Metrics for Dataset Demographic Bias in Machine Learning: A case study on Facial Expression Recognition

Iris Dominguez-Catena*, Mikel Galar, Daniel Paternáin

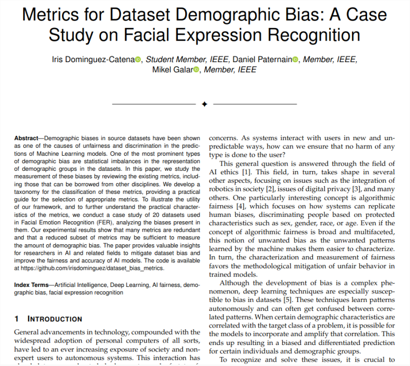
September 2024

Instituto de Smart Cities (ISC), Departmento de Estadística, Informática y Matemáticas
Universidad Pública de Navarra (UPNA)

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

*iris.dominguez@unavarra.es, irisai.neocities.org

The article



Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition

Iris Dominguez-Catena, *Student Member, IEEE*, Daniel Paternain, *Member, IEEE*, Mikel Galar, *Member, IEEE*

Abstract—Demographic biases in source datasets have been shown as one of the causes of unfairness and discrimination in the predictions of Machine Learning models. One of the most prominent types of demographic bias are statistical imbalances in the representation of demographic groups in the datasets. In this paper, we study the measurement of these biases by reviewing the existing metrics, including those that can be borrowed from other disciplines. We develop a taxonomy for the classification of these metrics, providing a practical guide for the selection of appropriate metrics. To illustrate the utility of our framework, and to further understand the practical characteristics of the metrics, we conduct a case study of 20 datasets used in Facial Emotion Recognition (FER), analyzing the biases present in them. Our experimental results show that many metrics are redundant and that a reduced subset of metrics may be sufficient to measure the amount of demographic bias. The paper provides valuable insights for researchers in AI and related fields to mitigate dataset bias and improve the fairness and accuracy of AI models. The code is available at https://github.com/irisdominguez/dataset_bias_metrics.

Index Terms—Artificial Intelligence, Deep Learning, AI fairness, demographic bias, facial expression recognition

1 Introduction

General advancements in technology, compounded with the widespread adoption of personal computers of all sorts, have led to an ever increasing exposure of society and non-expert users to autonomous systems. This interaction has

concerns. As systems interact with users in new and unpredictable ways, how can we ensure that no harm of any type is done to the user?

This general question is answered through the field of AI ethics [1]. This field, in turn, takes shape in several other aspects, focusing on issues such as the integration of robotics in society [2], issues of digital privacy [3], and many others. One particularly interesting concept is algorithmic fairness [4], which focuses on how systems can replicate human biases, discriminating people based on protected characteristics such as sex, gender, race, or age. Even if the concept of algorithmic fairness is broad and multifaceted, this notion of unwanted bias as the unwanted patterns learned by the machine makes them easier to characterize. In turn, the characterization and measurement of fairness favors the methodological mitigation of unfair behavior in trained models.

Although the development of bias is a complex phenomenon, deep learning techniques are especially susceptible to bias in datasets [5]. These techniques learn patterns autonomously and can often get confused between correlated patterns. When certain demographic characteristics are correlated with the target class of a problem, it is possible for the models to incorporate and amplify that correlation. This ends up resulting in a biased and differentiated prediction for certain individuals and demographic groups.

To recognize and solve these issues, it is crucial to

Published in the
IEEE Transactions on Pattern Analysis and Machine Intelligence
(Q1, Impact Factor: 23.6)

`https://doi.org/10.1109/TPAMI.2024.3361979`

Table of contents

Intro to AI bias

Fairness - Bias

- **Unwanted** patterns
  - Both in *data* and *model predictions*
- Based on **protected attributes**
  - Gender, race, age
  - Inherent and immutable
- **Quantifiable**
  - *Group bias* metrics
- Fairness gives us **constraints**, bias gives us **metrics**

Bias sources

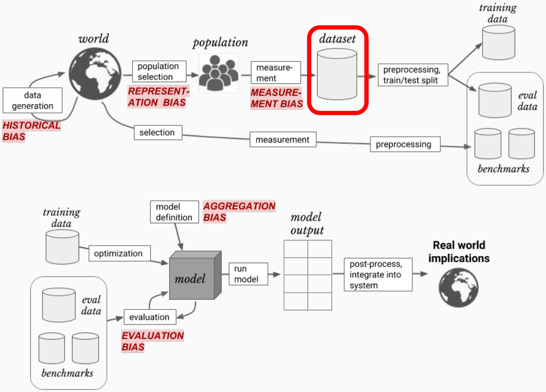


Figure 1: Bias source in the machine learning pipeline[1]

[1] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization. – NY USA: ACM, Oct. 5, 2021, pp. 1–9. ISBN: 978-1-4503-8553-4. DOI: 10.1145/3465416.3483305.

Facial Expression Recognition

Modalities and applications

Modalities

- **Image** or video
- **RGB**, IR, Depth...
- **Discrete** (Ekman's basic emotions) or continuous (NRC-VAD) labeling...

Applications

- Interactive multimedia
- Healthcare [2]
- Assistive robotics [3]
- Public safety [4]

[2] Philipp Werner et al. "Automatic Recognition Methods Supporting Pain Assessment: A Survey". In: *IEEE Transactions on Affective Computing* 13.1 (Jan. 2022), pp. 530–552. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2019.2946774

[3] Ritvik Nimmagadda, Kritika Arora, and Miguel Vargas Martin. "Emotion Recognition Models for Companion Robots". In: *The Journal of Supercomputing* (Mar. 24, 2022). ISSN: 1573-0484. DOI: 10.1007/s11227-022-04416-4

[4] **Mou2023**

A real example

Figure 2: A sample of FER2013/FER+, a popular FER dataset[5].

[5] Emad Barsoum et al. “Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ICMI ’16: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION. Tokyo Japan: ACM, Oct. 31, 2016, pp. 279–283. ISBN:

FER and FER-related known biases

- Gender and skin tone (Fitzpatrick Skin Type) in gender classification[6]
- FER research models[7]: capacitism, age, race and gender
- Commercial FER systems[8]: age, race and gender

[6] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 23–24, 2018, pp. 77–91.

[7] Jacqueline J. Greene et al. “The Spectrum of Facial Palsy: The MEEI Facial Palsy Photo and Video Standard Set”. In: *The Laryngoscope* 130.1 (2020), pp. 32–37. ISSN: 1531-4995. DOI: `10.1002/lary.27986`; Tian Xu et al. “Investigating Bias and Fairness in Facial Expression Recognition”. In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien Bartoli and Andrea Fusiello. Cham: Springer International Publishing, 2020, pp. 506–523. ISBN: 978-3-030-65414-6. DOI: `10.1007/978-3-030-65414-6\_35`.

[8] Eugenia Kim et al. “Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, July 21, 2021, pp. 638–644. ISBN: 978-1-4503-8473-5; Khurshid Ahmad et al. “Comparing the Performance of Facial Emotion Recognition Systems on Real-Life Videos: Gender, Ethnicity and Age”. In: *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1*. Ed. by Kohei Arai. Vol. 358. Cham: Springer International Publishing, 2022, pp. 193–210. ISBN: 978-3-030-89905-9 978-3-030-89906-6. DOI: `10.1007/978-3-030-89906-6\_14`.

FER datasets over time

Short name	Year	Collection	Images	Videos	Subjects
POFA	1976	Lab	110	-	16
JACFEE	1988	Lab	56	-	56
AR-Face	1998	Lab	4,000	-	126
JAFFE	1998	Lab	213	-	10
KDEF	1998	Lab	4,900	-	70
CK	2000	Lab	8,795	486	97
CK+	2010	Lab	10,727	593	123
MUG	2010	Lab	70,654	-	52
Multi-PIE	2010	Lab	750,000	-	337
RaFD	2010	Lab	8,040	-	67
SFEW	2011	ITW-M	1,766	-	330
FER2013	2013	ITW-I	32,298	-	-
WSEFEP	2014	Lab	210	-	30
ADFES	2016	Lab	-	648	22
FERPlus	2016	ITW-I	32,298	-	-
Aff-Wild2	2017	ITW-I	-	558	-
AffectNet	2017	ITW-I	291,652	-	-
ExpW	2017	ITW-I	91,793	-	-
RAF-DB	2017	ITW-I	29,672	-	-
CAER-S	2019	ITW-M	70,000	-	
SEWA	2019	ITW-I	-	199	398
MMAFEDB	2020	ITW-I	128,000	-	-
NHFIER	2020	ITW-I	5,558	-	-

- **Laboratory-gathered** (*Lab*): limited selection of subjects under controlled conditions. High-quality, low quantity.
- **In The Wild** (*ITW*): unknown subject identities and demographies. Low-quality, high quantity.
    - From **Internet queries** (*ITW-I*).
    - From **Motion Pictures** (*ITW-M*).

Important note

⚠

The demographic information for these datasets is not available.

We employ an approximation as predicted by FairFace[9].

[9] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1547–1557. ISBN: 978-1-66540-477-8. DOI: 10.1109/WACV48630.2021.00159.

Representational bias

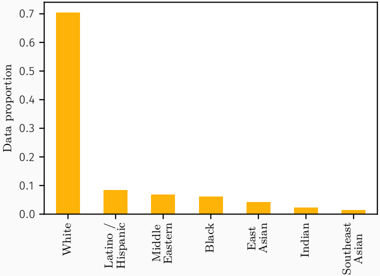


Figure 3: Apparent race distribution in FER+.

[9] (Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (AISafety 2022).* Thirty-First International Joint Conference on Artificial Intelligence and the Twenty-Fifth European Conference on Artificial Intelligence (IJCAI-ECAI-2022). Vienna, Austria, July 24–25, 2022)

Representational bias

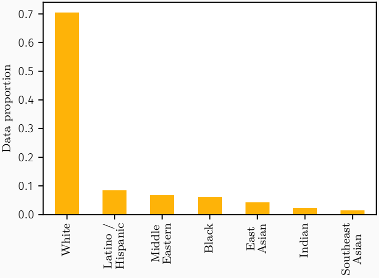


Figure 3: Apparent race distribution in FER+.

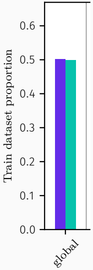


Figure 4: Apparent *per-label* gender distribution in FER+.

[9] (Dominguez-Catena, Paternain, and Galar, "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition")

Representational bias

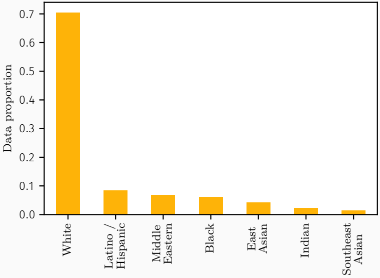


Figure 3: Apparent race distribution in FER+.

Stereotypical bias

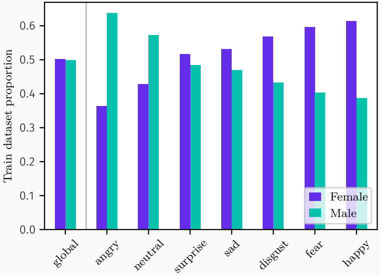


Figure 4: Apparent *per-label* gender distribution in FER+.

[9] (Dominguez-Catena, Paternain, and Galar, “Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition”)

Stereotypical bias, an example I

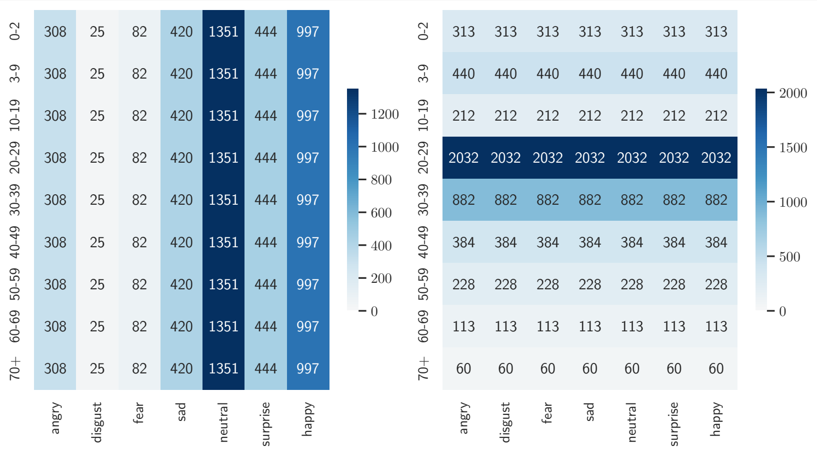


Figure 5: Contingency tables of two datasets **without** stereotypical bias

Stereotypical bias, an example II

	angry	disgust	fear	sad	neutral	surprise	happy
0-2	186	15	49	254	817	268	603
3-9	261	21	69	356	1146	376	846
10-19	126	10	33	172	554	182	409
20-29	1207	100	323	1648	5296	1741	3908
30-39	524	43	140	715	2299	755	1696
40-49	228	18	61	311	1000	328	738
50-59	136	11	36	185	596	196	440
60-69	67	5	18	92	295	97	218
70+	35	2	9	48	156	51	115

Figure 6: Contingency tables of a datasets **without** stereotypical bias

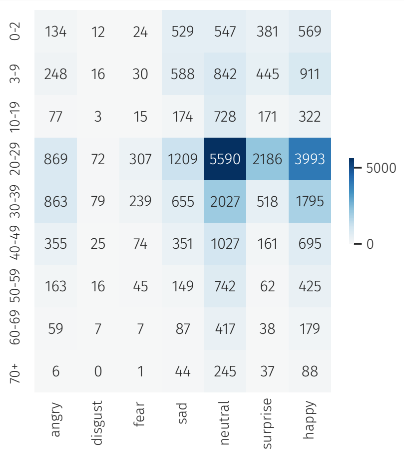


Figure 7: Real contingency table of a FER dataset **with** stereotypical bias

Bias metrics

- Representational
  - Richness
    - Richness
  - Evenness
    - Shannon Evenness Index
    - Normalized Standard Deviation
  - Dominance
    - Imbalance Ratio
    - Berger-Parker Index
  - Combined
    - Effective Number of Species
    - Simpson Index
    - Simpson’s Reciprocal
    - Simpson’s Index of Diversity
    - Shannon Entropy

- Stereotypical
  - Global
    - Cramer’s V
    - Tschuprow’s T
    - Pearson’s Contingency Coefficient
    - Theil’s U
    - Normalized Mutual Information
  - Local
    - Normalized Pointwise Mutual Information
    - Ducher’s Z

---

[9] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition*. Mar. 28, 2023. DOI: `10.48550/arXiv.2303.15889`. arXiv: `2303.15889 [cs]`. URL: `http://arxiv.org/abs/2303.15889` (visited on 05/26/2023). preprint

Methodology

1. Dataset preprocessing, homogenize images and labels
2. Demographic analysis of the datasets
    - FairFace[10]
3. Measure bias with all metrics
    - 3 demographic axis: age, gender, and race
    - + intersectional axis (Cartesian product)
4. Analyze the correlation between metrics
    - Discard *redundant* metrics, prioritizing *interpretable* metrics

[10] Karkkainen and Joo, “FairFace”.

Metric selection

Representational bias metric coherence

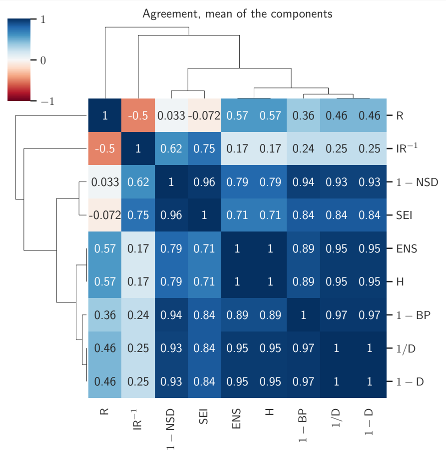


Figure 8: Spearman's ρ correlation between representational bias metrics

[10]Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

Stereotypical bias metric coherence

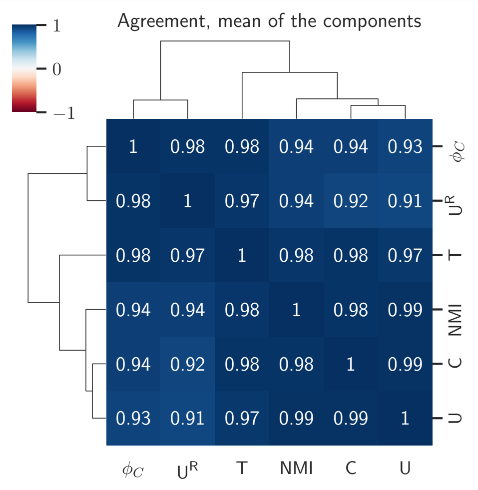


Figure 9: Spearman's ρ correlation between stereotypical bias metrics

[10] Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

Best choices

- General representational
  - **Effective Number of Species** (ENS)
- Evenness between represented groups
  - **Shannon Evenness Index** (SEI)
- Good approximation: Dominance
  - **Berger-Parker Index** (BP)
- Stereotypical bias (global)
  - **Cramer's V** (ϕ_C)
- Stereotypical bias (local)
  - **Ducher's Z** (Z)

Metrics in action

		Laboratory: Average	ITW-I: Average
Age	9 − ENS	7.067 ± 0.932	3.334 ± 0.286
	1 − SEI	0.409 ± 0.200	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.104 ± 0.027
Race	7 − ENS	5.168 ± 0.634	3.724 ± 0.321
	1 − SEI	0.384 ± 0.151	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.063 ± 0.018
Gender	2 − ENS	0.139 ± 0.280	0.005 ± 0.005
	1 − SEI	0.039 ± 0.052	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.167 ± 0.018

Figure 10: Average representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of lab (left) and ITW-I (right) datasets.

[10]Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

Local stereotypical bias

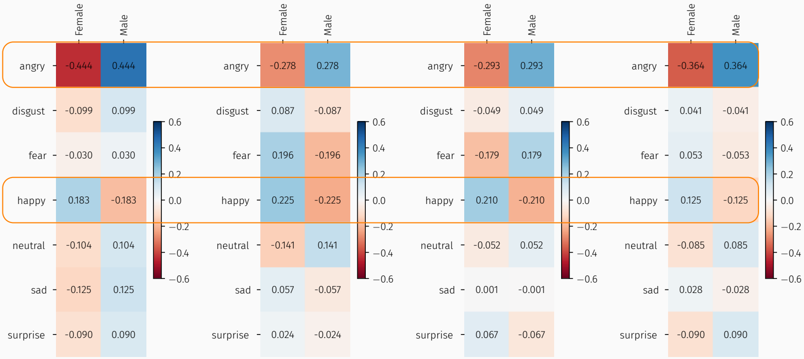


Figure 11: Local stereotypical bias for gender in Affectnet, Fer+, NHFIER y Raf-DB (Ducher's Z). (F: Female, M: Male)

Conclusion

Conclusion

- Dataset bias measurement is necessary for a more **fair** AI
- A reduced set of bias metrics is enough to characterize bias **in practice**
- Datasets are biased, and the biases **are changing over time**

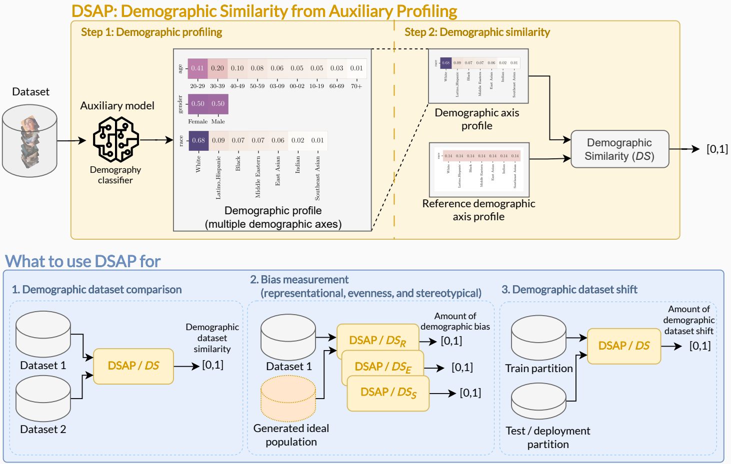
¿Questions?

✉ iris.dominguez@unavarra.es

https://irisai.neocities.org

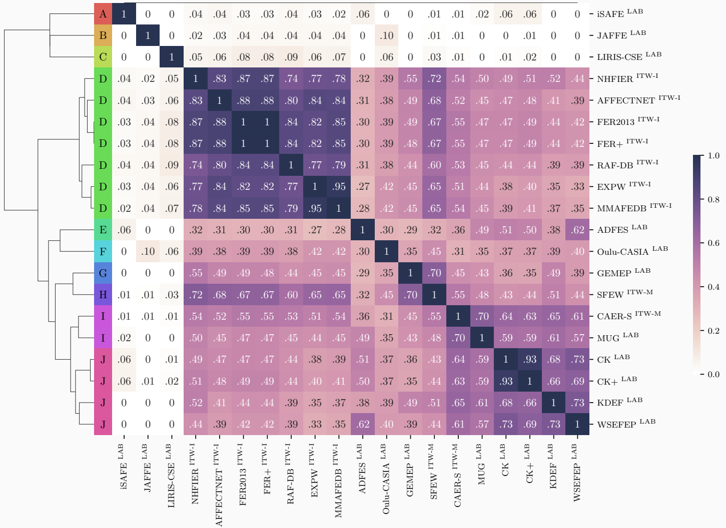
What’s next?

Dataset comparison



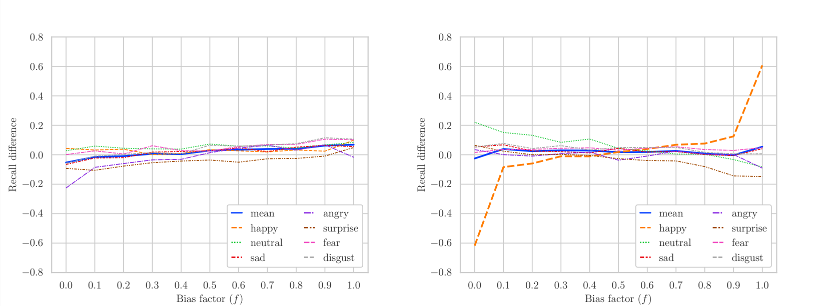
[10]Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *DSAP: Analyzing Bias Through Demographic Comparison of Datasets*. Dec. 22, 2023. DOI: `10.48550/arXiv.2312.14626`. arXiv: `2312.14626 [cs]`. URL: `http://arxiv.org/abs/2312.14626` (visited on 01/24/2024). preprint

Dataset comparison



[10] Dominguez-Catena, Paternain, and Galar, *DSAP*

Results on propagation



(a) **Difference in recalls** (F-M) under representational bias

(b) **Difference in recalls** (F-M) under stereotypical bias (happy)

Figure 12: Recall difference (female recall minus male recall) for the representationally (a) and stereotypically (b) biased datasets.

---

[10] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. “Gender Stereotyping Impact in Facial Expression Recognition”. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Vol. 1752. Cham: Springer Nature Switzerland, 2023, pp. 9–22. ISBN: 978-3-031-23617-4 978-3-031-23618-1. DOI: `10.1007/978-3-031-23618-1_1`

Formulas

Effective Number of Species (ENS)[11]:

$$\mathrm{ENS}(X) = \exp\left(-\sum_{g\in G} p_g \ln p_g\right) \ . \tag{1}$$

Adjusted entropy. *Effective* number of represented group.

[11] Lou Jost. “Entropy and Diversity”. In: *Oikos* 113.2 (May 2006), pp. 363–375. ISSN: 00301299. DOI: 10.1111/j.2006.0030-1299.14714.x.

Shannon Evenness Index (SEI)[12]:

$$\mathrm{SEI}(X) = \frac{\mathrm{H}(X)}{\ln(\mathrm{R}(X))}\ , \tag{2}$$

where $\mathrm{H}(X)$ is Shannon entropy.

Group evenness.

[12] E.C. Pielou. “The Measurement of Diversity in Different Types of Biological Collections”. In: *Journal of Theoretical Biology* 13 (Dec. 1966), pp. 131–144. ISSN: 00225193. DOI: 10.1016/0022-5193(66)90013-0.

Berger-Parker Index (BP)[13]:

$$\mathrm{BP}(X) = \frac{\max_{g \in G} n_g}{n} \,. \tag{3}$$

Ratio between the most represented group and the whole population.

[13] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (June 12, 1970), pp. 1345–1347. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.168.3937.1345.

Cramer's V (ϕ_C)[14]:

$$\chi^2(X) = \sum_{g \in G} \sum_{y \in Y} \frac{(n_{g \wedge y} - \frac{n_g n_y}{n})^2}{\frac{n_g n_y}{n}}, \tag{4}$$

$$\phi_C(X) = \sqrt{\frac{\chi^2(X)/n}{\min(|G| - 1, |Y| - 1)}}, \tag{5}$$

[14] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. Princeton: Princeton university press, 1991, p. 282. ISBN: 978-0-691-08004-8.

Ducher's Z (Z)[15]:

$$Z(X,g,y)=\begin{cases}\frac{p_{g\wedge y}-p_g p_y}{\min[p_g,p_y]-p_g p_y} & \text{if } p_{g\wedge y}-p_g p_y>0\\ \frac{p_{g\wedge y}-p_g p_y}{p_g p_y-\max[0,p_g+p_y-1]} & \text{if } p_{g\wedge y}-p_g p_y<0\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

[15] M. Ducher et al. "Statistical Relationships between Systolic Blood Pressure and Heart Rate and Their Functional Significance in Conscious Rats". In: *Medical & Biological Engineering & Computing* 32.6 (Nov. 1994), pp. 649–655. ISSN: 0140-0118, 1741-0444. DOI: 10.1007/BF02524241.