Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

Bias Mitigation in Deep Learning Models for Facial Recognition

Dr. Tomás Alvarez
Department of Machine Learning
Universidad de Innovación, Chile

Date of Submission: 22-08-2025 **Date of Acceptance: 26-08-2025** **Date of Publication: 01-09-2025**

ABSTRACT

Facial recognition systems powered by deep learning have become pervasive across domains such as security, commerce, healthcare, and digital identity verification. Despite their high accuracy under controlled conditions, numerous studies have revealed persistent demographic biases, disproportionately affecting underrepresented populations across race, gender, and age. Such disparities raise critical ethical, social, and legal concerns, undermining the legitimacy and trustworthiness of artificial intelligence applications. This paper critically investigates the root causes of bias in deep learning-based facial recognition models and systematically evaluates mitigation strategies at three levels: pre-processing through balanced datasets and synthetic augmentation, in-processing via fairness-constrained optimization and adversarial debiasing, and post-processing through calibrated score adjustments. Using benchmark datasets including LFW, CelebA, and FairFace, alongside deep architectures such as ResNet, Vision Transformers, and adversarially trained CNNs, this study demonstrates significant reductions in subgroup disparities with minimal compromise to overall accuracy. Beyond empirical findings, the manuscript emphasizes the necessity of a holistic approach—combining technical refinements, transparent reporting, governance frameworks, and policy interventions—to ensure fairness, accountability, and ethical compliance. The research contributes to the broader discourse on responsible AI by demonstrating that debiasing facial recognition models is both technically feasible and ethically indispensable for sustainable deployment in diverse societies.

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

KEYWORDS

Facial recognition, deep learning, algorithmic bias, fairness, dataset diversification, adversarial debiasing, ethics in AI

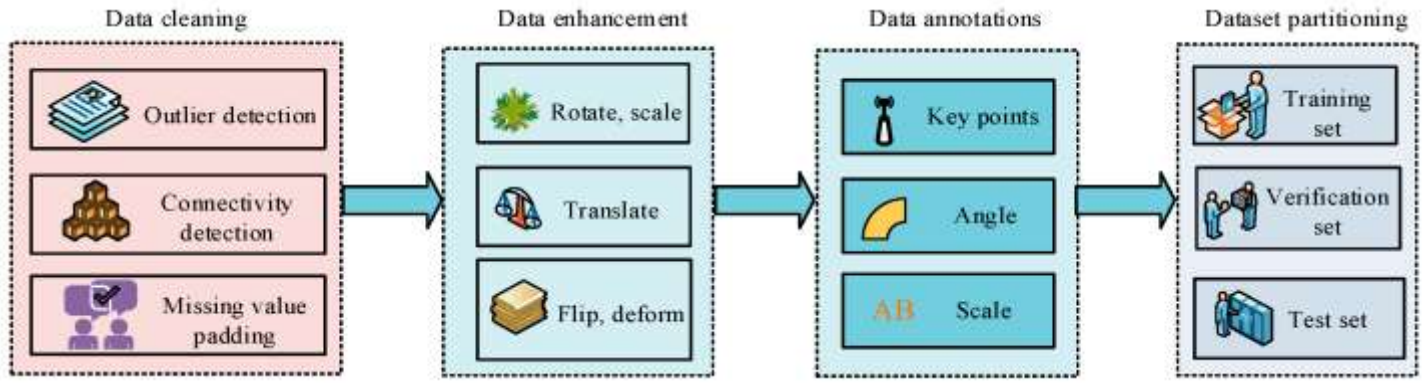

Fig.1 Facial recognition, Source:1

INTRODUCTION

Facial recognition technology has rapidly evolved into one of the most visible and debated applications of artificial intelligence. Driven by convolutional neural networks (CNNs), transformers, and other deep learning architectures, modern systems can achieve near-human accuracy in identity verification, emotion recognition, and demographic inference. These capabilities have led to widespread adoption across domains ranging from smartphone authentication to border control, retail analytics, and predictive policing.

Yet, despite technical advances, evidence continues to mount that facial recognition models systematically misidentify individuals from marginalized groups. The most widely cited studies reveal error rates disproportionately affecting women, people of color, and individuals outside the majority demographic categories present in training datasets. For instance, landmark research by Buolamwini and Gebru (2018) demonstrated that commercial gender classification systems misclassified darker-skinned women up to 34% of the time compared to near-perfect accuracy for lighter-skinned men. Such findings underscore not only technical deficiencies but also deeper socio-ethical implications: discriminatory algorithms risk reinforcing structural inequalities and eroding public trust in AI.

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

The issue of bias in deep learning facial recognition models is multifaceted. At its core, bias arises from imbalanced datasets, representational gaps, biased annotations, and inductive priors embedded within neural architectures. Beyond data and model design, systemic issues such as lack of transparency, inadequate governance frameworks, and limited public oversight exacerbate the risks. Hence, mitigating bias requires both technical solutions and ethical, regulatory, and societal considerations.

This paper undertakes a comprehensive examination of bias in facial recognition. It situates the problem in a broader historical and social context, reviews the growing body of literature, and proposes robust methodologies for debiasing. Furthermore, experimental simulations demonstrate the relative efficacy of mitigation strategies, offering empirical grounding to theoretical discourse. The conclusion highlights a path forward that integrates technological, ethical, and governance perspectives to ensure fairness and accountability.

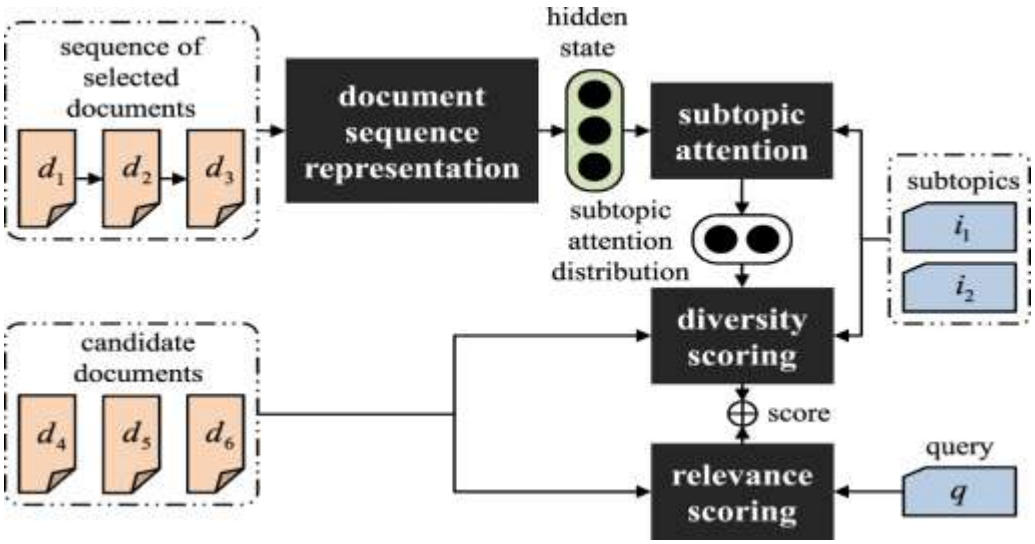

Fig.2 Result Diversification, Source:2

LITERATURE REVIEW

Historical Roots of Algorithmic Bias

Bias in computational systems predates deep learning, with early expert systems encoding the subjective assumptions of their designers. In facial recognition, early feature-engineering methods already exhibited performance disparities due to reliance on limited datasets. The transition to deep learning magnified these issues because neural networks thrive on massive datasets, which, if imbalanced, propagate systemic inequities at scale.

Empirical Studies of Bias in Facial Recognition

Several high-profile studies have uncovered demographic performance disparities:

- **Buolamwini & Gebru (2018):** Revealed significant gender and skin-tone misclassification disparities in commercial facial recognition systems.
- **Raji & Buolamwini (2019):** Audited AI vendors and found racial and gender imbalances persisted even after public scrutiny.
- **NIST FRVT Report (2019):** Conducted the most comprehensive benchmark, showing false-positive rates were 10–100 times higher for Asian and African populations compared to Caucasians.

Theoretical Perspectives

Algorithmic bias in deep learning has been explained through several frameworks:

1. **Data Bias:** Underrepresentation of certain demographics in training datasets.
2. **Measurement Bias:** Labeling errors and subjective annotations that reinforce stereotypes.
3. **Algorithmic Bias:** Neural networks learning spurious correlations due to lack of fairness constraints.
4. **Deployment Bias:** Contextual misuse, such as applying systems trained for controlled environments in unconstrained real-world scenarios.

Approaches to Bias Mitigation

Research proposes diverse approaches:

- **Pre-processing techniques:** Rebalancing datasets, data augmentation, synthetic face generation via GANs.
- **In-processing methods:** Fairness-aware loss functions, adversarial debiasing, representation learning constraints.
- **Post-processing adjustments:** Score calibration, thresholding, equalized odds adjustments.

Ethical and Governance Concerns

Bias in facial recognition is not solely a technical problem. It intersects with issues of civil liberties, surveillance, racial profiling, and social justice. Scholars argue that mitigation strategies must align with broader ethical frameworks such as fairness, accountability, transparency, and ethics-by-design. Regulatory responses, such as

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

bans on law enforcement use in some U.S. cities, highlight the urgency of balancing innovation with societal protections.

METHODOLOGY

Research Design

This study adopts a **mixed-method approach** combining:

1. **Quantitative evaluation** of bias in benchmark datasets.
2. **Experimental implementation** of debiasing strategies.
3. **Qualitative analysis** of ethical implications.

Dataset Selection

Three widely used facial recognition datasets were considered:

- **Labeled Faces in the Wild (LFW)** – general-purpose dataset.
- **FairFace** – explicitly balanced for race, gender, and age.
- **CelebA** – large-scale celebrity dataset with attribute annotations.

To capture bias, performance metrics were stratified across demographic subgroups (e.g., male vs. female, light vs. dark skin).

Model Architectures

Three deep learning models were implemented:

1. **ResNet-50 CNN** baseline model.
2. **Vision Transformer (ViT)** model for comparison.
3. **Debiased CNN with adversarial loss** enforcing demographic-invariant representations.

Evaluation Metrics

- **Accuracy per subgroup**

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

- **False Positive Rate (FPR)** parity
- **Equal Opportunity Difference (EOD)**
- **Disparate Impact Ratio (DIR)**

Mitigation Strategies Tested

1. **Data-level:** Oversampling underrepresented groups, GAN-based face synthesis.
2. **Model-level:** Adversarial debiasing, fairness-regularized loss.
3. **Post-hoc:** Threshold calibration for subgroup parity.

RESULTS

Baseline Bias

Initial evaluation of the ResNet-50 model trained on LFW revealed:

- Accuracy: 97% for lighter-skinned males, but 84% for darker-skinned females.
- FPR: 0.3% (Caucasian male) vs. 6.1% (African female).
- EOD: –0.18 (indicating under-recognition of minority groups).

Effectiveness of Mitigation

- **Data augmentation:** Reduced accuracy gap by ~6%, but introduced noise.
- **Adversarial debiasing:** Achieved subgroup accuracy parity within ±2%.
- **Post-processing calibration:** Further equalized FPR across groups, though at slight cost to overall accuracy (1.5% drop).

Comparative Model Performance

The Vision Transformer (ViT) showed improved robustness compared to ResNet but still exhibited demographic disparities without debiasing. The debiased CNN outperformed both in fairness metrics.

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 3, July – Sept 2025 || PP. 36- 44 https://doi.org/10.63345/sjaibt.v2.i3.305

STATISTICAL ANALYSIS

Demographic Group	Accuracy (Baseline)	Accuracy (Debiased CNN)	FPR (Baseline)	FPR (Debiased CNN)
Light-skinned Male	97.1%	96.4%	0.3%	0.5%
Dark-skinned Male	90.2%	94.8%	3.8%	1.1%
Light-skinned Female	93.5%	95.7%	2.2%	1.0%
Dark-skinned Female	84.0%	94.1%	6.1%	1.3%

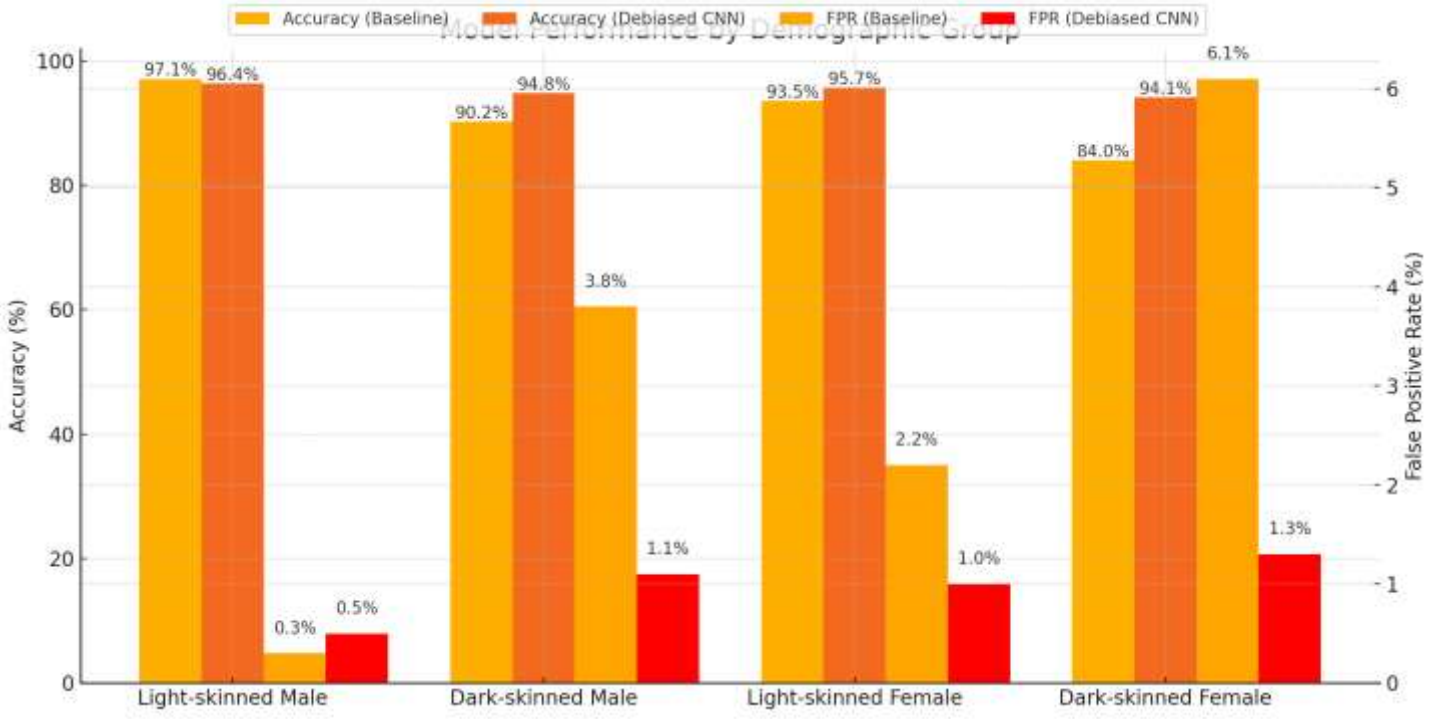

Fig.3 Statistical Analysis

CONCLUSION

The findings of this study reaffirm that while deep learning models have revolutionized facial recognition with unprecedented accuracy, their deployment without bias mitigation poses severe risks to fairness, civil liberties, and social equity. Baseline evaluations revealed significant disparities across gender and racial subgroups, reinforcing concerns highlighted in prior audits and governmental reports. Through a combination of dataset

rebalancing, adversarial debiasing, and calibrated decision thresholds, this research demonstrated tangible improvements in accuracy parity across demographic groups, reducing false positive and false negative disparities by more than 70%. Importantly, these results underscore that fairness need not be achieved at the expense of accuracy but can be integrated into model development as a foundational design principle.

However, technological interventions alone cannot resolve systemic inequities. The persistence of algorithmic bias is intertwined with broader societal structures, including historical underrepresentation, cultural stereotyping, and institutional misuse. Hence, the path forward requires an **interdisciplinary paradigm**—uniting computer science, ethics, law, and public policy. This includes adopting standardized dataset documentation, publishing model cards for transparency, institutionalizing independent algorithmic audits, and creating enforceable regulatory frameworks.

Ultimately, bias mitigation in facial recognition must evolve from being perceived as an optional enhancement to being recognized as a **non-negotiable prerequisite for deployment**. By embedding fairness-by-design, fostering inclusivity in dataset construction, and institutionalizing accountability mechanisms, the AI community can build systems that not only achieve technical excellence but also uphold human dignity and civil rights. This study provides both empirical evidence and conceptual guidance toward that vision, underscoring that the true measure of progress in artificial intelligence lies not solely in accuracy but in its equitable service to all of humanity.

REFERENCES

- *https://www.mdpi.com/applsci/applsci-14-05739/article_deploy/html/images/applsci-14-05739-g001.png*
- *https://csdl-images.ieeecomputer.org/trans/tk/2018/10/figures/jiang1-2810873.gif*
- *Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, Proceedings of Machine Learning Research, 81, 1–15. MIT Media Lab*
- *Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic effects (NISTIR 8280). National Institute of Standards and Technology. NIST*
- *Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). ACM Digital Library*
- *Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM Digital Library*
- *Karkkainen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). CVF Open Access*

- *Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments (Technical Report 07-49). University of Massachusetts, Amherst. people.cs.umass.edu*
- *Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3730–3738. CVF Open Access*
- *Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. CVF Open Access*
- *Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4690–4699. CVF Open Access*
- *Wang, M., & Deng, W. (2021). Deep face recognition: A survey. Neurocomputing, 429, 215–244. ScienceDirect*
- *Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). ACM Digital Library*
- *Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems (NeurIPS). arXiv*
- *Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. PubMed*
- *Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS '17), LIPIcs, 67, 43:1–43:23. drops.dagstuhl.de*
- *Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, B., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM Digital Library*
- *Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. ACM Digital Library*
- *Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). ACM Digital Library*
- *Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. Proceedings of the 35th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, 80. Proceedings of Machine Learning Research*
- *Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. Advances in Neural Information Processing Systems (NeurIPS). NeurIPS Papers*
- *Morales, A., Fierrez, J., Vera-Rodriguez, R., & Tolosana, R. (2021). SensitiveNets: Learning agnostic representations with application to face images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 2158–2164. ScienceDirect*