Évaluation de l'Équité des Systèmes Biométriques

Kaïra Neily SANON[12] Joël DI MANNO[12] Tanguy GERNOT[1]
Christophe CHARRIER[1] Christophe ROSENBERGER[1]

[1]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR6072, F-14000 Caen, France

[2]FIME EMEA, 14000 Caen, France

Résumé – Les systèmes de reconnaissance biométrique basés sur l'apprentissage automatique sont largement déployés pour authentifier ou identifier des individus. Toutefois, ces systèmes peuvent présenter des variations de performance selon les profils d'utilisateurs ou les conditions d'usage, soulevant des enjeux d'équité. Plusieurs travaux ont proposé des métriques pour mesurer ces biais, mais leur complexité d'interprétation limite souvent leur adoption en pratique. Dans cet article, nous proposons une méthodologie expérimentale permettant d'analyser la sensibilité de ces métriques face à différents biais démographiques, ainsi qu'un nouvel indice de mesure, plus simple et opérationnel. Les résultats obtenus sur trois bases de données faciales et trois modèles de reconnaissance faciale démontrent la pertinence et la facilité d'utilisation de l'approche proposée.

Abstract – Biometric recognition systems based on machine learning are widely deployed to authenticate or identify individuals. However, these systems may exhibit performance variations depending on user profiles or usage conditions, raising concerns about fairness. Several studies have proposed metrics to measure these biases, but their complexity often limits their practical adoption. In this article, we propose an experimental methodology to analyze the sensitivity of these metrics to various demographic biases, along with a new, simpler, and more practical measurement index. The results obtained on three facial datasets and three facial recognition models demonstrate the relevance and ease of use of the proposed approach.

1 Introduction

Les systèmes biométriques sont aujourd'hui largement utilisés pour la sécurité, le contrôle d'accès ou l'identification. Basés sur l'apprentissage automatique, ces systèmes offrent d'excellentes performances, mais plusieurs études ont révélé qu'ils peuvent générer des biais selon les groupes d'utilisateurs ou les conditions d'usage. Ces inégalités proviennent souvent d'un déséquilibre des données ou de facteurs contextuels, et pas uniquement de variables démographiques. De nombreuses métriques ont été proposées pour évaluer ces biais, mais leur complexité limite leur utilisation en pratique.
Dans cet article, nous proposons une méthodologie pour analyser la sensibilité de ces métriques face à différents biais, ainsi qu'un nouvel indice plus simple et opérationnel. Nos expérimentations, menées sur trois bases de données et trois modèles de reconnaissance faciale, confirment la pertinence et la facilité d'utilisation de cette approche.

2 Contexte

Un système biométrique a pour objectif de vérifier ou d'identifier un individu à partir de caractéristiques morphologiques, biologiques ou comportementales propres à chacun. Parmi les modalités les plus courantes, on peut retrouver les empreintes digitales, le visage, l'iris ou la voix. Dans le cas particulier de la reconnaissance faciale, qui constitue le cadre de cette étude, le fonctionnement d'un tel système repose sur plusieurs étapes successives décrites dans [9] et qui se simplifie en ces points : (1) la capture des données, où une image du visage de l'utilisateur est acquise ; (2) l'extraction des caractéristiques, réalisée à l'aide d'un réseau neuronal convolutif (CNN) qui génère un vecteur numérique décrivant le visage ; (3) le stockage de ces vecteurs dans une base de données lors de la phase d'enrôlement ; (4) la comparaison du vecteur extrait lors d'une tentative d'identification ou de vérification avec les vecteurs enregistrés, à l'aide d'une mesure de distance ou de similarité ; et enfin (5) la prise de décision, qui consiste à accepter ou rejeter l'identité en fonction d'un seuil appliqué au score obtenu. Ce processus peut parfois engendrer des discriminations, on dit alors que le système est biaisé.
Dans les systèmes décisionnels, le biais correspond à un écart régulier entre les résultats produits par un modèle et une situation jugée équitable. Ce biais peut avoir plusieurs origines : un déséquilibre dans les données d'apprentissage, un choix de modèle, ou encore des conditions spécifiques d'utilisation. Dans le cas des systèmes biométriques, il se traduit par des variations de performance selon certains groupes d'utilisateurs ou contextes d'acquisition. Ces différences peuvent concerner des critères démographiques (genre, âge, origine), mais aussi des facteurs liés à la qualité des images ou à l'environnement d'utilisation [10]. Un système biométrique est donc considéré comme équitable lorsqu'il garantit un niveau de performance comparable entre les différents groupes ou conditions d'utilisation. De nombreux travaux se sont penchés sur l'évaluation de ces biais afin d'en analyser l'impact.

3 État de l'art

L'évaluation de l'équité des systèmes biométriques a fait l'objet de nombreux travaux ces dernières années, en particulier avec la généralisation des modèles basés sur l'apprentissage automatique. En reconnaissance faciale, plusieurs études ont montré que les performances peuvent varier selon les groupes d'utilisateurs (genre, origine, âge) ou les conditions d'usage. Ces travaux ont cherché à identifier les causes de ces biais,

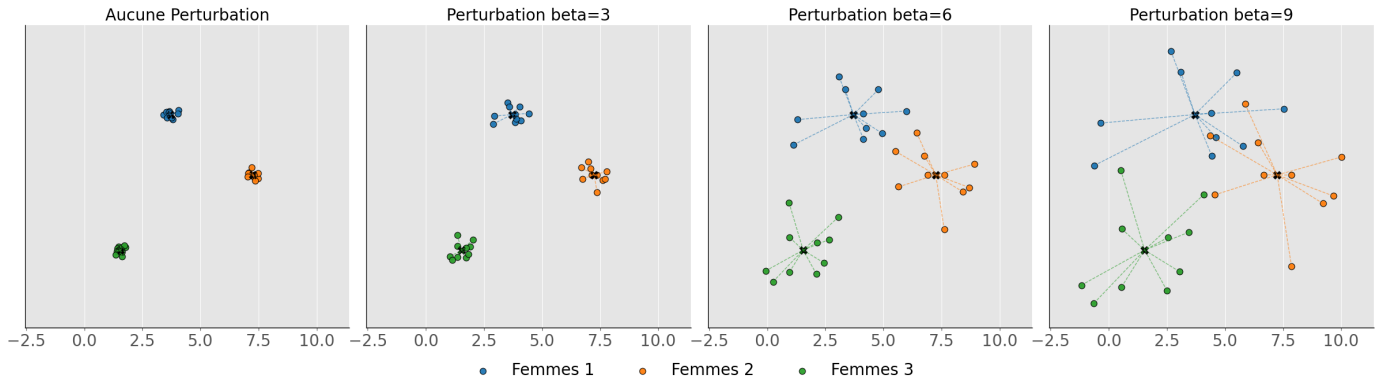

FIGURE 1 : Evolution de la distribution des caractéristiques sous différentes altérations appliquées au groupe cible(femme par exemple)

qu'il s'agisse d'un déséquilibre des données ou de choix algorithmiques favorisant certains profils. D'autres recherches ont proposé des approches statistiques pour déterminer si les écarts observés relèvent d'une fluctuation aléatoire ou d'un déséquilibre réel [13]. Parallèlement, plusieurs contributions ont introduit des indices pour quantifier ces biais. Par exemple, Fang et al. [4] ont proposé un indice conciliant précision et équité en détection d'attaques, tandis que d'autres travaux se sont concentrés sur les systèmes d'authentification, en analysant les distributions de scores ou les taux d'erreur entre groupes [11, 6, 3]. Certaines de ces métriques ont la particularité d'être bornées, facilitant ainsi l'évaluation.

La première métrique considérée est le **Fairness Discrepancy Rate (FDR)** proposée par [1]. Elle vise à mesurer les écarts maximaux de performance entre groupes démographiques, en se basant sur les taux de fausses acceptations (FMR) et de fausses non-correspondances (FNMR) pour un seuil donné τ. Cet indice combine ces deux écarts à l'aide d'un paramètre de pondération α, qui permet d'ajuster l'importance relative accordée à la sécurité (FMR) ou à l'expérience utilisateur (FNMR). La valeur de FDR varie entre 0 (absence d'équité) et 1 (équité parfaite). Le FDR est défini par :

$$\begin{aligned} FDR(\alpha,\tau) &= 1-(\alpha\times A(\tau)+(1-\alpha)\times B(\tau)) \\ A(\tau) &= \max_{i,j}|FMR^{d_i}(\tau)-FMR^{d_j}(\tau)| \\ B(\tau) &= \max_{i,j}|FNMR^{d_i}(\tau)-FNMR^{d_j}(\tau)| \end{aligned} \quad (1)$$

Le **Gini Aggregation Rate for Biometric Equitability (GARBE)** [6] propose une approche différente. Il repose sur le coefficient de Gini, utilisé pour mesurer l'inégalité de répartition des erreurs entre groupes démographiques. Le coefficient de Gini pour un ensemble de valeurs $x = x_1, x_2, \ldots, x_n$ est donné par :

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{2n^2\bar{x}}\right) \quad (2)$$

Dans le cadre du GARBE, deux coefficients de Gini sont calculés, l'un pour les FMR et l'autre pour les FNMR, puis combinés selon un facteur α :

$$GARBE(\alpha,\tau) = \alpha\times A(\tau)+(1-\alpha)\times B(\tau) \quad (3)$$

où $A(\tau)$ et $B(\tau)$ correspondent aux coefficients de Gini associés aux FMR et FNMR pour un seuil τ.

Enfin, le **Separation Fairness Index (SFI)** proposé par [11] évalue la capacité d'un système biométrique à séparer les scores légitimes et imposteurs de façon homogène entre les groupes démographiques. Pour chaque groupe d_i, on considère la différence moyenne entre les scores légitimes μ_{Gi} et imposteurs μ_{Ii}, notée $z_{Si} = |\mu_{Gi} - \mu_{Ii}|$ selon :

$$SFI_N = 1 - \frac{2}{K}\sum_{i=1}^{K}|z_{Si} - z_{S\text{mean}}| \quad (4)$$

où $z_{S\text{mean}}$ est la moyenne des z_{Si} sur l'ensemble des K groupes démographiques.

Ces métriques sont devenues des références pour évaluer l'équité des systèmes biométriques, mais leur complexité et leur sensibilité aux paramètres limitent leur usage en pratique. Nous proposons dans la suite une méthodologie pour analyser leur comportement.

4 Méthodologie proposée

Nous proposons, pour l'évaluation de l'équité, une approche expérimentale consistant, à partir d'un système donné, dont on connaît les performances, d'introduire progressivement un biais contrôlé afin de mesurer l'évolution de ses performances. Plus précisément, considérant un système de reconnaissance faciale composé de n groupes d'utilisateurs, nous ciblons un groupe démographique g_k et altérons progressivement les caractéristiques de ses membres. L'objectif est d'introduire un déséquilibre susceptible d'affecter la reconnaissance des individus de ce groupe. Pour cela, nous appliquons une altération dont l'objet est d'éloigner les représentations des utilisateurs de leur centroïde de groupe sans altérer ce centroïde.

Pour chaque capture faciale A_{ij} réalisée par un individu U_i appartenant au groupe démographique g_k, nous calculons le centroïde global $c(g_k)$ du groupe, correspondant à la moyenne des vecteurs caractéristiques de tous les individus de g_k. L'altération consiste à éloigner progressivement les vecteurs F_{ij}^{original} de ce centroïde, qui reste stable. Concrètement, nous appliquons un facteur d'éloignement β, variant linéairement de 0 à 50, correspondant à dix niveaux de perturbation permettant de se rapprocher de 50% d'erreur de reconnaissance. La transformation appliquée est la suivante :

$$F_{ij}^{\text{altéré}} = F_{ij}^{\text{original}} + \beta\times(F_{ij}^{\text{original}} - c(g_k)) \quad (5)$$

Lorsque $\beta = 0$, les caractéristiques d'origine sont conservées. À mesure que β augmente, les vecteurs s'éloignent de plus en plus du centroïde de leur groupe, augmentant artificiellement la variabilité intra-groupe et dégradant les performances de reconnaissance pour ce groupe spécifique. La figure 1 illustre ce processus d'altération sur deux individus appartenant au même groupe. Pour chaque scénario d'altération, nous extrayons les caractéristiques faciales, calculons les scores de correspondance (légitime et imposteur) et mesurons les performances via les taux FNMR et FMR pour 1 000 valeurs de seuil τ. Une fois les perturbations appliquées, nous calculons les différentes métriques d'équité présentées dans l'état de l'art, pour l'ensemble des niveaux d'altération. Nous mesurons ensuite la sensibilité de chaque métrique en évaluant la corrélation de Spearman entre son évolution et la dégradation des performances du système.
Parallèlement à cette analyse, nous proposons une nouvelle métrique, intitulée Area Max Differential Rate (AMDR), conçue pour fournir une mesure plus intuitive de l'équité.

5 Métrique AMDR

Afin de compléter les métriques existantes, nous proposons une nouvelle mesure d'équité, appelée **Area Max Differential Rate (AMDR)**. Cet indice vise à quantifier simplement et de manière directe l'écart maximal observé entre les performances des différents groupes démographiques, sans nécessiter de paramètres de pondération supplémentaires. Plus précisément, l'AMDR repose sur l'observation des différences relatives entre les taux de fausses acceptations (FMR) et de fausses non-correspondances (FNMR) des groupes considérés. Pour un seuil donné x, l'indice est défini comme suit :

$$\text{AMDR}(x) = \begin{cases} B(x) & \text{si FMR}_{\text{global}} = 0 \\ A(x) & \text{si FNMR}_{\text{global}} = 0 \\ \max\{A(x), B(x)\} & \text{sinon} \end{cases}$$

où :

$$A(x) = \frac{|\text{FMR}^{(g_i)}(x) - \text{FMR}^{(g_j)}(x)|}{\text{FMR}_{\text{global}}}$$

et

$$B(x) = \frac{|\text{FNMR}^{(g_i)}(x) - \text{FNMR}^{(g_j)}(x)|}{\text{FNMR}_{\text{global}}}.$$

L'AMDR reflète ainsi l'écart relatif maximal, ramené aux taux globaux du système, entre deux groupes démographiques g_i et g_j. Sa formulation permet d'identifier rapidement les situations où un groupe est nettement désavantagé par rapport à un autre, en termes de taux d'erreur. Contrairement aux métriques FDR et GARBE, l'AMDR ne requiert aucun paramètre de pondération (*ex.* α) et s'interprète de manière directe : plus sa valeur est élevée, plus l'écart de performance est marqué entre les groupes. Dans la suite de l'article, nous intégrons l'AMDR et les autres métriques à notre protocole expérimental afin d'évaluer sa capacité à détecter les déséquilibres introduits volontairement sur les données.

6 Protocole expérimental

Pour évaluer le comportement des métriques d'équité présentées, nous avons conduit des expériences sur trois systèmes de reconnaissance faciale : 1) **Inception-ResNet V1 [14]** : réseau profond combinant les architectures Inception et ResNet, pré-entraîné sur la base VGGFace2., 2) **ArcFace [2]** : modèle utilisant la fonction de perte Additive Angular Margin Loss, optimisé pour la séparation des identités et 3) **Dlib-ResNet** : réseau léger dérivé de ResNet34 [5], fournissant des vecteurs de dimension 128. Les performances des métriques d'équité ont été évaluées sur trois bases de données faciales, résumées dans le tableau 1) **LFW10** : sous-ensemble de la base LFW [7], composé de 158 sujets disposant chacun d'au moins 10 images, 2) **DemogPairs [8]** : base équilibrée en genre et en ethnicité (Asiatique, Noir, Blanc), comprenant 600 sujets et 3) **AgeDB [12]** : base annotée par tranche d'âge et genre, contenant 1 581 sujets.

Bases	M	F	A	B	W	18-30	31-50	51+	Essais	Annotée
LFW10	124	25	–	–	–	–	–	–	$\geq$10	Oui
DemogPairs	300	300	33%	33%	33%	–	–	–	18	Non
AgeDB	352	220	–	–	–	493	557	531	9	Non

TABLE 1 : Caractéristiques des bases de données (M : Homme, F : Femme, A : Asiatique, B : Noir, W : Blanc).

Pour chaque base et système, nous appliquons les perturbations définies précédemment et calculons, à chaque niveau β, l'AUC ainsi que la métrique d'équité correspondante. Cela nous permet d'obtenir deux vecteurs de 10 valeurs, l'un pour l'AUC et l'autre pour la métrique :

$$\overrightarrow{\text{AUC}} = \begin{bmatrix} \text{AUC}_{\beta_1} \\ \vdots \\ \text{AUC}_{\beta_{10}} \end{bmatrix}, \quad \overrightarrow{\text{Métrique}} = \begin{bmatrix} \text{Métrique}_{\beta_1} \\ \vdots \\ \text{Métrique}_{\beta_{10}} \end{bmatrix}$$

Nous évaluons la sensibilité de chaque métrique à l'aide du coefficient de corrélation de Spearman (SP) entre ces deux vecteurs. Pour les métriques à seuil, celui-ci est fixé à un FMR de 6 %, selon la recommandation de la FIDO Alliance.

7 Résultats expérimentaux

Le tableau 2 illustre l'impact des perturbations intra-variation sur les performances des trois modèles. Sans perturbation (Before), les AUC sont faibles, traduisant une faible séparabilité intra-groupe. Après perturbation (After), les AUC augmentent fortement, indiquant un éloignement contrôlé des représentations utilisateur. On observe toutefois que, pour ArcFace, cette augmentation reste plus limitée sur l'attribut âge. Cela suggère que, pour ce modèle, les représentations initiales sont plus compactes et que la perturbation intra-groupe produit un effet moindre. Ces résultats confirment l'efficacité du protocole pour simuler des déséquilibres, tout en révélant des différences d'impact selon l'architecture du modèle.
Pour chaque base de données et système biométrique, nous mesurons la corrélation entre les valeurs des métriques d'équité et l'AUC aux différents niveaux d'altération. Les résultats de la figure 2 mettent en évidence plusieurs tendances : les métriques sans paramètre de pondération α (SFI, AMDR) montrent une plus grande stabilité face aux perturbations. L'indice AMDR se démarque par ses performances, capturant efficacement les

	LFW10		Demogp Genre		Demogp Ethnie			AgeDB Genre		AgeDB Age		
	M	F	M	F	A	B	W	M	F	18-30	31-50	51+
Inception ResNetv1												
Initial	2.07%	4.24%	1.24%	1.86%	2.02%	1.46%	2.48%	4.28%	7.35%	2.29%	1.32%	1.26%
Altéré	42.96%	40.43%	40.16%	41.12%	40.63%	40.17%	40.92%	37.84%	39.49%	28.15%	38.54%	35.39%
Dlib-Resnet												
Initial	1.07%	3.70%	2.42%	3.04%	8.01%	3.13%	2.43%	8.29%	11.18%	2.89%	1.93%	1.90%
Altéré	41.94%	46.37%	44.11%	44.66%	45.15%	44.15%	44.03%	45.52%	45.21%	34.01%	42.62%	42.33%
ArcFace												
Initial	1.04%	0.25%	1.34%	2.34%	2.59%	1.93%	1.90%	1.83%	2.82%	1.99%	1.99%	0.87%
Altéré	39.10%	38.57%	40.08%	40.82%	40.97%	41.11%	39.57%	9.48%	14.11%	14.18%	11.77%	11.77%

TABLE 2 : Valeur AUC pour les trois modèles profonds et les différents ensembles de données (performance sans aucune altération (Initial), et avec les altérations maximales intraclasse (Altéré)).

variations induites par les altérations. Les métriques intégrant le paramètre α (FDR, GARBE) se révèlent sensibles à son réglage. Une valeur élevée de α devrait privilégier la prise en compte des fausses acceptations (FMR), tandis qu'une valeur faible favoriserait les faux rejets (FNMR). Cette sensibilité est plus marquée pour GARBE que pour FDR. Toutefois, en termes de praticité et d'interprétation, ces métriques peuvent être moins adaptées à un usage opérationnel.

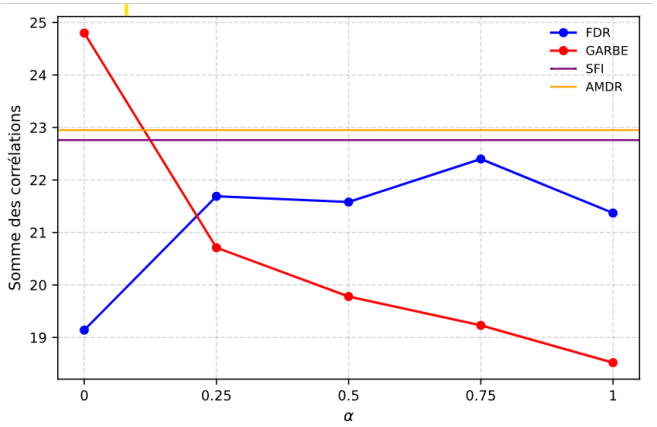

FIGURE 2 : Comparaison des métriques suivant la totalités des bases et des systèmes obtenues par aggrégation des correlations inividuelles de chaque couple bases+systèmes .

8 Conclusion et perspectives

Dans ce travail, nous avons proposé une méthodologie systématique pour évaluer la sensibilité/pertinence des métriques d'équité dans les systèmes biométriques. Nous avons également introduit un nouvel indice, l'AMDR, conçu pour être à la fois simple, interprétable et robuste. Les expériences menées sur plusieurs bases de données et systèmes biométriques ont montré que l'AMDR offre une meilleure stabilité face aux perturbations contrôlées.

Ces résultats offrent des perspectives concrètes pour les industriels et les organismes de certification souhaitant disposer d'indicateurs fiables et facilement exploitables. Un prolongement naturel de cette étude serait d'intégrer de telles métriques directement dans les fonctions d'apprentissage, afin de favoriser l'équité dès la conception des modèles biométriques.

Références

[1] Tiago DE FREITAS PEREIRA et Sebastien MARCEL : Fairness in Biometrics : A Figure of Merit to Assess Biometric Verification Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, janvier 2020.

[2] Jiankang DENG, Jia GUO, Niannan XUE et Stefanos ZAFEIRIOU : ArcFace : Additive Angular Margin Loss for Deep Face Recognition. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, juin 2019. ISSN : 2575-7075.

[3] Alaa ELOBAID, Nathan RAMOLY, Lara YOUNES, Symeon PAPADOPOULOS, Eirini NTOUTSI et Ioannis KOMPATSIARIS : Sum of group error differences : A critical examination of bias evaluation in biometric verification and a dual-metric measure, 2024.

[4] Meiling FANG, Wufei YANG, Arjan KUIJPER, Vitomir STRUC et Naser DAMER : Fairness in face presentation attack detection. *Pattern Recognition*, 147:110002, mars 2024.

[5] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN : Deep Residual Learning for Image Recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, juin 2016. IEEE.

[6] John J. HOWARD, Eli J. LAIRD, Rebecca E. RUBIN, Yevgeniy B. SIROTIN, Jerry L. TIPTON et Arun R. VEMURY : Evaluating Proposed Fairness Models for Face Recognition Algorithms. *In* Jean-Jacques ROUSSEAU et Bill KAPRALOS, éditeurs : *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, Lecture Notes in Computer Science, pages 431–447, Cham, 2023. Springer Nature Switzerland.

[7] Gary B. HUANG, Marwan MATTAR, Tamara BERG et Eric LEARNED-MILLER : Labeled Faces in the Wild : A Database forStudying Face Recognition in Unconstrained Environments. *In Workshop on Faces in 'Real-Life' Images : Detection, Alignment, and Recognition*, Marseille, France, octobre 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

[8] Isabelle HUPONT et Carles FERNÁNDEZ : Demogpairs : Quantifying the impact of demographic imbalance in deep face recognition. *In 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[9] ISO ISO : Iec 19795-1 : Information technology–biometric performance testing and reporting-part 1 : Principles and framework. *ISO/IEC, Editor*, 1(3):5, 2006.

[10] Geraldine JECKELN, Selin YAVUZCAN, Kate A. MARQUIS, Prajay Sandipkumar MEHTA, Amy N. YATES, P. Jonathon PHILLIPS et Alice J. O'TOOLE : Human-Machine Comparison for Cross-Race Face Verification : Race Bias at the Upper Limits of Performance ?, mai 2023. arXiv :2305.16443 [cs].

[11] Ketan KOTWAL et Sebastien MARCEL : Fairness Index Measures to Evaluate Bias in Biometric Recognition, juin 2023.

[12] Stylianos MOSCHOGLOU, Athanasios PAPAIOANNOU, Christos SAGONAS, Jiankang DENG, Irene KOTSIA et Stefanos ZAFEIRIOU : Agedb : the first manually collected, in-the-wild age database. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.

[13] Michael SCHUCKERS, Sandip PURNAPATRA, Kaniz FATIMA, Daqing HOU et Stephanie SCHUCKERS : Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups, août 2022. arXiv :2208.10948 [stat].

[14] Christian SZEGEDY, Sergey IOFFE, Vincent VANHOUCKE et Alex ALEMI : Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, août 2016. arXiv :1602.07261 [cs].