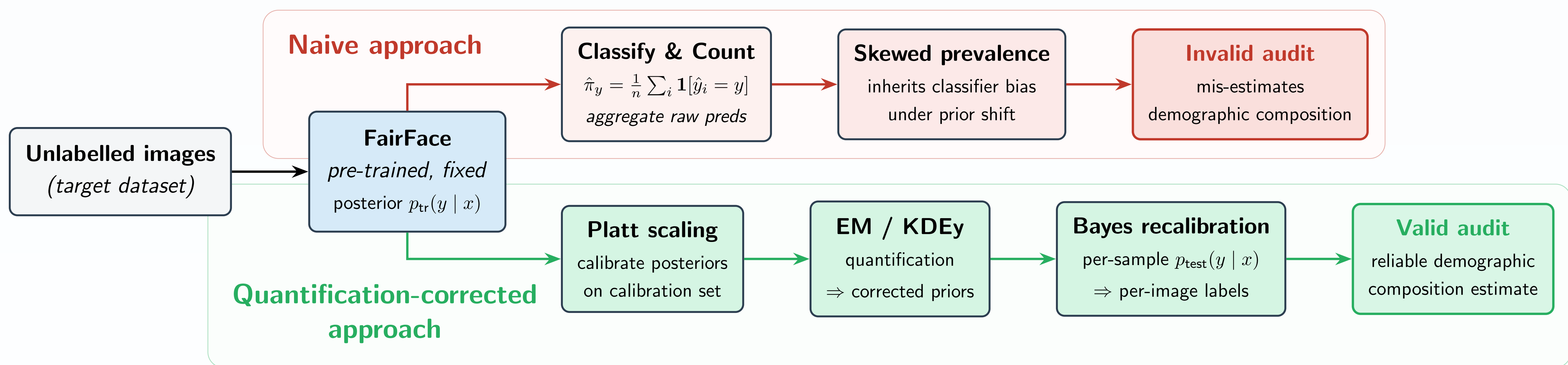


Better Demographic Profiling of Datasets through Quantification

Iris Dominguez-Catena¹, Pablo González², Juan José del Coz², Daniel Paternain¹, Mikel Galar¹

¹Universidad Pública de Navarra ²Artificial Intelligence Center, University of Oviedo
iris.dominguez@unavarra.es



1 Motivation

Modern image datasets like LAION (~5 B images) are too large for manual demographic labelling, so we rely on auxiliary demographic classifiers like FairFace to annotate characteristics such as race, gender, and age. The problem is that **classifier accuracy degrades under prior probability shift** (when the target population differs from the training distribution) so naive aggregate counts inherit the classifier's class-conditional error rates and become biased.

Our goal is to correct these biases, yielding reliable demographic profiles and per-sample labels for bias auditing in large datasets.

2 Methodology

- Start from a **fixed, pre-trained** demographic model (FairFace).
- Calibrate** posteriors via Platt scaling on a calibration set.
- Apply **quantification** (EM or KDEy) to estimate class prevalences $\hat{\pi}_y$ on the target dataset $\Rightarrow$ improved global demographic profiles.
- Recalibrate** per-sample posteriors [1] $\Rightarrow$ improved per-image demographic predictions.

Two quantification methods:

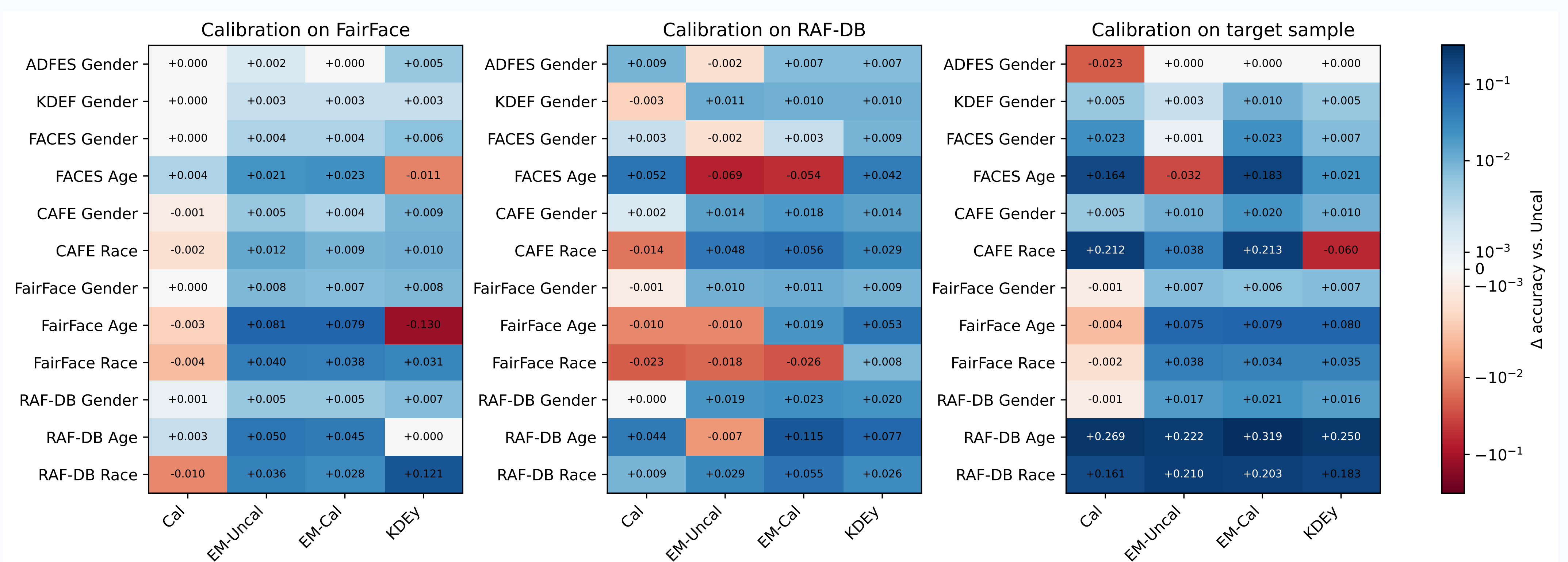
- EM (Expectation-Maximization) [2]**: iteratively re-estimates priors from posterior probabilities. Sensitive to calibration quality.
- KDEy [3]**: models class-conditional score distributions via kernel density estimation. Robust to poor calibration but higher compute cost.

3 Experimental setup

- 6 face datasets**: FairFace, RAF-DB, ADFES, KDEF, FACES, CAFE
- 3 attributes**: race (nominal), gender (binary), age (ordinal)
- 3 calibration scenarios**:
 - Calibrate on FairFace train (same as model training)
 - Calibrate on RAF-DB train (external, coarser labels)
 - Calibrate on a 40% labeled subset of the target dataset
- 5 methods compared**: Uncalibrated (baseline), Platt calibration, EM-Uncal, EM-Cal, KDEy
- Evaluation**: Artificial Prevalence Protocol (APP), mean accuracy across bags with diverse group compositions.

4 Results

Each cell shows the **accuracy change (Δ)** relative to the uncalibrated model baseline. Positive values (blue) indicate improvement; negative values (red) indicate degradation. The three panels correspond to the three calibration scenarios.

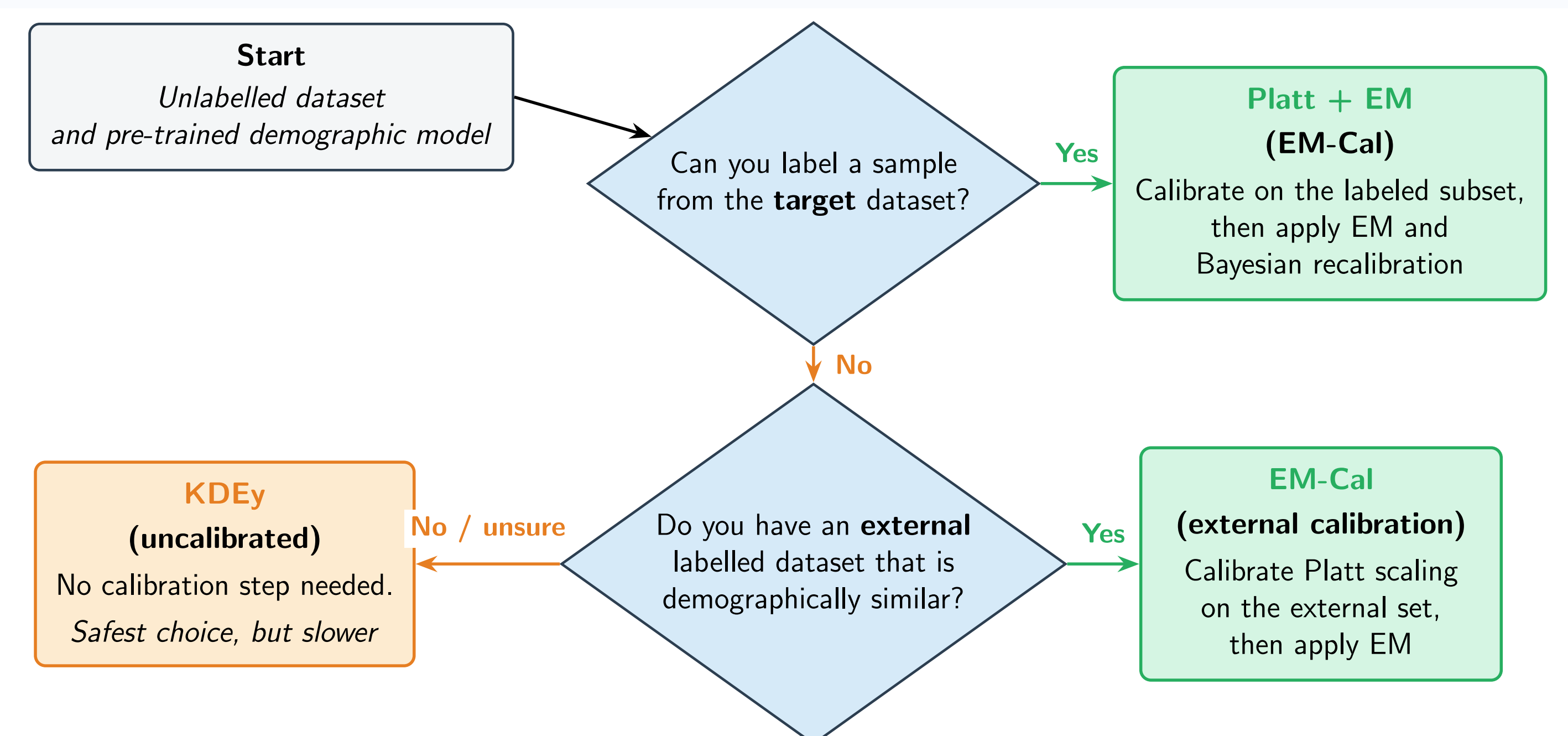


5 Key findings

- Quantification consistently helps**: EM and KDEy improve over both Uncal and Cal baselines across nearly all datasets and attributes.
- Calibration alignment matters**: EM-Cal excels when the calibration set resembles the target (Scenarios 2–3). Under poor alignment (Scenario 1), EM amplifies calibration error.
- KDEy: robust runner-up**. Higher compute cost but seems more robust to calibration scenario.
- Works beyond binary**: gains observed for all attribute codifications.

6 Conclusion & practical recommendations

Quantification techniques, originally designed for aggregate prevalence estimation, also improve **per-sample** demographic predictions from a fixed pre-trained model, enabling more reliable **bias audits** of large, unlabeled image datasets.



Acknowledgments

This work has been partially funded by projects PID2022-136627NB-I00 and PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033 and FEDER, EU).

Bibliography

- Amr Alexandari et al. "Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation". In: *International conference on machine learning*. PMLR. 2020, pp. 222–232.
- Marco Saerens et al. "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure". In: *Neural computation* 14.1 (2002), pp. 21–41.
- Alejandro Moreo et al. "Kernel density estimation for multiclass quantification". In: *Machine Learning* 114.4 (2025), p. 92.