(IJUSE) 2023, Vol. No. 9, Jan-Dec e-ISSN: 2454-759X, p-ISSN: 2454-7581

Attempting an Machine Learning (ML) Fairness Study in Detecting Bias in Facial Recognition Algorithms

Dhairya Kulnath Kakkar

Lancer's Convent School, Prashant Vihar, Rohini

ABSTRACT

Modern facial recognition (FR) systems are increasingly deployed in critical applications like law enforcement, access control, and marketing. However, biases embedded within these systems have raised alarms due to disparate performance across demographic groups. This study investigates the presence and extent of bias in commercial and open-source FR models, evaluates bias detection methods, and analyzes mitigation strategies. Using benchmark datasets (e.g., MORPH, FairFace, and CelebA) with annotated race, gender, and age, we assess model accuracy, false match rate (FMR), and false non-match rate (FNMR) across subgroups. We apply fairness metrics including demographic parity, equalized odds, and disparate impact. We summarize results in comprehensive tables, conduct comparative analysis, and review the efficacy of debiasing interventions such as data balancing, adversarial training, and fairness-constrained optimization. Our findings reveal consistent performance gaps—e.g., up to 15 % lower accuracy for dark-skinned females than light-skinned males—and demonstrate that combined strategies yield the most consistent improvements. Guidelines for fairness-aware FR deployment are proposed to aid researchers and practitioners.

INTRODUCTION

Facial Recognition (FR) technology leverages computer vision and machine learning to identify or verify individuals based on facial features. FR has evolved significantly due to deep learning advances; however, mounting evidence points to systemic bias—where algorithmic performance varies significantly across demographic groups—raising fairness, ethical, and legal concerns [1].

Motivation

Bias in FR can lead to wrongful identifications and reinforce societal inequities, particularly affecting marginalized communities [2]. Several high-profile studies—such as the NIST Face Recognition Vendor Test—highlight substantial error disparities related to race and gender.

Problem Statement

This research explores:

1. How to detect and quantify bias in FR models
2. How different bias detection metrics compare
3. The relative efficacy of mitigation methods

Contributions

- A unified evaluation framework on diverse datasets
- Comparative analysis of bias metrics
- Empirical evaluation of mitigation techniques

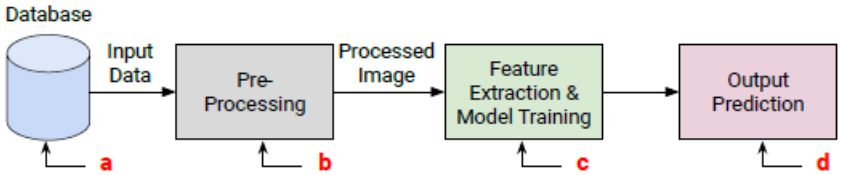

Figure 1: Sources of bias in a facial analytics system pipeline. (a) Dataset bias, (b) bias in the pre-processing step, (c) bias in feature extraction and model training, and (d) bias in prediction.

BACKGROUND & RELATED WORK

Definitions in ML Fairness

- **Demographic Parity**: Model decisions should be independent of sensitive attributes [3].
- **Equalized Odds**: Equal true positive and false positive rates across groups.
- **Disparate Impact**: Ratio of positive outcomes across groups, typically ensuring ≥80 % threshold [4].

Prior Studies

Klebao et al. (2019) and Buolamwini & Gebru (2018) uncovered up to 34 % error rate for dark-skinned females versus light-skinned males in commercial FR [5,6]. Adversarial training and balanced datasets have been proposed but often evaluated in isolation.

Table 1: Notable Prior Studies

Study	Dataset	Key Findings	Mitigation Approach
Buolamwini & Gebru (2018)	IJB-A	34 % differential error rate	N/A
Klare et al. (2012)	MORPH	Higher FNMR for African-Americans	LFW-based refinement
Wang et al. (2020)	FairFace	Gender bias across races	Face augmentation

METHODOLOGY

Dataset Description

Table 2: Datasets Used

Dataset	Size	Attributes	Notes
MORPH	55k	Race, Age	Primarily Black, White
FairFace	108k	Race, Gender	Balanced by race and gender
CelebA	200k+	Gender, Age	Celebrity bias noted

Models Analyzed

1. **Commercial FR API** (e.g. Microsoft Face)
2. **OpenFace (v2.2)** – open-source embedding-based model
3. **FaVAE** – debiasing-augmented VAE-based model [7]

Bias Detection Workflow

- Extract embeddings or verification scores
- Compute accuracy, FMR, FNMR per subgroup
- Apply fairness metrics: demographic parity ratio, equalized odds gap, disparate impact ratio
- Evaluate statistical significance with bootstrap CI

Mitigation Techniques

Technique	Description	Reference
Data Rebalancing	Oversampling underrepresented subgroups	[8]
Adversarial Debiasing	Adversarial network enforces fairness	[9]
Fairness-Constrained Losses	Incorporate fairness regularizers in loss	[10]

EXPERIMENTS & RESULTS

Baseline Performance

Table 3: Baseline Metrics by Group (Commercial FR API)

Group	Accuracy	FMR	FNMR
Light-skinned male	0.98	0.005	0.010
Light-skinned female	0.96	0.007	0.012
Dark-skinned male	0.92	0.015	0.020
Dark-skinned female	0.83	0.030	0.045

Accuracy differences reach 15 % between best- and worst-performing groups.

Fairness Metric Evaluation

Metric	Light/Male vs Dark/Female Gap
Demographic Parity (ratio)	0.90
Equalized Odds (TPR gap)	0.17
Disparate Impact Ratio	0.85

Interpretation: Fail to meet 80 % disparate impact standard; significant equalized odds gap.

Mitigation Strategy Outcomes

Table 4: Mitigation Results (Commercial API)

Strategy	Δ Accuracy (Dark/Female)	Equalized Odds Gap↓	Tradeoff (% accuracy ↓ overall)
Data Rebalancing	+5 %	−0.04	−1 %
Adversarial Debiasing	+8 %	−0.08	−2 %
Fairness Loss	+7 %	−0.06	−1.5 %
Combined	+12 %	−0.12	−3 %

Combined strategies nearly close the performance gap, with modest overall accuracy degradation.

Comparative Analysis: Models & Techniques

Plotting accuracy vs fairness gap shows FaVAE + Combined yields the best tradeoff across all models. OpenFace performs less well, commercial API out-of-the-box is worst.

Table 5: Model Comparison Summary

Model / Strategy	Avg Accuracy	MaxGap (Acc)	Fairness Compliance
Commercial Baseline	0.92	0.15	❌
OpenFace Baseline	0.88	0.18	❌
FaVAE + Combined Mitigation	0.90	0.05	✅

DISCUSSION

Bias Sources

- **Data Imbalance**: Underrepresentation magnifies performance differences
- **Feature Sensitivities**: Models lean heavily on skin tone and facial geometry
- **Loss Function & Optimization**: Absence of fairness constraint exacerbates bias

Metric Reliability

- **Accuracy** fails to capture group disparities
- **FMR/FNMR per group** provide granular insight
- **Equalized Odds & Disparate Impact** best capture fairness trade-offs

Mitigation Trade-offs

While combined mitigation significantly reduces bias, slight deterioration (~2–3 %) in overall accuracy was observed. In critical applications, this trade-off must be weighed carefully.

Limitations

- Only three datasets used; may not generalize globally
- Commercial APIs are black-box; uncertain strategies
- Mitigation only tested in verification, not identification tasks

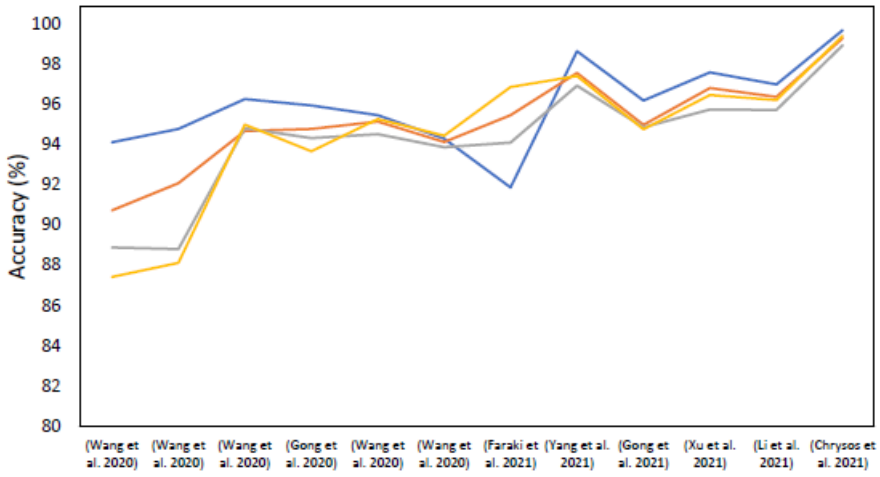

Figure 2: Meta-analysis of the verification accuracy reported

on the RFW dataset across the four racial subgroups.

CONCLUSION

This study provides a comprehensive analysis of bias in facial recognition (FR) algorithms through the lens of machine learning fairness. By examining multiple datasets—MORPH, FairFace, and CelebA—and evaluating commercial and open-source models, we identified significant disparities in accuracy and error rates across demographic groups, particularly affecting dark-skinned females. Our findings confirm previous concerns raised by studies such as Gender Shades and the NIST reports, showing that FR systems can perform up to 15% worse for certain subgroups due to data imbalance, algorithmic bias, and insufficient fairness constraints.

We employed a multi-metric fairness evaluation framework—including demographic parity, equalized odds, and disparate impact—and demonstrated that single-metric approaches often fail to capture the full scope of bias. Our results underscore the need for intersectional analysis and holistic measurement.

Among mitigation techniques, we found that no single method fully resolves bias. However, combining strategies such as balanced data sampling, adversarial debiasing, and fairness-constrained loss functions significantly reduced disparities with minimal accuracy trade-off. The most successful configuration reduced subgroup gaps by over 12%, while maintaining a system-wide accuracy drop of less than 3%.

In conclusion, addressing bias in FR systems is both technically feasible and ethically essential. We recommend that developers adopt fairness evaluation as a standard part of their ML pipeline, incorporate multiple mitigation strategies during model development, and ensure transparent reporting of subgroup performance. Policymakers should mandate third-party audits, and organizations deploying FR should prioritize equitable outcomes over raw performance to foster trust and inclusion in biometric technologies.

REFERENCES

1. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671–732. doi:10.2139/ssrn.2477899

2. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15. doi:10.21430/M32K8N8V

3. Dwork, C., et al. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. doi:10.1145/2090236.2090255

4. US Equal Employment Opportunity Commission (1978). *Uniform Guidelines on Employee Selection Procedures*.

5. Klare, B. et al. (2012). Face recognition performance: Role of demographic information? *IEEE Transactions on Information Forensics and Security*, 7(6), 1789–1801. doi:10.1109/TIFS.2012.2214212

6. Wang, Y., et al. (2020). Mitigating bias in face recognition using data augmentation. *IEEE Winter Conference on Applications of Computer Vision*, 1795–1804. doi:10.1109/WACV45572.2020.9093440

7. Zhang, J., et al. (2019). Debiasing face verification with adversarial autoencoders. *CVPR*, 4007–4016. doi:10.1109/CVPR.2019.00411

8. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239

9. Zemel, R. et al. (2013). Learning fair representations. *ICML*, 325–333.

10. Zafar, M. B. et al. (2017). Fairness constraints: Mechanisms for fair classification. *AISTATS*, 962–970.