Siddhi Suresh, Harry L. Sanders IV, Joshua A. Riznyk
Faculty Advisor: Winnie Yu, EdD
{anantakriss1, sandersivh1, riznykj1, yuw1} @southernct.edu

THE EFFECTS OF BIASED TRAINING DATA ON INTERSECTIONAL CLASSIFICATION ACCURACY IN FACIAL ANALYSIS

INTRODUCTION

Most facets of our society, including scientific disciplines, have biases and other forms of discrimination. Computing and Artificial Intelligence (AI) are two areas where this issue is at the forefront. Bias and discrimination within this technology enable it to spread to other areas, facilitate social injustice and imbalance, and have societal repercussions for research or the larger society. Bias can be introduced into the development of AI, its learning methods, data collection, and/or analysis produced from these algorithms when researchers and developers fail to take into consideration their own biases, often unwillingly. Our primary goal in doing this research is to examine the relationship between bias in datasets and bias in classification algorithms.

DATASETS

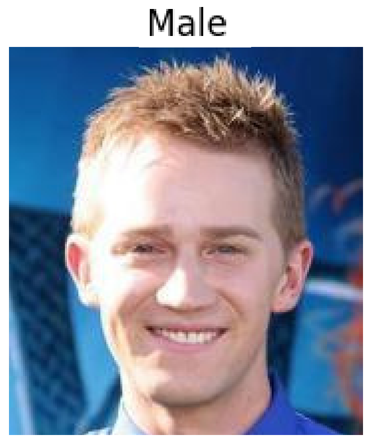
Table I: Data Partitioning

Dataset	Fairness	Size	Attributes
Fairface Balance	1.000	15704	Male, Female, Black, White, East Asian, Indian, Southeast Asian, Latino/Hispanic, Middle Eastern
Fairface Somewhat Balance	0.199	15663	Male, Female, Black, White, East Asian, Indian, Southeast Asian, Latino/Hispanic, Middle Eastern
Fairface Completely Unbalanced	0.042	15705	Male, Female, Black, White, East Asian, Indian, Southeast Asian, Latino/Hispanic, Middle Eastern
CelebA Balance	1.000	15682	Male, Not Male, Young, Old
CelebA Somewhat Balance	0.024	15680	Male, Not Male, Young, Old
CelebA Completely Unbalanced	0.001	12626	Male, Not Male, Young, Old

Images I–III: Example Data (CelebA)

Not Male

Male



Not Male

METHODOLOGY

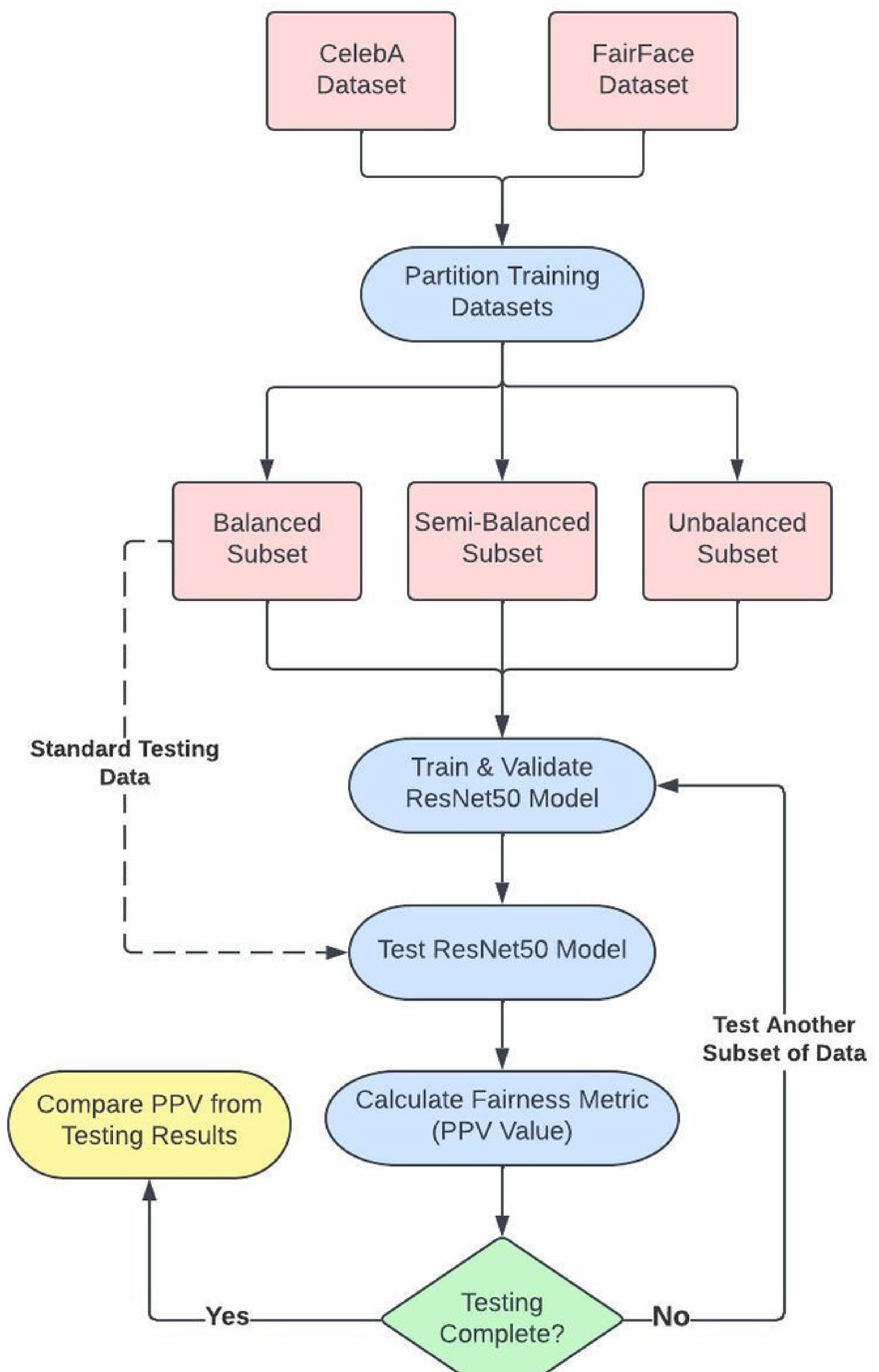


Image IV: Methodology Block Diagram

$$\mathrm{PPV} = \frac{TP}{TP+FP}$$

$$\mathrm{Fairness} = \frac{min(PPV)}{max(PPV)}$$

PPV=Positive Predictive Value
TP = True Postive
FP = False Positive

RESULTS

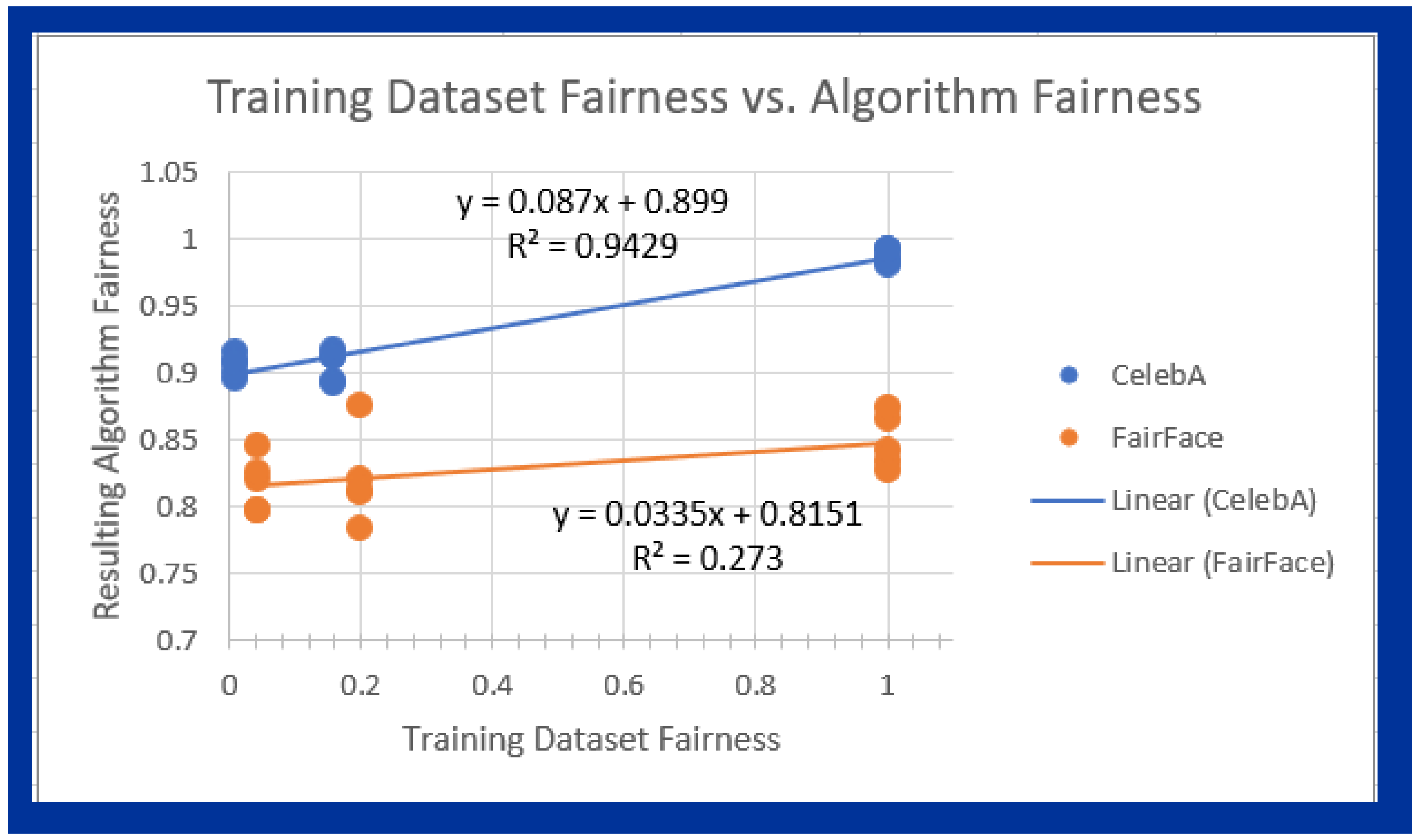


Image V: Fairness Results for FairFace & CelebA

Regression Statistics for CelebA		*Regression Statistics for FairFace*	
Multiple R	0.971051	**Multiple R**	0.522502
R Square	0.94294	**R Square**	0.273009
Adjusted R Square	0.938551	**Adjusted R Square**	0.217086
Standard Error	0.010014	**Standard Error**	0.024623
Observations	15	**Observations**	15
Coefficients	0.087045	**Coefficients**	0.033494
Standard Error	0.005939	**Standard Error**	0.015159
t Stat	14.6571	**t Stat**	2.209506
P-value	1.84E-09	**P-value**	0.045692

Table II: Statistical Results

CONCLUSIONS

- Image V shows the positive correlation calculated between the fairness of the FairFace & CelebA datasets and the fairness of the algorithm's performance
- Table II shows that the variance in the dataset fairness had a 92.2% and 27.3% impact on the fairness of the classifier for CelebA and FairFace respectively
- Correlations for both datasets are significant, as determined by the p-values calculated for both datasets; 0.045 for CelebA, >0.000 for FairFace
- For both datasets, we can reject the null hypothesis, which stated that there was no correlation between dataset fairness and algorithm fairness
- **We can successfully support the claim that dataset fairness does impact algorithm fairness**

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Conference on fairness, accountability and transparency. PMLR, 2018, pp. 77–91.
[2] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in Conference on Fairness, Accountability and Transparency. PMLR, 2018, pp. 107–118.
[3] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 12, pp. 2483–2509, 2014.

ACKNOWLEDGEMENT

We would like to acknowledge and thank Dr. Mohammad Islam, Ph.D & Dr. Amal El-Raouf, Ph.D, for their support, guidance, and inspiration throughout this project.