Titre: Title:	Assess and quantify DNN classifier bias using likelihood prediction
Auteurs: Authors:	Benjamin Prosper Paul Djian, Ettore Merlo, Sébastien Gambs, & Rosin Claude Ngueveu
Date:	2024
Type:	Communication de conférence / Conference or Workshop Item
Référence: Citation:	Djian, B. P. P., Merlo, E., Gambs, S., & Ngueveu, R. C. (mai 2024). Assess and quantify DNN classifier bias using likelihood prediction [Affiche]. 5e Forum Mobilit.AI, Montréal, Qc, Canada (1 page). https://publications.polymtl.ca/59631/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/59631/
Version:	Version officielle de l'éditeur / Published version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	

Nom de la conférence: Conference Name:	5e Forum Mobilit.AI
Date et lieu: Date and Location:	2024-05-28 - 2024-05-29, Montréal, Qc, Canada
Maison d'édition: Publisher:	
URL officiel: Official URL:	
Mention légale: Legal notice:	

ASSESS AND QUANTIFY DNN CLASSIFIER BIAS USING LIKELIHOOD PREDICTIONS

FORUM MobiliT.AI

AUTHORS

Benjamin Djian - benjamin.djian@polymtl.ca
Ettore Merlo - ettore.merlo@polymtl.ca
Sébastien Gambs - gambs.sebastien@uqam.ca
Rosin Claude Ngueveu - rosin.ngueveu@polymtl.ca

PARTENAIRES

POLYTECHNIQUE MONTRÉAL

UQÀM

INTRODUCTION

Fully connected deep neural networks are trained on an abundant amount of data. In some contexts, these data may hold **sensitive information**.
Machine learning models may differentiate outputs on sensitive attributes, like **race**, **sex**, or **age**.

OBJECTIVES

With a database of individuals and a classifier designed to classify these profiles based on their salaries, we aim to :

- Investigate the use of **Computational Profile Likelihood** [1] to show that ethically sensitive profiles are distinguishable.
- Measure the impact of a bias mitigation technique [2] on CPL measurements.

METHODOLOGY

For each neuron of the penultimate layer of the classifier, histograms of activation levels are constructed [1].

The **Computational Profile Likelihood (CPL)** of a profile is computed by considering where corresponding activation levels "fall" into the histograms. A higher CPL for a profile means that the profile is more likely to belong to the distribution used to construct the histograms.

We considered the **Adult Census database** [3] and a binary classifier trained to assign class "**Low Revenue**" or "**High Revenue**" to profiles of individuals. We have focused on possible variations between **Male** and **Female** individuals.

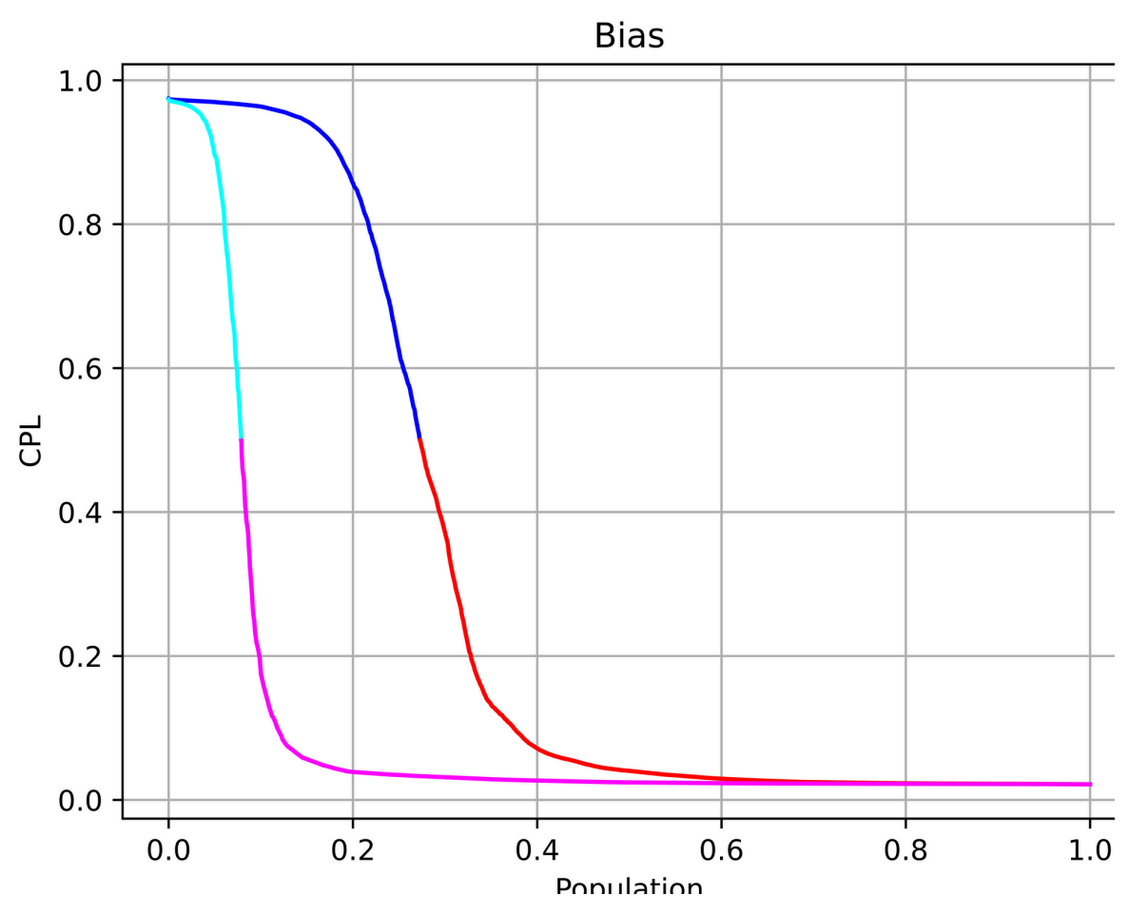

Figure 1
Original data

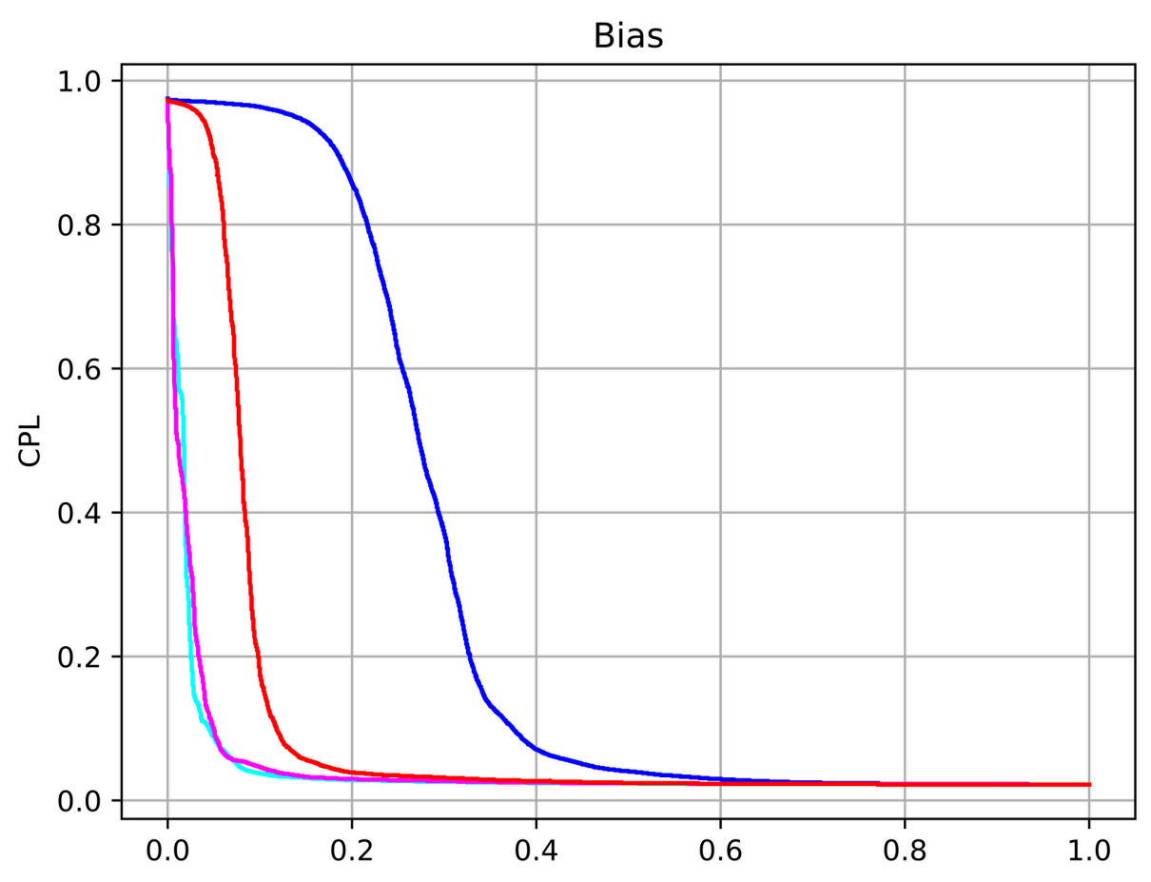

Figure 2
Moderate bias mitigation

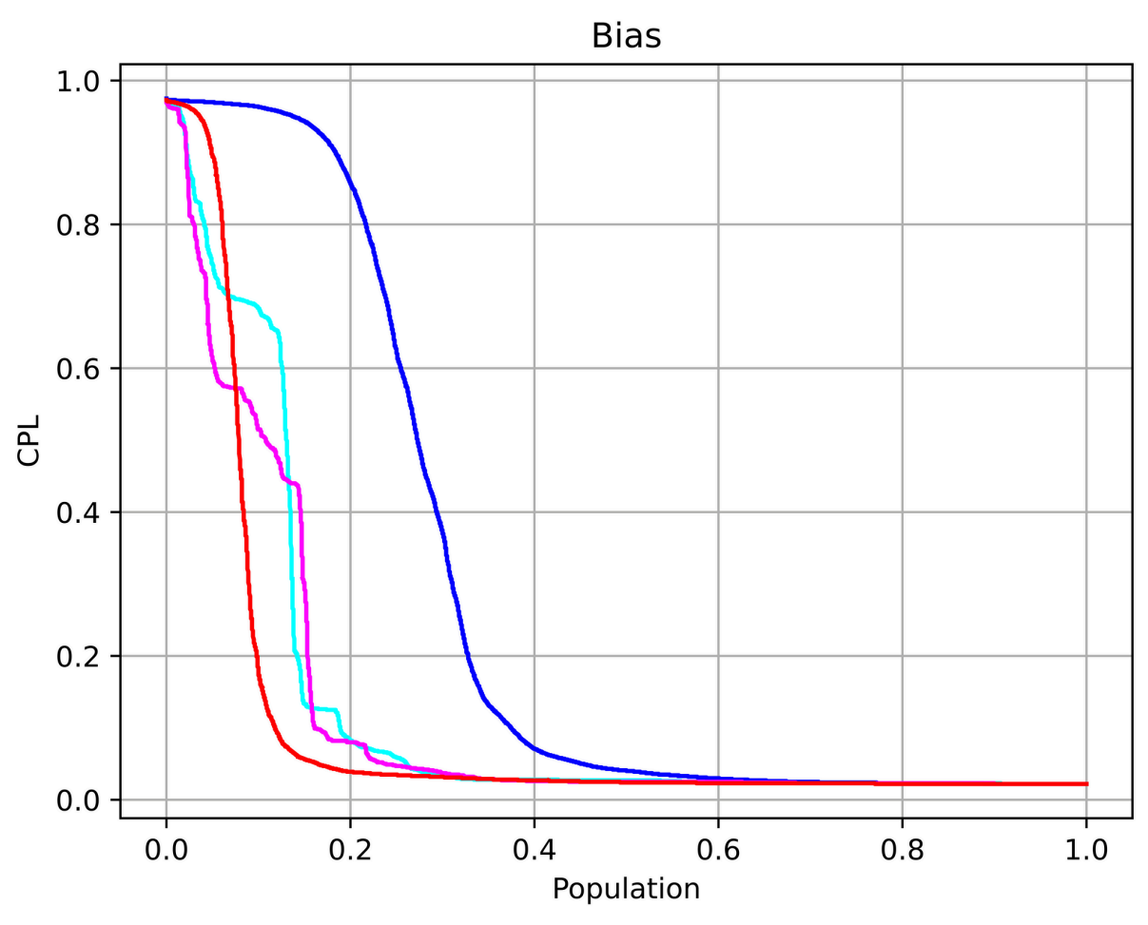

Figure 3
Strong bias mitigation

RESULTS

Figure 1 to Figure 3 : Normalized CPL against portion of the respective population.

Figure 1 : Male curve in red and blue, Female curve in cyan and magenta
Figure 2 and 3 : Male curve in blue, sanitized Male curve in cyan, Female curve in red, sanitized Female curve in magenta

Figure 4 and 5 : Scatter representation of all inputs, where X-axis is the distribution of Female Low Revenue, and Y-axis is the distribution of Male Revenue

ANALYSIS

Figure 1 clearly distinguishes between the curve of Male individuals and the curve of Female individuals. A very restrictive proportion of Female profiles are more likely to have high revenues than low revenues (approximately **10% of the population**). Males are much more likely to have high revenues because **30% of the population** has higher CPL for this class.

In **Figure 2**, cyan and magenta curves are less distinguishable than blue and red curves. Cyan and magenta are also skewed toward the low-revenue class.

In **Figure 3**, cyan and magenta are less distinguishable than in Figure 2, but still more than blue and red curves. The distributions are also less skewed towards the low revenues class than in Figure 2, but still more than the blue and red curves.

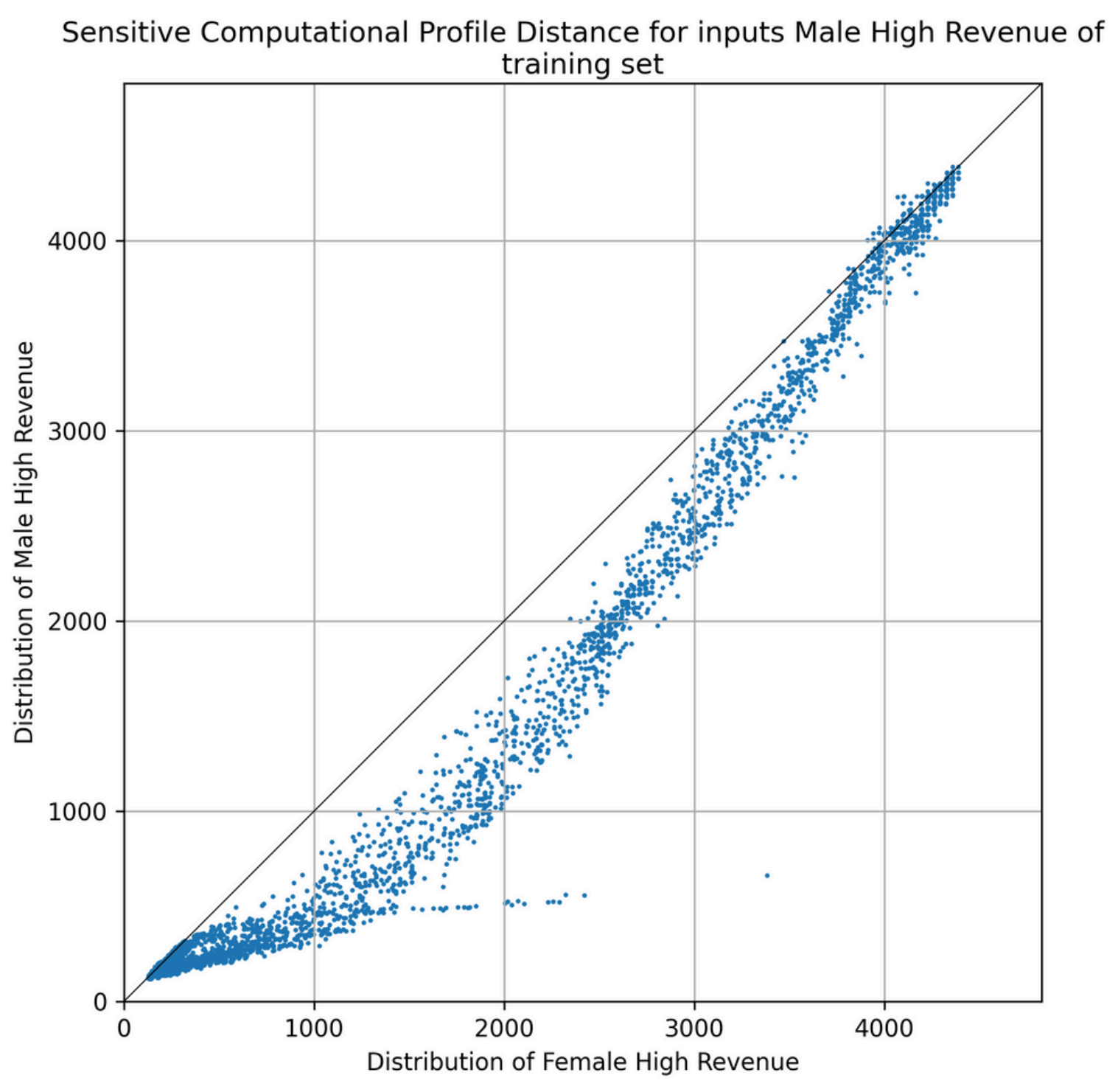

Figure 4
Male High Revenues inputs

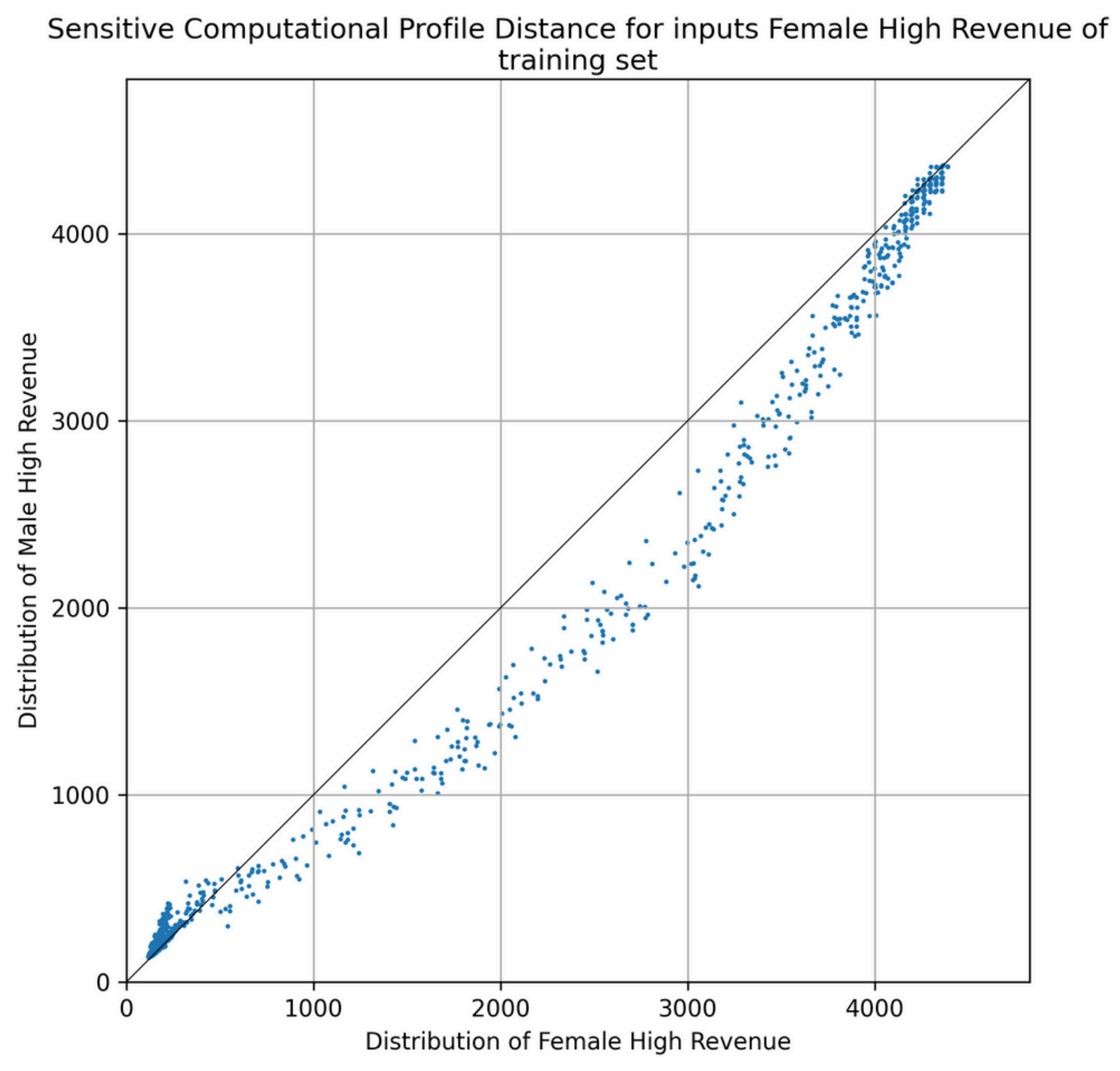

Figure 5
Female High Revenues inputs

Figure 4 and 5 use **Computational Profile Distance** (CPD), defined as the negative log of CPL. Higher CPD means lower CPL.

On **figure 4**, 93% of individuals has lower CPD towards Male High Revenues class than Female High Revenues class.
On **figure 5**, only 6% of all individuals has a lower CPD towards Male High Revenues class than Female High Revenues class.

The distributions of figure 4 and 5 are **skewed towards Male individuals**, especially for inputs that have CPD higher than 500.

CONCLUSION

Computational Profile Likelihood is adaptable to any fully connected deep network without requiring model re-training or an additional trained model.

We have used this method to exhibit variations between Male and Female individuals in the Adult Census database.

We also studied the impact of bias mitigation technique on CPL calculations and observed that differences between Male and Female profiles were diminished.

CPL offers reliability and robustness against unusual profiles [1] and brings promising results in algorithmic fairness. However, further studies are necessary.

Future research should aim to generalize these results on various model architectures and diverse databases. Studies on multi-dimensional attributes could also be considered.

REFERENCES

[1] Merlo, Ettore & Marhaba, Mira & Khomh, Foutse & Braiek, Houssem & Antoniol, Giuliano. (2021). Models of Computational Profiles to Study the Likelihood of DNN Metamorphic Test Cases.

[2] Aïvodji, Ulrich, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu, and Alain Tapp. 2021. "Local Data Debiasing for Fairness Based on Generative Adversarial Training" Algorithms 14, no. 3: 87

[3] Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

ACKNOWLEDGEMENTS

DEEL
DEpendable & Explainable Learning

The authors wish to thank the industrial and academic partners of the "DEpendable & Explainable Learning" (DEEL) project, CRIAQ, and the National Science and Engineering Research Council of Canada (NSERC) for funding the project CRDPJ 537462-18.