0.117)*	*0.105 (0.095-0.115)*
FNMR	***0.2 (0.183-0.219)***	***0.246 (0.226-0.265)***	***0.211 (0.193-0.229)***
Black Female			
Accuracy	*0.957 (0.914-0.986)*	*1.0 (1.0-1.0)*	*0.963 (0.927-0.988)*
TLR	***0.0 (0.0-0.525)***	***0.0 (0.0-0.0)***	0.0 (0.0-1.0)
FMR	*0.029 (0.0-0.06)*	*0.0 (0.0-0.0)*	*0.038 (0.012-0.077)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]	xxx [xxx - xxx]

C.4 Depth choice

Depth	16	32 (base)	64
Male			
Accuracy	*0.837 (0.831-0.843)*	*0.847 (0.841-0.852)*	*0.831 (0.824-0.837)*
TLR	***0.883 (0.877-0.89)***	***0.889 (0.883-0.896)***	***0.882 (0.875-0.888)***
FMR	*0.181 (0.173-0.189)*	*0.176 (0.169-0.184)*	*0.184 (0.177-0.192)*
FNMR	***0.136 (0.128-0.145)***	***0.122 (0.114-0.129)***	***0.147 (0.139-0.156)***
Female			
Accuracy	*0.872 (0.858-0.883)*	*0.894 (0.882-0.906)*	*0.857 (0.843-0.869)*
TLR	***0.679 (0.629-0.726)***	***0.651 (0.599-0.706)***	***0.649 (0.593-0.701)***
FMR	*0.086 (0.074-0.098)*	*0.064 (0.054-0.075)*	*0.103 (0.089-0.117)*
FNMR	***0.288 (0.251-0.325)***	***0.269 (0.231-0.31)***	***0.316 (0.277-0.357)***

Depth	16	32 (base)	64
White			
Accuracy	0.857 (0.852-0.863)	0.866 (0.86-0.872)	*0.839 (0.833-0.845)*
TLR	***0.875 (0.868-0.881)***	***0.886 (0.879-0.892)***	***0.872 (0.865-0.879)***
FMR	*0.152 (0.145-0.16)*	*0.147 (0.14-0.155)*	*0.163 (0.155-0.171)*
FNMR	***0.129 (0.12-0.138)***	***0.115 (0.107-0.124)***	***0.156 (0.148-0.166)***
Non-White			
Accuracy	0.857 (0.843-0.87)	0.879 (0.867-0.891)	*0.867 (0.854-0.879)*
TLR	***0.731 (0.685-0.774)***	***0.736 (0.693-0.781)***	***0.719 (0.668-0.769)***
FMR	*0.087 (0.075-0.099)*	*0.082 (0.07-0.093)*	*0.079 (0.067-0.091)*
FNMR	***0.276 (0.248-0.306)***	***0.222 (0.193-0.252)***	***0.267 (0.237-0.298)***
Black			
Accuracy	0.857 (0.825-0.89)	0.899 (0.869-0.93)	*0.879 (0.85-0.912)*
TLR	0.5 (0.167-0.881)	0.5 (0.143-1.0)	0.5 (0.125-1.0)
FMR	*0.027 (0.012-0.047)*	*0.011 (0.0-0.023)*	*0.034 (0.018-0.056)*
FNMR	***0.672 (0.566-0.772)***	***0.695 (0.564-0.816)***	***0.609 (0.491-0.726)***

Depth	16	32 (base)	64
White Male			
Accuracy	*0.84 (0.833-0.847)*	*0.853 (0.846-0.859)*	*0.827 (0.82-0.834)*
TLR	***0.833 (0.822-0.844)***	***0.844 (0.835-0.855)***	***0.831 (0.82-0.841)***
FMR	*0.161 (0.153-0.17)*	*0.156 (0.147-0.164)*	*0.169 (0.16-0.179)*
FNMR	***0.157 (0.146-0.169)***	***0.131 (0.121-0.142)***	***0.181 (0.168-0.193)***
Non-White Female			
Accuracy	*0.866 (0.857-0.875)*	*0.878 (0.869-0.887)*	*0.858 (0.849-0.867)*
TLR	***0.795 (0.773-0.815)***	***0.794 (0.772-0.815)***	***0.779 (0.757-0.8)***
FMR	*0.102 (0.093-0.112)*	*0.096 (0.087-0.105)*	*0.111 (0.102-0.121)*
FNMR	***0.194 (0.177-0.213)***	***0.173 (0.155-0.19)***	***0.204 (0.186-0.222)***
Black Female			
Accuracy	*0.972 (0.93-1.0)*	*0.962 (0.924-0.987)*	*0.959 (0.918-0.986)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.028 (0.0-0.056)*	*0.025 (0.0-0.064)*	*0.028 (0.0-0.058)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]	xxx [xxx - xxx]

C.5 Margin choice

Margin	1	0.5
Male		
Accuracy	*0.84 (0.835-0.846)*	*0.832 (0.826-0.838)*
TLR	***0.886 (0.88-0.892)***	***0.877 (0.871-0.883)***
FMR	*0.184 (0.176-0.191)*	*0.185 (0.177-0.193)*
FNMR	***0.124 (0.116-0.132)***	***0.144 (0.136-0.153)***
Female		
Accuracy	*0.872 (0.859-0.884)*	*0.876 (0.862-0.889)*
TLR	***0.663 (0.608-0.713)***	***0.658 (0.6-0.712)***
FMR	*0.077 (0.066-0.089)*	*0.072 (0.062-0.084)*
FNMR	***0.323 (0.286-0.363)***	***0.313 (0.274-0.353)***
Margin	**0.2 (base)**	**0.1**
Male		
Accuracy	*0.847 (0.841-0.852)*	0.849 (0.844-0.856)
TLR	***0.889 (0.883-0.896)***	***0.894 (0.887-0.9)***
FMR	*0.176 (0.169-0.184)*	*0.169 (0.162-0.177)*
FNMR	***0.122 (0.114-0.129)***	***0.123 (0.115-0.131)***
Female		
Accuracy	*0.894 (0.882-0.906)*	0.857 (0.844-0.871)
TLR	***0.651 (0.599-0.706)***	***0.664 (0.609-0.715)***
FMR	*0.064 (0.054-0.075)*	*0.086 (0.074-0.098)*
FNMR	***0.269 (0.231-0.31)***	***0.339 (0.302-0.38)***

Margin	**1**	**0.5**
White		
Accuracy	0.852 (0.846-0.858)	*0.848 (0.843-0.854)*
TLR	***0.875 (0.869-0.882)***	***0.873 (0.866-0.88)***
FMR	*0.154 (0.147-0.162)*	*0.153 (0.146-0.161)*
FNMR	***0.14 (0.13-0.148)***	***0.149 (0.14-0.158)***
Non-White		
Accuracy	0.86 (0.846-0.872)	*0.867 (0.855-0.879)*
TLR	***0.743 (0.698-0.788)***	***0.72 (0.673-0.763)***
FMR	*0.093 (0.08-0.106)*	*0.085 (0.074-0.097)*
FNMR	***0.254 (0.225-0.283)***	***0.253 (0.222-0.283)***
Black		
Accuracy	0.878 (0.845-0.908)	*0.888 (0.855-0.914)*
TLR	0.5 (0.119-1.0)	0.5 (0.0-1.0)
FMR	*0.044 (0.023-0.066)*	*0.03 (0.015-0.05)*
FNMR	***0.694 (0.562-0.821)***	***0.673 (0.553-0.794)***
Margin	**0.2 (base)**	**0.1**
White		
Accuracy	0.866 (0.86-0.872)	0.86 (0.855-0.866)
TLR	***0.886 (0.879-0.892)***	***0.88 (0.873-0.887)***
FMR	*0.147 (0.14-0.155)*	*0.147 (0.14-0.155)*
FNMR	***0.115 (0.107-0.124)***	***0.129 (0.12-0.137)***
Non-White		
Accuracy	0.879 (0.867-0.891)	0.878 (0.866-0.889)
TLR	***0.736 (0.693-0.781)***	***0.717 (0.668-0.762)***
FMR	*0.082 (0.07-0.093)*	*0.077 (0.066-0.089)*
FNMR	***0.222 (0.193-0.252)***	***0.246 (0.217-0.276)***
Black		
Accuracy	0.899 (0.869-0.93)	0.877 (0.847-0.91)
TLR	0.5 (0.143-1.0)	0.5 (0.0-1.0)
FMR	*0.011 (0.0-0.023)*	*0.019 (0.008-0.036)*
FNMR	***0.695 (0.564-0.816)***	***0.686 (0.561-0.796)***

Margin	**1**	**0.5**
White Male		
Accuracy	*0.835 (0.827-0.842)*	*0.83 (0.823-0.838)*
TLR	***0.832 (0.822-0.842)***	***0.833 (0.822-0.844)***
FMR	*0.166 (0.157-0.176)*	*0.164 (0.155-0.173)*
FNMR	0.162 (0.151-0.175)	0.179 (0.167-0.191)
Non-White Female		
Accuracy	*0.873 (0.865-0.882)*	*0.875 (0.866-0.883)*
TLR	***0.799 (0.777-0.819)***	***0.789 (0.768-0.808)***
FMR	*0.104 (0.094-0.114)*	*0.096 (0.087-0.106)*
FNMR	0.172 (0.154-0.189)	0.184 (0.167-0.204)
Black Female		
Accuracy	*0.971 (0.942-1.0)*	*1.0 (1.0-1.0)*
TLR	***0.0 (0.0-0.5)***	0.0 (0.0-1.0)
FMR	*0.029 (0.0-0.057)*	*0.0 (0.0-0.0)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]
Margin	**0.2 (base)**	**0.1**
White Male		
Accuracy	*0.853 (0.846-0.859)*	*0.854 (0.848-0.861)*
TLR	***0.844 (0.835-0.855)***	***0.842 (0.831-0.852)***
FMR	*0.156 (0.147-0.164)*	*0.152 (0.143-0.16)*
FNMR	***0.131 (0.121-0.142)***	***0.134 (0.124-0.145)***
Non-White Female		
Accuracy	*0.878 (0.869-0.887)*	*0.878 (0.869-0.887)*
TLR	***0.794 (0.772-0.815)***	***0.788 (0.765-0.809)***
FMR	*0.096 (0.087-0.105)*	*0.094 (0.084-0.103)*
FNMR	***0.173 (0.155-0.19)***	***0.179 (0.161-0.196)***
Black Female		
Accuracy	*0.962 (0.924-0.987)*	*0.933 (0.88-0.973)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.025 (0.0-0.064)*	*0.042 (0.014-0.087)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]

C.6 Learning rate choice

lr	5.10-4, scheduler 200 (base)	10-3	10-4
	Male		
Accuracy	*0.847 (0.841-0.852)*	0.831 (0.826-0.837)	*0.841 (0.835-0.846)*
TLR	***0.889 (0.883-0.896)***	***0.881 (0.875-0.888)***	***0.888 (0.881-0.894)***
FMR	*0.176 (0.169-0.184)*	*0.182 (0.174-0.19)*	*0.181 (0.173-0.188)*
FNMR	***0.122 (0.114-0.129)***	***0.15 (0.142-0.159)***	***0.129 (0.121-0.136)***
	Female		
Accuracy	*0.894 (0.882-0.906)*	0.843 (0.829-0.856)	*0.874 (0.861-0.886)*
TLR	***0.651 (0.599-0.706)***	***0.652 (0.597-0.705)***	***0.657 (0.603-0.712)***
FMR	*0.064 (0.054-0.075)*	*0.093 (0.081-0.106)*	*0.079 (0.067-0.091)*
FNMR	***0.269 (0.231-0.31)***	***0.374 (0.334-0.413)***	***0.306 (0.267-0.345)***

lr	scheduler 100	scheduler 300
	Male	
Accuracy	0.847 (0.841-0.853)	0.847 (0.841-0.852)
TLR	***0.887 (0.88-0.894)***	***0.893 (0.888-0.899)***
FMR	*0.171 (0.162-0.179)*	*0.171 (0.164-0.179)*
FNMR	***0.13 (0.121-0.138)***	***0.129 (0.121-0.137)***
	Female	
Accuracy	0.855 (0.841-0.869)	0.858 (0.844-0.871)
TLR	***0.654 (0.596-0.712)***	***0.659 (0.608-0.713)***
FMR	*0.082 (0.071-0.095)*	*0.092 (0.08-0.106)*
FNMR	***0.365 (0.325-0.407)***	***0.339 (0.299-0.382)***

lr	5.10-4, scheduler 200 (base)	10-3	10-4
White			
Accuracy	0.866 (0.86-0.872)	***0.855 (0.85-0.861)***	*0.853 (0.848-0.859)*
TLR	***0.886 (0.879-0.892)***	***0.874 (0.868-0.881)***	***0.878 (0.871-0.885)***
FMR	*0.147 (0.14-0.155)*	*0.152 (0.145-0.16)*	*0.152 (0.144-0.159)*
FNMR	***0.115 (0.107-0.124)***	***0.134 (0.125-0.143)***	***0.14 (0.131-0.148)***
Non-White			
Accuracy	0.879 (0.867-0.891)	*0.832 (0.819-0.846)*	*0.889 (0.877-0.901)*
TLR	***0.736 (0.693-0.781)***	***0.714 (0.665-0.764)***	***0.738 (0.688-0.782)***
FMR	*0.082 (0.07-0.093)*	*0.074 (0.063-0.087)*	*0.084 (0.073-0.097)*
FNMR	***0.222 (0.193-0.252)***	***0.362 (0.33-0.392)***	***0.19 (0.162-0.219)***
Black			
Accuracy	0.899 (0.869-0.93)	*0.743 (0.7-0.783)*	*0.891 (0.861-0.921)*
TLR	0.5 (0.143-1.0)	0.5 (0.0-0.857)	0.5 (0.0-0.881)
FMR	*0.011 (0.0-0.023)*	*0.028 (0.01-0.047)*	*0.034 (0.015-0.053)*
FNMR	***0.695 (0.564-0.816)***	***0.835 (0.768-0.897)***	***0.675 (0.537-0.804)***

lr	scheduler 100	scheduler 300
White		
Accuracy	0.865 (0.86-0.871)	*0.86 (0.854-0.866)*
TLR	***0.879 (0.872-0.886)***	***0.883 (0.877-0.89)***
FMR	*0.141 (0.134-0.148)*	*0.145 (0.138-0.153)*
FNMR	***0.126 (0.117-0.134)***	***0.132 (0.124-0.141)***
Non-White		
Accuracy	0.868 (0.855-0.88)	*0.885 (0.874-0.897)*
TLR	***0.738 (0.69-0.783)***	***0.731 (0.685-0.777)***
FMR	*0.091 (0.077-0.104)*	*0.071 (0.06-0.083)*
FNMR	***0.235 (0.206-0.264)***	***0.23 (0.2-0.259)***
Black		
Accuracy	0.875 (0.845-0.905)	0.889 (0.859-0.918)
TLR	0.5 (0.0-0.857)	0.5 (0.167-0.876)
FMR	*0.038 (0.019-0.059)*	*0.023 (0.008-0.038)*
FNMR	***0.684 (0.558-0.8)***	***0.642 (0.52-0.75)***

lr	5.10-4, scheduler 200 (base)	10-3	10-4
White Male			
Accuracy	*0.853 (0.846-0.859)*	0.846 (0.838-0.853)	*0.842 (0.835-0.849)*
TLR	***0.844 (0.835-0.855)***	***0.833 (0.822-0.844)***	***0.842 (0.831-0.852)***
FMR	*0.156 (0.147-0.164)*	*0.158 (0.149-0.166)*	*0.162 (0.153-0.17)*
FNMR	***0.131 (0.121-0.142)***	***0.148 (0.137-0.16)***	***0.153 (0.141-0.165)***
Non-White Female			
Accuracy	*0.878 (0.869-0.887)*	0.848 (0.839-0.857)	*0.884 (0.875-0.893)*
TLR	***0.794 (0.772-0.815)***	***0.784 (0.764-0.806)***	***0.798 (0.778-0.818)***
FMR	*0.096 (0.087-0.105)*	*0.102 (0.092-0.112)*	*0.093 (0.084-0.103)*
FNMR	***0.173 (0.155-0.19)***	***0.243 (0.223-0.263)***	***0.162 (0.145-0.179)***
Black Female			
Accuracy	*0.962 (0.924-0.987)*	*0.933 (0.88-0.973)*	*0.987 (0.962-1.0)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.025 (0.0-0.064)*	*0.042 (0.013-0.087)*	*0.0 (0.0-0.0)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]	xxx [xxx - xxx]

lr	scheduler 100	scheduler 300
White Male		
Accuracy	0.85 (0.843-0.857)	*0.847 (0.84-0.854)*
TLR	***0.834 (0.824-0.845)***	***0.844 (0.834-0.854)***
FMR	*0.15 (0.142-0.159)*	*0.153 (0.144-0.162)*
FNMR	***0.151 (0.139-0.162)***	***0.153 (0.142-0.165)***
Non-White Female		
Accuracy	0.863 (0.854-0.872)	*0.873 (0.864-0.882)*
TLR	***0.794 (0.773-0.813)***	***0.788 (0.768-0.809)***
FMR	*0.104 (0.094-0.114)*	*0.096 (0.087-0.106)*
FNMR	***0.199 (0.182-0.216)***	***0.189 (0.172-0.207)***
Black Female		
Accuracy	*0.971 (0.929-1.0)*	*1.0 (1.0-1.0)*
TLR	***0.0 (0.0-0.0)***	0.0 (0.0-1.0)
FMR	*0.029 (0.0-0.071)*	*0.0 (0.0-0.0)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]

C.7 Threshold choice

Threshold	Low (0.48)	Base (0.81)	High (1.15)
Male			
Accuracy	*0.807 (0.801-0.813)*	*0.847 (0.841-0.852)*	*0.802 (0.795-0.808)*
TLR	***0.89 (0.884-0.896)***	***0.889 (0.883-0.896)***	***0.892 (0.886-0.898)***
FMR	*0.081 (0.074-0.088)*	*0.176 (0.169-0.184)*	*0.258 (0.25-0.267)*
FNMR	***0.27 (0.262-0.28)***	***0.122 (0.114-0.129)***	***0.061 (0.055-0.068)***
Female			
Accuracy	*0.752 (0.736-0.77)*	*0.894 (0.882-0.906)*	*0.879 (0.866-0.891)*
TLR	***0.655 (0.602-0.712)***	***0.651 (0.599-0.706)***	***0.653 (0.599-0.706)***
FMR	*0.047 (0.036-0.057)*	*0.064 (0.054-0.075)*	*0.115 (0.103-0.129)*
FNMR	***0.547 (0.516-0.578)***	***0.269 (0.231-0.31)***	***0.156 (0.117-0.194)***

Threshold	Low (0.48)	Base (0.81)	High (1.15)
White			
Accuracy	***0.804 (0.797-0.81)***	0.866 (0.86-0.872)	*0.836 (0.83-0.842)*
TLR	***0.882 (0.875-0.888)***	***0.886 (0.879-0.892)***	***0.883 (0.876-0.889)***
FMR	*0.073 (0.066-0.079)*	*0.147 (0.14-0.155)*	*0.217 (0.209-0.225)*
FNMR	***0.287 (0.277-0.297)***	***0.115 (0.107-0.124)***	***0.042 (0.036-0.049)***
Non-White			
Accuracy	*0.777 (0.761-0.793)*	0.879 (0.867-0.891)	*0.858 (0.845-0.871)*
TLR	***0.734 (0.693-0.779)***	***0.736 (0.693-0.781)***	***0.72 (0.671-0.761)***
FMR	*0.045 (0.035-0.056)*	*0.082 (0.07-0.093)*	*0.134 (0.12-0.148)*
FNMR	***0.442 (0.416-0.467)***	***0.222 (0.193-0.252)***	***0.171 (0.142-0.202)***
Black			
Accuracy	*0.712 (0.673-0.754)*	0.899 (0.869-0.93)	*0.9 (0.87-0.927)*
TLR	0.5 (0.167-1.0)	0.5 (0.143-1.0)	0.5 (0.0-1.0)
FMR	*0.024 (0.009-0.044)*	*0.011 (0.0-0.023)*	*0.034 (0.015-0.054)*
FNMR	***0.792 (0.726-0.856)***	***0.695 (0.564-0.816)***	***0.619 (0.485-0.762)***

Threshold	Low (0.48)	Base (0.81)	High (1.15)
White Male			
Accuracy	*0.798 (0.791-0.806)*	*0.853 (0.846-0.859)*	*0.819 (0.811-0.826)*
TLR	***0.84 (0.83-0.851)***	***0.844 (0.835-0.855)***	***0.846 (0.835-0.856)***
FMR	*0.074 (0.066-0.082)*	*0.156 (0.147-0.164)*	*0.226 (0.217-0.236)*
FNMR	***0.31 (0.297-0.322)***	***0.131 (0.121-0.142)***	***0.051 (0.042-0.06)***
Non-White Female			
Accuracy	*0.782 (0.771-0.791)*	*0.878 (0.869-0.887)*	*0.863 (0.854-0.872)*
TLR	***0.79 (0.768-0.812)***	***0.794 (0.772-0.815)***	***0.778 (0.756-0.799)***
FMR	*0.055 (0.047-0.064)*	*0.096 (0.087-0.105)*	*0.149 (0.139-0.16)*
FNMR	***0.379 (0.361-0.397)***	***0.173 (0.155-0.19)***	***0.101 (0.085-0.117)***
Black Female			
Accuracy	*0.91 (0.859-0.962)*	*0.962 (0.924-0.987)*	*0.972 (0.93-1.0)*
TLR	0.0 (0.0-1.0)	0.0 (0.0-1.0)	0.0 (0.0-1.0)
FMR	*0.028 (0.0-0.069)*	*0.025 (0.0-0.064)*	*0.029 (0.0-0.072)*
FNMR	xxx [xxx - xxx]	xxx [xxx - xxx]	xxx [xxx - xxx]