A Comprehensive Analysis of AI Biases in Deepfake Detection With Massively Annotated Databases

Ying Xu*, Philipp Terhörst*, Kiran Raja, Marius Pedersen

Colourlab

Colourlab, Department of Computer Science,
Norwegian University of Science and Technology, Gjøvik, Norway

Abstract

This study examines the bias issue in public Deepfake datasets and its impact on Deepfake detection models. It provides annotations for 47 attributes in five popular Deepfake datasets and analyses the bias of three state-of-the-art Deepfake detection backbone models. The investigation covers demographic and non-demographic attributes such as gender, ethnicity, hair, skin, and accessories. The results reveal a lack of diversity in the datasets and significant bias in the detection models towards various attributes. These biased models may produce incorrect detection results, posing challenges in generalizability, fairness, and security. The study aims to raise awareness and offer annotation databases to help address bias in future Deepfake detection techniques.

Experimental setup

Deepfake Datasets

- Celeb-DF, DFD, FF++, DF1.0, DFDC
- 30 frames per video with 10-frame intervals

Deepfake Detection Backbones

- EfficientNet-B0
- Xception
- Capsule-Forensics-v2

Evaluation Metrics

- Balanced error
- Corrected relative performance

Project Page

Database Annotations

Massive and diverse annotations for five widely-used Deepfake detection datasets

- MAAD-Face classifier[1]
- **47** attributes: Demographic & non-demographic
- Over **65.3**M annotations

Male	-1	Bangs	-1	Round Face	0	Smiling	1
Young	1	Sideburns	-1	Double Chin	0	Big Lips	0
Middle Aged	-1	Black Hair	0	High Cheekbones	0	Big Nose	0
Senior	-1	Blond Hair	-1	Chubby	0	Pointy Nose	-1
Asian	0	Gray Hair	-1	Obstructed Forehead	0	Heavy Makeup	1
White	-1	No Beard	1	Fully Visible Forehead	-1	Wearing Hat	0
Black	1	Mustache	-1	Brown Eyes	1	Wearing Necktie	-1
Shiny Skin	1	5'o Clock Shadow	-1	Bags Under Eyes	0	Wearing Lipstick	0
Bald	-1	Goatee	-1	Bushy Eyebrows	0	No Eyewear	0
Wavy Hair	0	Oval Face	0	Arched Eyebrows	0	Eyeglasses	1
Receding Hairline	0	Square Face	-1	Mouth Closed	-1	Attractive	0

1: Positive -1: Negative 0: Undefined

Measuring Bias on Unbalanced Data

Corrected relative performance (CRP)

$$CRP(a) = RP_{data}(a) - RP_{control}(a)$$

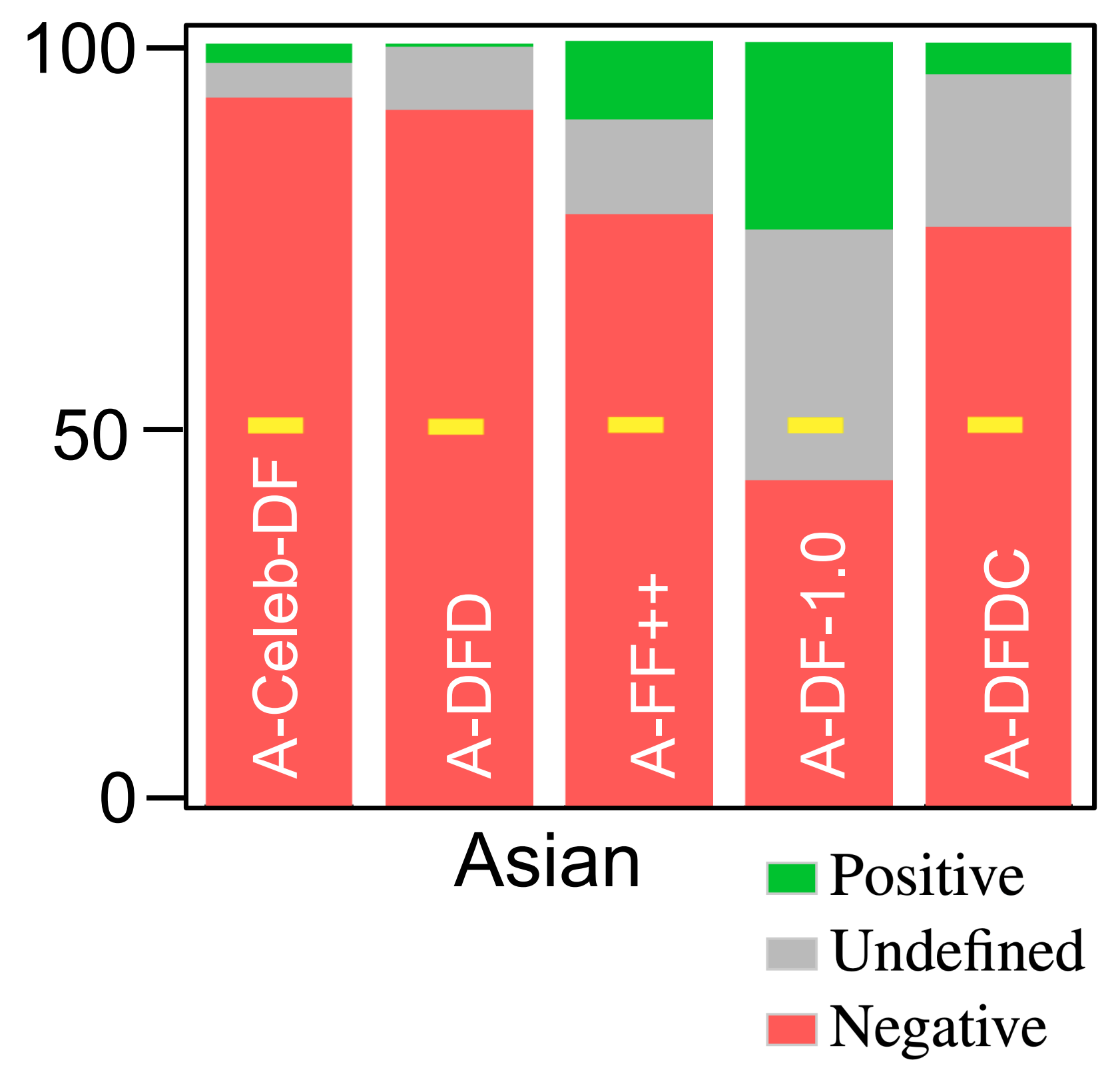
Where $RP_{type}(a) = 1 - \frac{err^{(+)}_{type}(a)}{err^{(-)}_{type}(a)}$, *type = {data, control}*

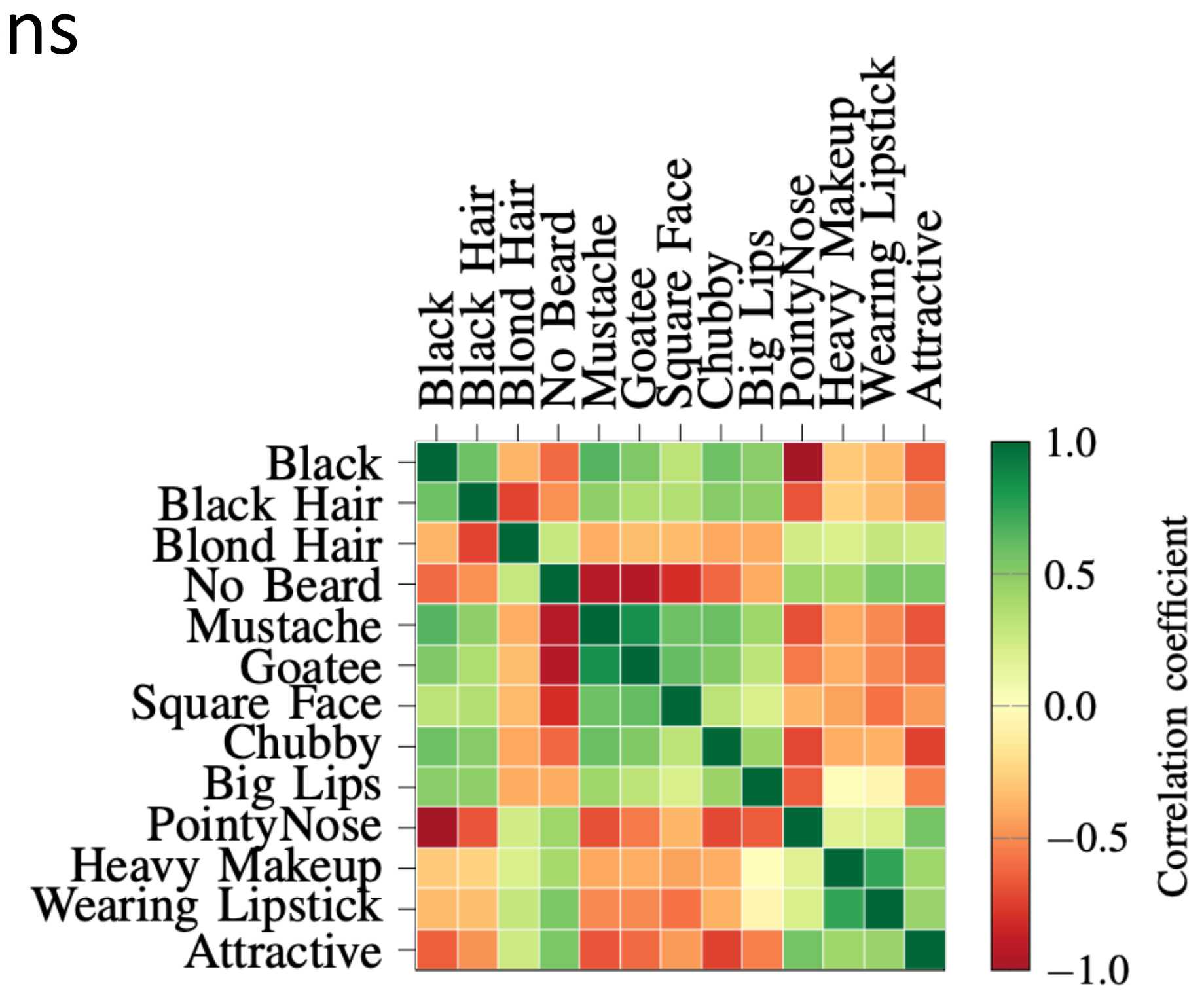
$RP_{type}(a)$ measures the performance differences for an attribute a based on the error rates for a positive $err^{(+)}_{type}(a)$ and a negative $err^{(-)}_{type}(a)$ group

$PDRP$ - CRP for pristine data, $DDRP$ - CRP for Deepfake data

Analysing Database Annotations

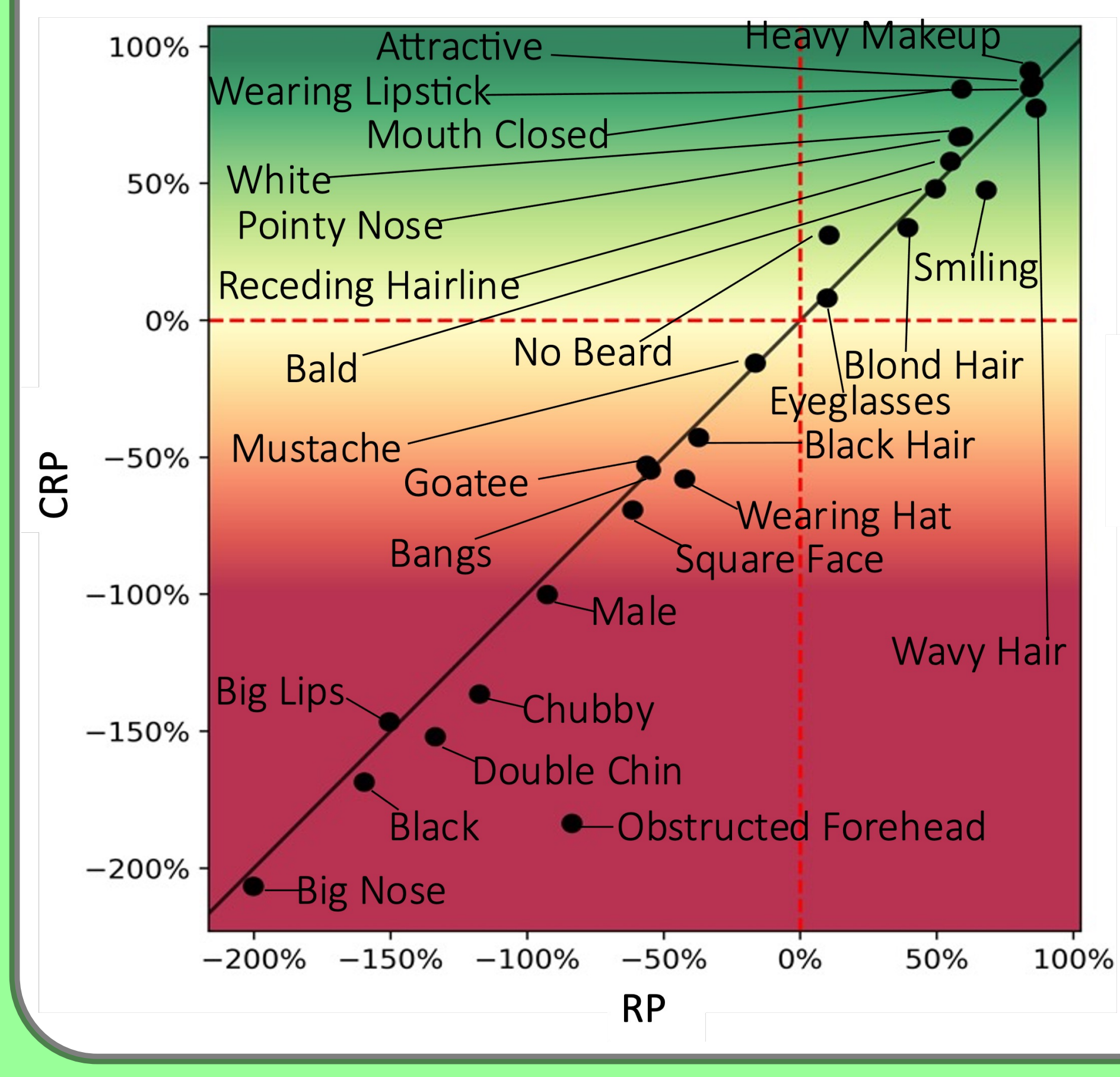
Strong imbalance and correlations

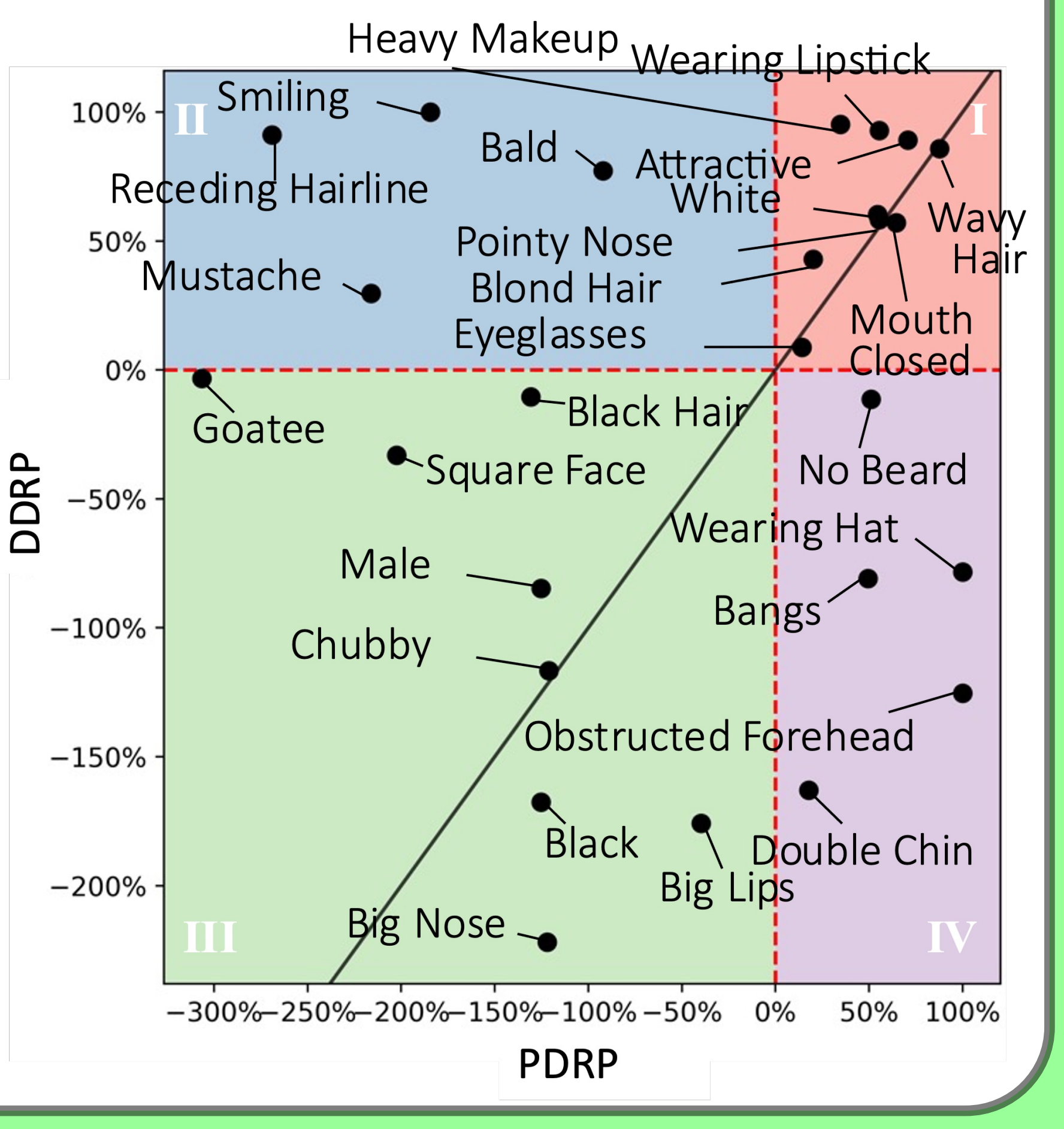




Analysing Bias in Deepfake Detection

EfficientNet-B0 with A-Celeb-DF





Conclusion

We provided large-scale annotations for five popular Deepfake detection datasets and used these to comprehensively analyse bias in Deepfake detection. Our main findings are

- Deepfake detection databases and strong imbalance.
- Current Deepfake detection databases contain some strongly correlating attribute pairs.
- The analysed Deepfake detection backbone models are strongly biased for many demographic and non-demographic attributes.
- For many of the investigated attributes, the biased performance similarly affects the pristine and Deepfake data.
- The results suggest that the model tends to learn questionable assumptions where a certain attribute is present.
- The presence of a certain attribute in a Deepfake image resulted in an increased error rate, several times higher than for a Deepfake without this attribute.

Future works:

- Creating more unbiased, balanced, and diverse Deepfake datasets
- Developing bias-mitigating Deepfake detection solutions

[1] "MAAD-Face: a massively annotated attribute dataset for face images." IEEE Transactions on Information Forensics and Security 16 (2021): 3942-3957.