Auditing and Mitigating Bias in Gender Classification Algorithms: A Data-Centric Approach

Tadesse K. Bahiru
Department of Computer Science
University of Houston
Houston, Texas, USA
tbahiru@uh.edu

Natnael Tilahun Sinshaw
Department of Computer Science
University of Houston
Houston, Texas, USA
ntilahun@uh.edu

Teshager Hailemariam Moges
Department of Computer Science
KIoT, Wollo University
Dessie, Ethiopia
hailemariam.moges@wu.edu.et

Dheeraj Kumar Singh
Department of Information Technology
ADIT, Charutar Vidya Mandal University
Vallabh Vidyanagar, Anand, Gujarat, India
it.dksingh@adit.ac.in

***Abstract*—Gender classification systems often inherit and amplify demographic imbalances in their training data. We first audit five widely used gender classification datasets, revealing that all suffer from significant intersectional underrepresentation. To measure the downstream impact of these flaws, we train identical MobileNetV2 classifiers on the two most balanced of these datasets, UTKFace and FairFace. Our fairness evaluation shows that even these models exhibit significant bias, misclassifying female faces at a higher rate than male faces and amplifying existing racial skew. To counter these data-induced biases, we construct BalancedFace, a new public dataset created by blending images from FairFace and UTKFace, supplemented with images from other collections to fill missing demographic gaps. It is engineered to equalize subgroup shares across 189 intersections of age, race, and gender using only real, unedited images. When a standard classifier is trained on BalancedFace, it reduces the maximum True Positive Rate gap across racial subgroups by over 50% and brings the average Disparate Impact score 63% closer to the ideal of 1.0 compared to the next-best dataset, all with a minimal loss of overall accuracy. These results underline the profound value of data-centric interventions and provide an openly available resource for fair gender classification research.**

***Index Terms*—algorithmic fairness, intersectional bias, gender classification, face datasets, data-centric AI**

I. Introduction

Automated face analysis has moved from research laboratories to daily life, shaping applications as diverse as border control, photo tagging, driver monitoring, and targeted advertising [1]. Within these systems, gender classification frequently operates as a first-stage filter that influences how subsequent decisions are made. When these classifiers work well for one population but fail for another, the resulting errors accumulate and can lead to unequal treatment. Such disparities are not only technical shortcomings, but also social risks, since they affect access to services and trust in public institutions. Early audits of commercial face analysis systems showed that women with darker skin tones experienced error rates up to ten times higher than men with lighter skin tones [2]. Later investigations confirmed that these discrepancies persist across multiple vendors and remain evident even as architectures have advanced [3], [4].

Much of this persistent disparity can be traced back to the data that fuel these models. Large-scale face datasets, often scraped from online sources, overrepresent young men with lighter skin while severely underrepresenting women, older adults, and minority racial groups [5]–[7]. These imbalances effectively encode historical inequities in the training process before the model learns its first weight. Once such skew is present, even well-intentioned pre-processing or architectural adjustments are rarely sufficient to fully undo its influence. Approaches that focus on model-side interventions, such as adversarial debiasing or fairness-constrained optimization, can reduce certain disparities, but typically involve trade-offs. They may sacrifice accuracy for specific subgroups, require fine-grained demographic labels that are not always available, or demand complex optimization schemes that limit their use in practice [8]–[11].

These challenges highlight the need for a complementary data-centric perspective that elevates representativeness and balance to the central goals of system design. Instead of relying solely on downstream fixes, this approach recognizes that equitable outcomes first depend on equitable inputsf [12]. By carefully auditing, curating, and balancing training datasets, it is possible to directly address the structural gaps that drive algorithmic bias and thereby create a stronger foundation for fair model behavior.

In this paper, we adopt this data-centric perspective and make three primary contributions. First, we introduce a transparent audit protocol that quantifies intersectional coverage across age, race, and gender for widely used face datasets, providing a reproducible baseline for future evaluations. Second, we show how these imbalances manifest in practice by training gender classification models and measuring disparities in true positive rates and positive outcome ratios, thus

isolating data bias as a principal driver of unfair outcomes. Finally, we propose a fairness-aware training framework that combines adversarial learning, equalised odds regularisation, and re-weighting to mitigate disparities while maintaining competitive accuracy. Through comprehensive experiments, we demonstrate that this framework significantly narrows gaps in equalised odds and disparate impact across demographic intersections, underscoring the value of data-centric interventions in building more inclusive gender classification systems.

II. Related Work

Research on algorithmic fairness has matured rapidly, moving from broad conceptual discussions to rigorous empirical studies that quantify and mitigate bias in machine learning pipelines [13]–[15]. Early work established that bias can emerge from several sources, including skewed sampling, imperfect labels, and imbalance in the feature space, each requiring distinct diagnostic tools and remedies. Hinnefeld *et al.* [16] provided one of the first systematic case studies of bias detection in observational data, benchmarking six common fairness metrics and synthesizing guidelines for metric selection based on the type of bias encountered. Their analysis underscored the importance of choosing evaluation criteria that align with the causal mechanisms underlying the data.

Subsequent surveys broadened the landscape. Le *et al.* [17] focused on tabular data and used Bayesian networks to uncover conditional dependencies that reveal bias propagation paths. Their results highlighted that detrimental bias can originate from the training data or the prediction function, motivating interventions at multiple stages of model development. Jiang *et al.* [18] proposed minimizing the Wasserstein distance between class-conditional feature distributions as a route to fair classification, demonstrating improved group parity without large sacrifices in accuracy. In parallel, Dwork *et al.* [19] introduced a decoupled classification framework in which separate models are trained for each protected subgroup, with transfer learning alleviating data scarcity for minorities. This line of work marked an early shift toward more fine-grained, group-aware learning strategies.

Data-centric mitigation remains pivotal for vision tasks because visual corpora often amplify societal imbalances. Wang *et al.* [8] curated Racial Faces in the Wild to reduce racial skew and demonstrated that balanced sampling can shrink accuracy gaps across skin tones. Krishnan *et al.* [20] examined gender classification specifically, showing that even architectures trained on balanced data sets can display disparate performance due to complex interactions between features and protected attributes. Their findings suggest that fairness assessments must consider intersectional slices such as gender–race combinations rather than isolated attributes.

Despite these advances, open problems remain. Comparative audits across face data sets are rare, making it difficult to isolate which design choices most effectively close demographic gaps. Moreover, existing mitigation methods sometimes require synthetic augmentation or costly post-processing that may introduce artifacts or degrade utility. Our work builds on this foundation by offering a unified audit protocol, quantifying how intersectional imbalances translate to measurable disparities, and validating a training pipeline that integrates adversarial debiasing with equalised-odds regularisation. This approach situates fairness as a first-class optimization goal and provides actionable guidance for constructing inclusive gender classification systems.

III. Methodology

This study introduces a structured four-stage pipeline for auditing and improving fairness in gender classification systems. The process begins with a dataset audit that diagnoses representational deficiencies across key demographic axes, followed by a targeted repair stage that addresses both missing subgroups and distributional imbalances. A fairness-aware training stage then integrates adversarial learning and error regularization to suppress demographic leakage and mitigate performance disparities. The final stage conducts a comprehensive fairness evaluation using both outcome- and error-based criteria. As illustrated in Fig. 1, each stage is designed to produce an interpretable artifact, enabling transparent and traceable fairness interventions from raw data to final model behavior.

A. Dataset Audit

Let $x \in \mathcal{X}$ represent a face image, $y \in \{\text{male}, \text{female}\}$ its ground-truth gender label, and a its sensitive attributes. For this study, we consider three primary demographic axes: gender, race, and age. The complete list of the subgroups considered for our fairness quantification is detailed in Table I. We computed two complementary indicators to characterize the demographic structure of a dataset based on these groups.

TABLE I
Demographic Subgroups Used for Fairness Evaluation [21]

Demography	Subgroups
Gender	Male, Female, Other
Race	White/Caucasian, Black/African, Southeast Asian, East Asian, Indian, Hispanic/Latino, Middle Eastern, Mixed-race, Other
Age	0–2, 3–7, 8–15, 16–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–100

1) Inclusivity: The goal of the Inclusivity score is to capture whether all subgroups that are expected in the dataset are actually present. This is particularly important for attributes such as age, gender, and race, which often carry both ethical and statistical weight in fairness evaluations. For each attribute $d \in D = \{\text{age}, \text{gender}, \text{race}\}$, we define a coverage ratio $r_d = |s_d|/|S_d|$, where $|s_d|$ is the number of subgroups observed in the dataset and $|S_d|$ is the total number of subgroups that should be represented. The Inclusivity score is then given by

$$R = \frac{1}{|D|} \sum_{d \in D} \frac{|s_d|}{|S_d|}, \tag{1}$$

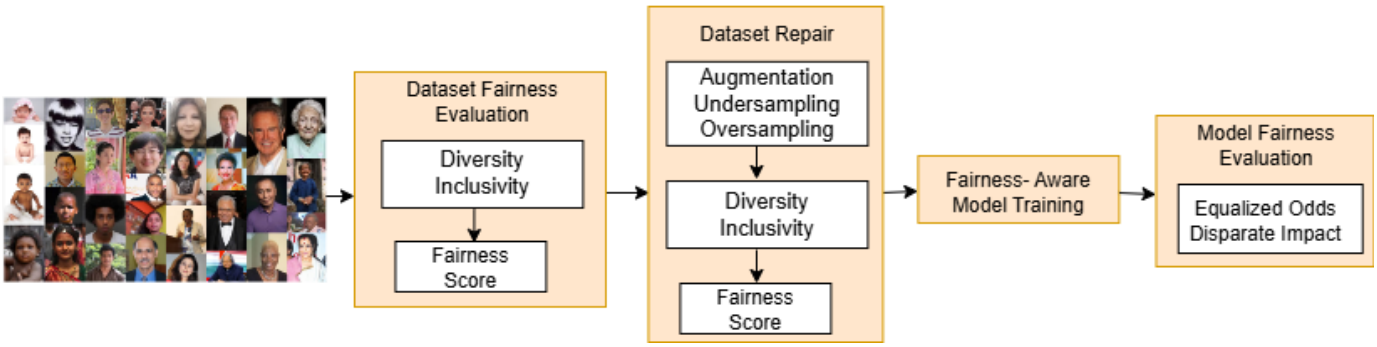

Fig. 1. Overview of the proposed method. Each stage produces outputs used by the next, from dataset diagnosis to final fairness evaluation.

with $R \in [0, 1]$. A value of 1.0 corresponds to perfect coverage, meaning every subgroup across all considered attributes is included. Lower values indicate missing subgroups and highlight possible blind spots in representation.

2) Diversity: While Inclusivity tells us whether subgroups exist in the dataset, Diversity examines how evenly they are represented. To do this, we consider intersectional groups defined across age, gender, and race. Let g_i denote each cell in the age–gender–race lattice, and define $\text{GRS}_i = |g_i|/N$, where $N = |\mathcal{X}|$ is the total sample size. The Diversity score is then calculated as

$$D = \frac{\min_i \text{GRS}_i}{\max_i \text{GRS}_i}, \tag{2}$$

where a value close to 1.0 indicates that all demographic cells are represented in a balanced manner. A value near zero, on the other hand, reveals large disparities between the most and least represented groups, pointing to uneven distribution that can bias downstream learning.

B. Fairness-Aware Training

To reduce bias in model predictions, we apply a fairness-aware training strategy based on adversarial learning. Our model builds on a MobileNetV2 backbone pretrained on ImageNet, where the convolutional layers are frozen to preserve general feature representations. On top of this backbone, we add two heads: a primary gender classifier and an adversary tasked with predicting joint sensitive attributes $\langle \text{age-bin}, \text{race} \rangle$. The adversary is connected through a Gradient Reversal Layer (GRL), which forces the shared representation to obscure demographic information and thereby limit leakage.

The training objective combines three components: the classification loss for accurate gender prediction, an adversarial loss that discourages demographic leakage, and a fairness regularizer that penalizes gaps in true positive rates between groups. The overall objective is expressed as

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(\theta) + \lambda_{\text{adv}} \max_{\phi} \mathcal{L}_{\text{adv}}(\theta, \phi) + \lambda_{\text{eo}} \left| \text{TPR}_{\max} - \text{TPR}_{\min} \right|, \tag{3}$$

where θ are the model parameters, ϕ are the adversary parameters, $\lambda_{\text{adv}} = 0.1$, and $\lambda_{\text{eo}} = 0.05$. This formulation guides the model to achieve accurate classification while simultaneously suppressing demographic leakage and reducing disparities in error rates. In this way, the training process explicitly balances predictive performance with fairness considerations.

C. Model Fairness Evaluation

We evaluate the trained models using two group fairness metrics: Equalized Odds and Disparate Impact. These metrics examine fairness both in terms of error distribution and outcome allocation across subgroups.

1) Equalized Odds: Equalized Odds requires that prediction outcomes be conditionally independent of sensitive attributes given the true label. For each class label $i \in \{\text{male}, \text{female}\}$ and demographic groups $j, k \in D$, Equalized Odds requires

$$\begin{aligned} P(\hat{Y} = i \mid Y = i, A = j) &= P(\hat{Y} = i \mid Y = i, A = k), \\ P(\hat{Y} \neq i \mid Y \neq i, A = j) &= P(\hat{Y} \neq i \mid Y \neq i, A = k). \end{aligned} \tag{4}$$

To simplify evaluation, we compute the maximum disparity in accuracy between subgroups as

$$\epsilon(\hat{Y}) = \max_{j,k \in D} \left| \log \frac{P(\hat{Y} = Y \mid A = j)}{P(\hat{Y} = Y \mid A = k)} \right|. \tag{5}$$

Lower values of $\epsilon(\hat{Y})$ reflect better fairness under Equalized Odds.

2) Disparate Impact: Disparate Impact is a widely used group fairness metric that examines whether positive predictions are distributed fairly between different demographic groups. It provides a way to quantify potential biases in decision-making systems by comparing the rate of favorable outcomes received by an unprivileged group relative to a privileged group. Formally, it is defined as

$$\text{DI} = \frac{P(\hat{Y} = 1 \mid A = \text{unprivileged})}{P(\hat{Y} = 1 \mid A = \text{privileged})}. \tag{6}$$

Here, $\hat{Y} = 1$ indicates a positive prediction, while A represents the sensitive attribute used to distinguish between groups. In practice, this probability is estimated by computing the proportion of predicted positives within each group. This

is typically approximated using both true positives and false positives over the total number of individuals in the group:

$$\text{Positive Outcome Ratio} = \frac{\text{TP} + \text{FP}}{\text{Total Cases}}. \tag{7}$$

A Disparate Impact value close to 1 suggests that both groups receive positive outcomes at nearly the same rate, which indicates equity in treatment. However, when the ratio falls below 0.8, the outcomes are considered potentially discriminatory according to the well-established "four-fifths rule" [22]. This threshold originates from employment law and is commonly adopted in algorithmic fairness evaluations to flag potential disparities in automated decision systems.

IV. Experimental Results

1) Experimental Setup: The experimental setup was designed to isolate the impact of data quality on model fairness. From our initial audit, we selected UTKFace and FairFace as our baseline training sets, as they demonstrated comparatively better inclusivity and diversity than other widely used datasets. We then trained three identical models to provide a clear comparison: one on UTKFace, one on FairFace, and one on our proposed BalancedFace. For a consistent and robust comparison, we employed a MobileNetV2 architecture pre-trained on ImageNet as our feature extractor. In each experiment, the convolutional base of the MobileNetV2 model was frozen, and a new classification head was attached. This head consisted of a global average pooling layer, a fully-connected dense layer with 256 units and ReLU activation, a dropout layer with a rate of 0.3, and a final sigmoid output unit for binary gender prediction. Each dataset was partitioned into a 70% training, 15% validation, and 15% testing split, ensuring that the intersectional distribution of the original set was maintained across the partitions. The models were trained for 30 epochs using the Adam optimizer with a learning rate of 1e-4, a batch size of 64, and a standard binary cross-entropy loss function. All experiments were conducted on a single NVIDIA RTX A6000 GPU with 49 GB of memory to ensure a consistent hardware environment. This setup allowed us to attribute any observed fairness or accuracy differences primarily to the dataset composition rather than architectural or optimization changes.

A. Dataset Inclusiveness and Diversity

Our audit of the five widely used face datasets presented in Fig 2 reveals significant demographic imbalances. The tables quantify this lack of representation across age and race, with our analysis showing that even the highest-scoring datasets, FairFace and UTKFace, still lacked certain age brackets (such as 0–2 and 70+) and racial categories (notably Southeast Asians). These missing groups are particularly concerning, as they create systematic blind spots in downstream models.

To address these deficiencies, we constructed the BalancedFace dataset. This was achieved by first creating a comprehensive image pool by combining images from FairFace and UTKFace with images from other public datasets to fill

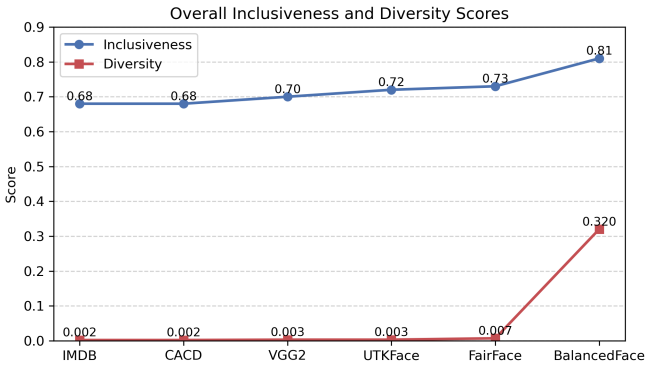

Fig. 2. Comparison of Inclusiveness and Diversity scores across six widely used face datasets. BalancedFace demonstrates markedly higher inclusiveness and especially improved diversity, highlighting its strength in addressing demographic underrepresentation compared to prior datasets.

specific demographic gaps, such as missing age brackets and racial categories. This curated set was then balanced using sampling and augmentation techniques to produce a more even distribution across all demographic subgroups. As a result, BalancedFace markedly improves both inclusivity and diversity scores, as reflected in Fig 2. Although we could not include the 'other' gender due to data limitations, coverage across race and age was substantially improved, particularly for the extreme age groups that were consistently absent in prior datasets.

B. Model Accuracy and Fairness Trade-off

Figure 3 reports the overall classification accuracy of gender classification models trained on each dataset. Models trained on UTKFace and FairFace achieved higher accuracy, largely because these datasets overrepresent certain subgroups, enabling the model to specialize in majority classes. However, this apparent performance comes at the cost of fairness.

In contrast, the model trained on BalancedFace achieved slightly lower overall accuracy, but this was a deliberate trade-off. By ensuring more balanced representation across intersectional groups, BalancedFace shifts the model away from majority-class overfitting and toward equitable performance across subgroups. In real-world fairness-critical applications, this balance is far more valuable than marginal gains in global accuracy.

C. Fairness Across Racial Groups

Fairness metrics in Table II reveal disparities between racial groups. The BalancedFace-trained model yields better Disparate Impact (DI) scores and more consistent True Positive Rates (TPR), particularly for Black and Southeast Asian females. These results demonstrate that improved dataset diversity can lead to more equitable model outcomes.

D. Fairness Across Age Groups

Table III compares fairness outcomes across age groups. Similar to racial fairness, UTKFace and FairFace struggle in

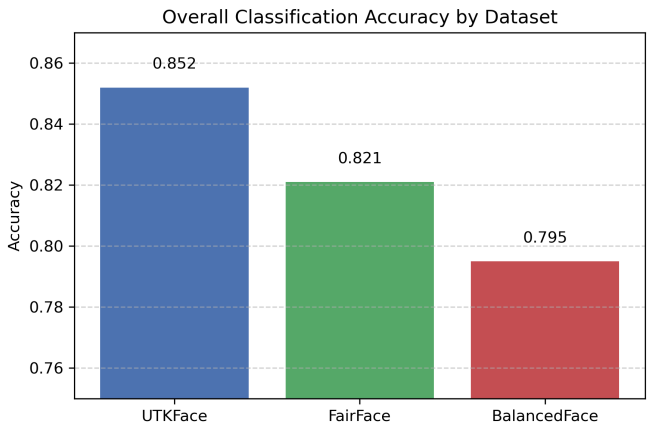

Fig. 3. Overall classification accuracy of gender classification models trained on UTKFace, FairFace, and BalancedFace datasets. While UTKFace and FairFace achieve higher raw accuracy, the BalancedFace model achieves more equitable performance across demographic groups with only a minor drop in overall accuracy.

TABLE II
RACIAL FAIRNESS METRICS ACROSS MODELS

Group	Metric	UTKFace	FairFace	BalancedFace
White (Ref)	TPR (F)	0.830	0.792	**0.785**
	TPR (M)	0.801	0.747	**0.760**
	DI	1.000	1.000	**1.000**
Black	TPR (F)	0.550	0.650	**0.721**
	TPR (M)	0.925	0.758	**0.735**
	DI	0.751	0.887	**0.952**
East Asian	TPR (F)	0.882	0.845	**0.810**
	TPR (M)	0.701	0.682	**0.741**
	DI	1.045	1.066	**1.020**
Indian	TPR (F)	0.760	0.755	**0.773**
	TPR (M)	0.890	0.761	**0.780**
	DI	0.965	0.953	**0.985**
Southeast Asian	TPR (F)	0.865	0.807	**0.790**
	TPR (M)	0.650	0.703	**0.715**
	DI	1.031	1.019	**1.008**

underrepresented age brackets, particularly at the extremes (0–2 years and 70+). BalancedFace fills these gaps by explicitly curating data for such groups. As a result, it enables reliable predictions for infants and elderly individuals, which were not possible with the baseline datasets.

For intermediate age groups, BalancedFace maintains performance while reducing disparities. For example, in the 60–69 age group, the TPR for females improves from 0.579 (FairFace) to 0.632, and the DI score increases from 0.691 to 0.772. These consistent improvements demonstrate that data balancing helps not only to include missing categories but also to equalize treatment across all age brackets. Importantly, BalancedFace expands fairness into previously excluded subgroups while avoiding large sacrifices in overall predictive power.

TABLE III
AGE FAIRNESS METRICS ACROSS MODELS

Age Group	Metric	UTKFace	FairFace	BalancedFace
20–29 (Ref)	TPR (F)	0.891	0.838	**0.819**
	TPR (M)	0.845	0.689	**0.730**
	DI	1.000	1.000	**1.000**
0–2	TPR (F)	–	–	**0.684**
	TPR (M)	–	–	**0.683**
	DI	–	–	**0.835**
3–9	TPR (F)	0.850	0.833	**0.802**
	TPR (M)	0.650	0.524	**0.548**
	DI	0.955	0.993	**0.979**
40–49	TPR (F)	0.795	0.724	**0.721**
	TPR (M)	0.910	0.835	**0.829**
	DI	0.882	0.864	**0.880**
60–69	TPR (F)	0.680	0.579	**0.632**
	TPR (M)	0.935	0.884	**0.855**
	DI	0.735	0.691	**0.772**
70+	TPR (F)	–	–	**0.724**
	TPR (M)	–	–	**0.800**
	DI	–	–	**0.884**

V. DISCUSSION

The findings from our study make it clear that the demographic makeup of a training dataset is a decisive factor in shaping fairness outcomes for gender classification models. The central lesson is that data-centric interventions, particularly those that deliberately balance representation across intersectional groups, can mitigate bias more effectively than many of the complex model-centric techniques proposed in recent years.

The comparison across UTKFace, FairFace, and BalancedFace highlights this point. Models trained on the first two datasets produced relatively high overall accuracy, yet this came at the cost of significant disparities across subgroups. For example, the True Positive Rate for Black females was substantially lower than that for majority groups, illustrating how overrepresentation of some categories leads to systematic disadvantages for others. In contrast, BalancedFace, which was intentionally constructed to fill demographic gaps and reduce skew, achieved far more consistent performance. The maximum TPR gap across racial subgroups dropped by more than half, and Disparate Impact values moved considerably closer to parity. Importantly, BalancedFace extended fairness to age groups that had been excluded or poorly represented in prior datasets, such as children under two and adults over seventy, thereby broadening coverage in ways that are practically meaningful.

These improvements were accompanied by a modest reduction in overall accuracy. However, this should not be seen as a weakness. The slightly higher accuracy of models trained on UTKFace and FairFace was inflated by their performance on majority groups. BalancedFace, in contrast, distributed predictive performance more evenly across subgroups, offering accuracy that is more honest and inclusive. In real-world deployments where fairness is not optional but essential, such as in government services, healthcare, or identity verification, this trade-off is not only acceptable but preferable.

Taken together, the results validate the importance of treating data balance and representativeness as primary design criteria. Rather than relying solely on post-hoc corrections, building fairness into the dataset itself produces models that are both more transparent and more accountable. The evidence suggests that this strategy provides a clearer and more durable path to equitable face analysis than approaches that attempt to correct bias only at the algorithmic level.

VI. Conclusion

This work demonstrated that careful attention to intersectional fairness during model development can yield substantial improvements in the equitable performance of gender classification systems. By integrating adversarial learning, equalised odds regularisation, and a fairness-aware training strategy, the proposed approach systematically reduced disparities in true positive rates and positive outcome ratios across demographic subgroups defined by age, race, and gender. Unlike models that prioritize overall accuracy at the expense of minority group performance, our method maintains competitive accuracy while significantly narrowing fairness gaps. These results highlight that performance parity across sensitive attributes is achievable when fairness is treated as a primary optimization objective rather than a post hoc evaluation. The findings also emphasize the importance of auditing models across intersectional slices, rather than isolated attributes, to surface compounded disparities that are often overlooked.

References

[1] Z. Sun, W. Liu, and K. Chen, "Applications of face recognition technology: A comprehensive review," *IEEE Access*, vol. 10, pp. 24578–24599, 2022.

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 77–91, 2018.

[3] I. D. Raji and J. Buolamwini, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151, 2020.

[4] A. Krishnan, A. Almadan, and A. Rattani, "Understanding fairness of gender classification algorithms across gender-race groups," *arXiv preprint arXiv:2009.11491*, 2020.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[6] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546, 2021.

[7] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1528–1538, 2021.

[8] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 692–702, 2019.

[9] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

[10] K. Kotwal and S. Marcel, "Review of demographic bias in face recognition," *arXiv preprint arXiv:2502.02309*, 2025.

[11] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, S. Ghosh, N. Tomasev, F. Schroff, J. Brandt, T. Pfister, and S. Ando, "Towards measuring fairness in ai: The casual conversations dataset," in *arXiv preprint arXiv:2104.02821*, 2021.

[12] T. K. Bahiru, H. Tibebu, and I. A. Kakadiaris, "Ai data development: A scorecard for the system card framework," *arXiv preprint arXiv:2506.02071*, 2025.

[13] T. De Vries, A. Sharma, and M. Patel, "Does fairness in machine learning require fairness in data? a case study," *IEEE Access*, vol. 7, pp. 173008–173021, 2019.

[14] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa, "Anatomizing bias in facial analysis," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12351–12358, 2022.

[15] L. Yang, W. Li, and Q. Chen, "Towards fair machine learning: Survey, perspective, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 11, pp. 2113–2131, 2020.

[16] J. H. Hinnefeld, P. Cooman, N. Mammo, and R. Deese, "Evaluating fairness metrics in the presence of dataset bias," *arXiv preprint*, 2018.

[17] A. Le, D. Phung, and T. Tran, "A survey on bayesian network approaches to fairness in machine learning," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–35, 2022.

[18] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, vol. 115 of *Proceedings of Machine Learning Research*, pp. 862–872, 2020.

[19] C. Dwork, N. Immorlica, A. Kalai, C. Leiserson, I. Mironov, and T. Pitassi, "Decoupled classifiers for group-fair and efficient machine learning," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 81 of *Proceedings of Machine Learning Research*, pp. 119–127, 2018.

[20] A. Krishnan, A. Almadan, and A. Rattani, "Understanding fairness of gender classification algorithms across gender–race groups," *arXiv preprint*, 2020.

[21] S. Mittal, K. Thakral, R. Singh, M. Vatsa, T. Glaser, C. Canton Ferrer, and T. Hassner, "On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare," *Nature Machine Intelligence*, vol. 6, pp. 936–949, 2024.

[22] U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, and U.S. Department of Justice, "Uniform guidelines on employee selection procedures." 43 Fed. Reg. 38290, 1978.