Analysis of Bias in AI Facial Beauty Regressors

Chandon Hamel

Regis University

MSDS 640

Dr. Ghulam Mujtaba

April 30, 2025

Analysis of Bias in AI Facial Beauty Regressors

Beauty is in the eye of the beholder, unless the beholder has no eyes. As computer vision systems become ubiquitous across industries, their growing influence on human self-perception warrants scrutiny, particularly in beauty assessment applications.

The deployment of facial beauty prediction models potentially impacts critical domains including:

- Automated talent screening in acting/modeling agencies
- Beauty-score-targeted marketing by cosmetic companies
- Post-surgical outcome simulation in aesthetic clinics
- AI-curated profile picture selection on social platforms

These applications risk codifying and amplifying subjective beauty standards through algorithmic reproduction. When such systems exhibit ethnic bias, they perpetuate harmful stereotypes and create unequal access to opportunities tied to perceived attractiveness.

This study conducts a comparative analysis of ethnic bias in facial beauty assessment models trained on two publicly available datasets: SCUT-FBP5500 (Liang et al., 2018) and MEBeauty (Lebedeva et al., 2021). Its purpose is to analyze disparities in prediction and error distributions across ethnic groups, thereby assessing the extent and nature of bias in models trained to predict beauty scores.

Motivation

Related Works

Following the bias analysis framework demonstrated in hiring algorithms by Quer et al. (2024), this study adapts their evaluation approach to continuous beauty score prediction through two key fairness metrics:

1. Distributional Parity
    - Analogous to Statistical Parity for classification

- Assessed via Mann-Whitney U/Kruskal-Wallis tests on prediction distributions
- Reveals systemic score disparities across demographic groups

2. Error Parity

- Analogous to Equal Opportunity Difference
- Evaluated through non-parametric tests on error distributions
- Identifies differential model performance between groups

These metrics build on a bias mitigation framework from Feldman and Peake (2021), extending their fairness validation techniques from discrete classification to continuous regression tasks. Where Quer et al. (2024) focused on categorical outcomes in talent matching, this work addresses fairness in subjective beauty score prediction.

The importance of this analysis is underscored by mounting evidence of representational harms in generative AI systems. Recent studies reveal that image-generation models:

- Produce less diverse outputs than human-curated content (Bogdanova et al., 2024)
- Underrepresent women and People of Color in depictions of power and success (Gengler, 2024)
- Systematically whitify non-white faces in image-to-image transformations (Yang, 2025)

These findings suggest beauty prediction models risk amplifying similar biases through automated scoring. This study bridges this critical gap by examining how regression-based scoring systems may perpetuate harmful norms through biased prediction patterns.

Ethical Implications

The ethical concerns regarding AI-driven beauty standards are evident in both empirical research and philosophical analysis. According to a recent survey by Ilyas (2024), 50% of respondents reported that exposure to AI-generated beauty standards negatively impacted their self-esteem, while 70% agreed that such standards promote unrealistic cultural and social ideals. Furthermore, 82% of informants felt that AI-based beauty images are less inclusive in promoting

diversity across cultures. These findings underscore growing concerns about the effects of AI on self-worth and cultural inclusivity.

Philosophical perspectives further illuminate the risks posed by algorithmic beauty standards. As Zhou (2024) argues, the introduction of AI into the domain of beauty risks "warping natural beauty standards and making the original human appear increasingly imperfect." Zhou situates this concern within a broader historical context, noting that while philosophers like Plato conceived of beauty as an abstract, eternal ideal, and Hume emphasized its subjectivity, AI systems operationalize beauty in ways that are both highly specific and potentially exclusionary. By learning from biased or narrow data, AI models may reinforce harmful stereotypes, objectify marginalized groups, and amplify unattainable ideals, especially for women and people of color.

In summary, the ethical implications of AI in beauty prediction extend beyond technical fairness to encompass broader social harms. These include the perpetuation of unrealistic and exclusionary beauty ideals, the erosion of self-esteem, and the marginalization of diverse cultural expressions of beauty. Addressing these challenges requires not only technical interventions to improve model fairness, but also industry-wide commitments to transparency, inclusivity, and the responsible use of AI in shaping societal standards.

Experiment

This study employs a three-phase computational pipeline to assess ethnic bias in facial beauty prediction models, structured as follows:

1. **Model Development:** Fine-tune ResNet-152 architectures (He et al., 2015) on the SCUT-FBP5500 (Liang et al., 2018) and MEBeauty (Lebedeva et al., 2021) datasets separately, maintaining dataset isolation to preserve their inherent bias profiles.
2. **Model Evaluation:** Generate beauty score predictions using each trained model on the alternate labeled dataset it was not trained on and the FairFace (Karkkainen & Joo, 2021) dataset—a large, demographically diverse, but unlabeled set of face images. This approach enables assessment of model generalization and bias across both curated and real-world data distributions.

3. **Bias Quantification:** Apply non-parametric statistical tests (Mann-Whitney U and Kruskal-Wallis) to prediction distributions and errors across ethnic groups.

This cross-dataset validation approach enables detection of both dataset-specific biases and model-generalization fairness issues. Subsequent sections detail implementation specifics for reproducibility.

Datasets

SCUT-FBP5500 (Liang et al., 2018)

The SCUT-FBP5500 dataset contains 5,500 high-quality frontal face images, divided into four demographic groups: 2,000 Asian females, 2,000 Asian males, 750 Caucasian females, and 750 Caucasian males. Most images were sourced from the internet and each was assigned a beauty score in the range [1, 5], computed as the average rating from 60 volunteers. An example selection of images from SCUT-FBP5500 is shown in Figure 1.

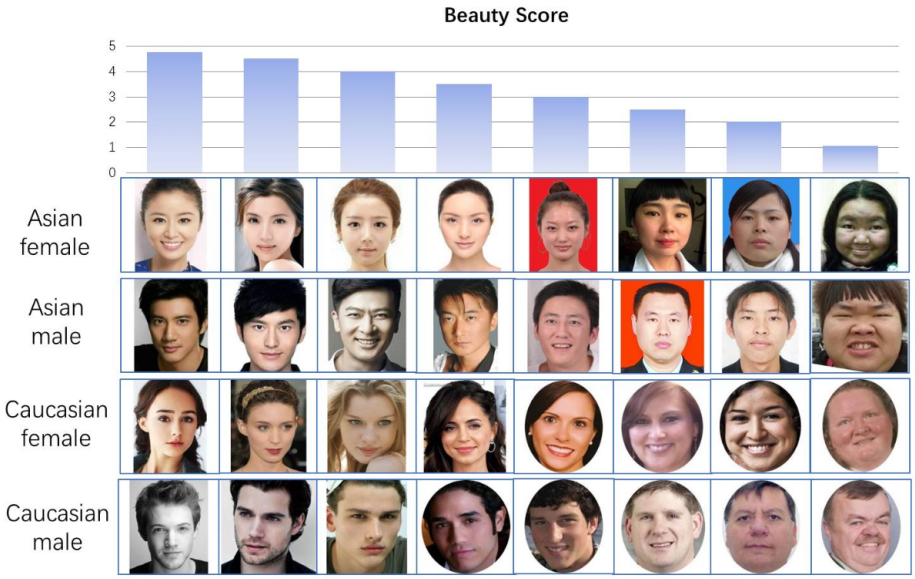

Figure 1

Sample images from the SCUT-FBP5500 dataset.

MEBeauty (Lebedeva et al., 2021)

The MEBeauty dataset comprises 2,550 images representing Black, Asian, Caucasian, Hispanic, Indian, and Mideastern female and male faces. Each image is rated on a [1, 10] beauty scale by approximately 300 individuals from diverse cultural and social backgrounds. Figure 2 presents sample images from MEBeauty.

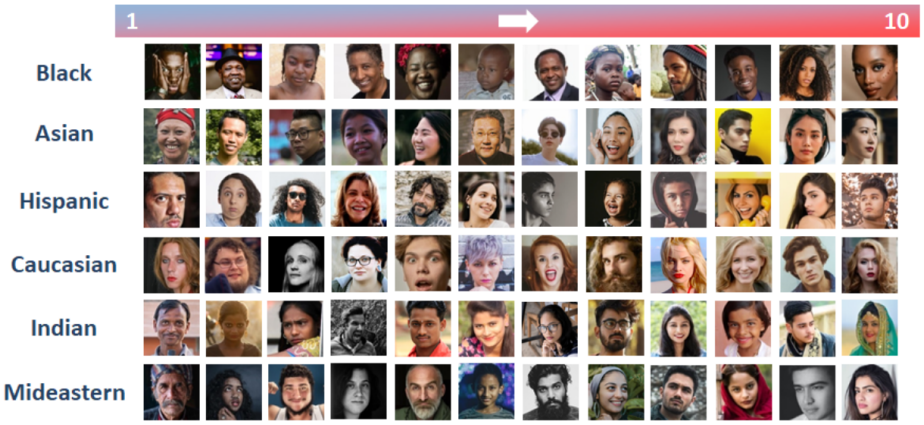

Figure 2

Sample images from the MEBeauty dataset.

FairFace Training Subset (Karkkainen & Joo, 2021)

The FairFace training subset consists of 86,744 images spanning seven race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Designed to mitigate race bias, FairFace emphasizes balanced representation across these groups. Images were collected from the YFCC-100M Flickr dataset (Thomee et al., 2016) and are annotated with race, gender, and age group labels.

Preprocessing

Metadata

For each dataset, a tabular metadata file was constructed with columns for the image filename, the subject's race, and, for the SCUT-FBP5500 and MEBeauty datasets, the corresponding labeled beauty score. To ensure consistency across datasets, all beauty scores were

normalized to the range [0, 1] prior to model training.

Images

Facial regions were extracted from each image using the Multi-task Cascaded Convolutional Networks (MTCNN) face detector (Zhang et al., 2016). This process removed background artifacts and ensured that only the subject's face was retained. To maintain aspect ratio and avoid distortion, images were padded with black borders to produce square images. MTCNN failed to detect faces in 2,015 images (2.3%) from the FairFace dataset, which were subsequently excluded from further analysis. Figure 3 illustrates examples of the face cropping and padding transformation.

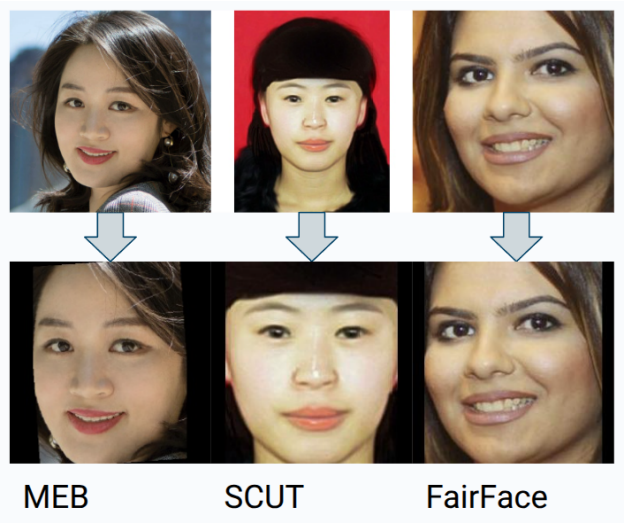

Figure 3

Examples of face cropping and padding using MTCNN.

Model Training

Data Loading

Training pipelines were implemented using PyTorch and PyTorch Lightning. A custom `Dataset` class performed the following operations for training data:

- **Augmentation:** Random horizontal flips, $\pm 10^{\circ}$ rotations, and resized crops (80-100% scale) of training images

- **Normalization:** Pixel values scaled using ImageNet means ($\mu = [0.485, 0.456, 0.406]$) and standard deviations ($\sigma = [0.229, 0.224, 0.225]$) (Deng et al., 2009)

Datasets were split into training ($\frac{2}{3}$), validation ($\frac{2}{9}$), and test ($\frac{1}{9}$) subsets. Separate dataloaders were created for:

- Training/validation/test splits of the source dataset
- External evaluation datasets (cross-dataset and FairFace)

Batch sizes were optimized per dataset—64 for SCUT-FBP5500 and 32 for MEBeauty—for computational efficiency and stability during training.

ResNet-152 Fine-tuning

The base ResNet-152 architecture (He et al., 2015) was adapted for regression by replacing its final fully connected layer with a single-output linear layer. Training proceeded in three progressive phases:

1. **Frozen Backbone:** Only the final layer trained using Adam optimizer ($\eta = 3 \times 10^{-3}$), ReduceLROnPlateau scheduler (factor=0.5, patience=2 epochs), and early stopping (patience=5 epochs)
2. **Unfreeze Conv5:** Resumed training with conv5 block unfrozen
3. **Unfreeze Conv4:** Final training phase with conv4 block unfrozen

Models were continuously monitored via validation MSE loss, with restoration to best-performing checkpoints between phases.

Bias Analysis

Model Performance

Table 1 presents the performance metrics for both models, evaluated on their respective test sets and cross-validated on the alternative dataset. The SCUT-trained model achieved a test Mean Squared Error (MSE) of 0.008, while the MEBeauty-trained model reached 0.013. These results are competitive with state-of-the-art performance on facial beauty prediction tasks.

Model	Test MSE	Cross-dataset MSE
SCUT-trained	0.008	0.024
MEBeauty-trained	0.013	0.028

Table 1

Model performance measured by Mean Squared Error (MSE) on normalized [0,1] scale. Test MSE represents performance on the held-out test set from the training dataset. Cross-dataset MSE shows performance when SCUT-trained model is evaluated on MEBeauty and vice versa.

For contextual comparison, Liang et al. (2018) reported an optimal RMSE of 0.302 on a [1, 5] scale using ResNeXt-50 (Xie et al., 2017). This converts to approximately 0.076 on this studies normalized [0, 1] scale ($0.302 \div 4 = 0.0755$). This study's SCUT-trained model's RMSE of 0.089 ($\sqrt{0.008} \approx 0.089$) demonstrates comparable performance despite using a different, slightly earlier architecture.

Cross-dataset evaluation shows expected performance degradation, with MSE increasing by factors of 3.0 and 2.2 for the SCUT and MEBeauty models respectively. This degradation likely reflects underlying differences in beauty annotation protocols, demographic composition, and visual characteristics between datasets.

MEBeauty Data Analysis

Figure 4 presents a pairplot visualization of the MEBeauty dataset, showing distributions of ground truth labels, model predictions from the SCUT-trained model, and prediction errors across ethnic groups. Visual inspection reveals notable distributional differences that suggest potential bias in both the dataset annotations and model predictions.

Quantitative analysis of the SCUT-trained model's performance on MEBeauty data is summarized in Table 2. The results reveal substantial disparities in beauty score predictions across ethnic groups. Caucasian and Middle Eastern faces received significantly higher mean predicted beauty scores (0.65 and 0.66 respectively) compared to Asian and Black faces (both 0.56). Error analysis further indicates systematic bias, with Middle Eastern faces showing the

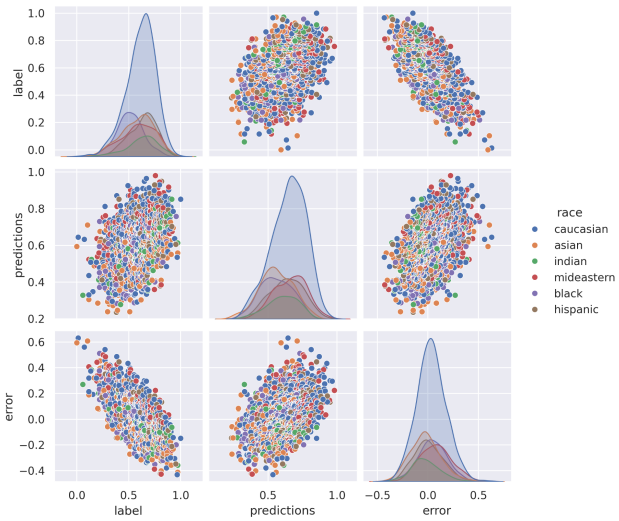

Figure 4

Pairplot of ground truth labels, model predictions, and prediction errors from the SCUT-trained model applied to the MEBeauty dataset, segmented by ethnic group.

largest positive error (0.07), suggesting consistent overestimation of beauty scores for this group.

Statistical validation via Kruskal-Wallis tests strongly rejects the null hypothesis of equal distribution across ethnic groups for both predictions ($H = 229$, $p = 1.9 \times 10^{-47}$) and errors ($H = 101$, $p = 3.6 \times 10^{-20}$). These findings were further confirmed through permutation-based ANOVA tests, which yielded consistent significance levels ($p = 2 \times 10^{-4}$) for both metrics.

To identify specific inter-group disparities, Post Hoc Dunn tests were conducted with Benjamini-Hochberg p-value adjustment to control for multiple comparisons. Figure 5 visualizes these pairwise comparisons for predictions, revealing that only 2 out of 15 ethnic group pairs (13.3%) satisfy the Distributional Parity criterion at $\alpha = 0.05$. Similarly, Figure 6 shows that only 3 pairs (20%) meet the Error Parity standard. These results provide strong statistical evidence of ethnic bias in the SCUT-trained model when applied to the MEBeauty dataset.

SCUT Data Analysis

Figure 7 presents a pairplot visualization of the SCUT-FBP5500 dataset, illustrating distributions of ground truth labels, model predictions, and prediction errors across Asian and Caucasian ethnic groups. Visual inspection reveals distinct distributional patterns that suggest

Table 2

Statistical analysis of predictions and errors from the SCUT-trained model on the MEBeauty dataset, segmented by ethnic group. Lower rows show Kruskal-Wallis test results assessing equality of distributions across groups.

	Predictions		Errors	
Race	Mean	Median	Mean	Median
Asian	0.56	0.56	-0.02	-0.03
Black	0.56	0.56	0.04	0.04
Caucasian	0.65	0.66	0.03	0.03
Hispanic	0.63	0.64	0.00	-0.00
Indian	0.60	0.61	-0.02	-0.03
Middle Eastern	0.66	0.66	0.07	0.07
KW Test Statistic	229		101	
KW Test P-value	$< 10^{-3}$		$< 10^{-3}$	

potential bias in both the dataset annotations and the MEBeauty-trained model's predictions.

Table 3 provides a quantitative analysis of the MEBeauty-trained model's performance on SCUT-FBP5500 data. The results indicate a notable difference in prediction patterns between Asian and Caucasian faces. While mean prediction values are similar (0.5929 for Asian vs. 0.5883 for Caucasian), the error distributions show more substantial disparities. Asian faces exhibit significantly higher mean errors (0.0722) compared to Caucasian faces (0.0380), suggesting systematic overestimation of beauty scores for Asian subjects relative to ground truth labels.

Statistical validation was performed using both Mann-Whitney U (MWU) and Kolmogorov-Smirnov (KS) tests. The MWU test, which assesses differences in distribution medians, rejects the null hypothesis of equal prediction distributions with moderate significance

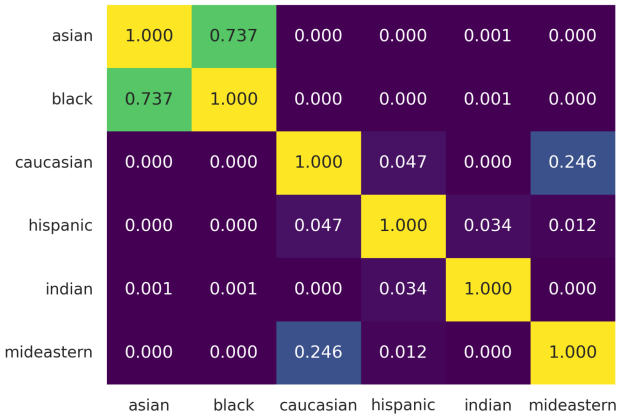

Figure 5

Heatmap of pairwise post hoc Dunn test results for model predictions across ethnic groups (SCUT-trained model on MEBeauty data). Each cell displays the adjusted p-value for the comparison between two groups after Benjamini-Hochberg correction.

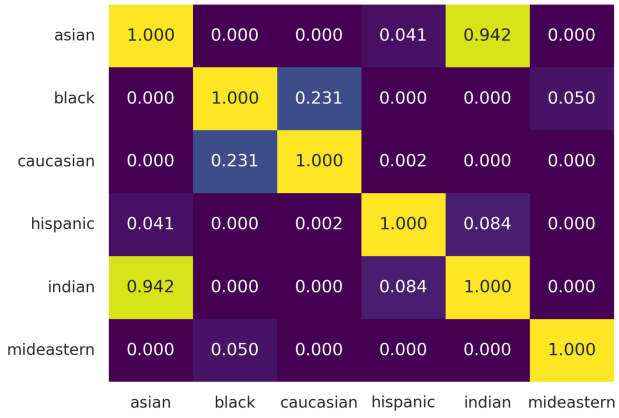

Figure 6

Heatmap of pairwise post hoc Dunn test results for prediction errors across ethnic groups (SCUT-trained model on MEBeauty data). Each cell displays the adjusted p-value for the comparison between two groups after Benjamini-Hochberg correction.

($U = 2,869,019$, $p = 0.0124$) and strongly rejects equal error distributions ($U = 2,562,921.5$, $p = 7.8 \times 10^{-17}$). The KS test, which is sensitive to differences in distribution shape and spread, provides even stronger evidence of distributional differences for both predictions ($D = 0.0702$, $p = 4.1 \times 10^{-5}$) and errors ($D = 0.146$, $p = 9.8 \times 10^{-21}$).

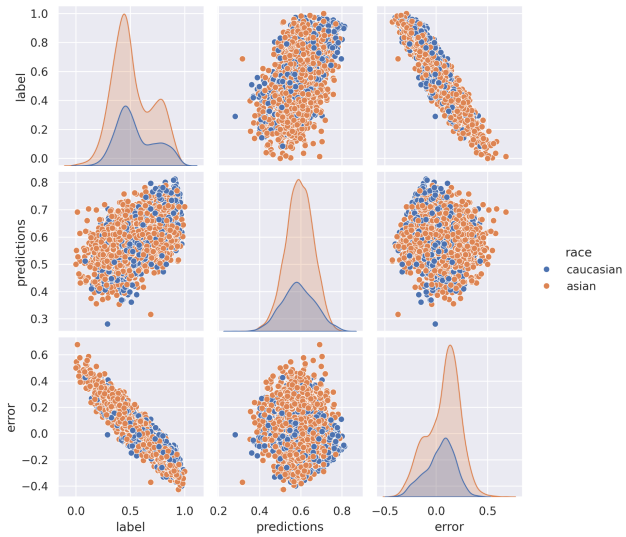

Figure 7

Pairplot of ground truth labels, model predictions, and prediction errors from the MEBeauty-trained model applied to the SCUT-FBP5500 dataset, segmented by ethnic group (Asian and Caucasian).

Table 3

Statistical analysis of predictions and errors from the MEBeauty-trained model on the SCUT-FBP5500 dataset, segmented by ethnic group. Lower rows show Mann-Whitney U and Kolmogorov-Smirnov test results assessing equality of distributions across groups.

	Predictions		Errors	
Ethnic Group	Mean	Median	Mean	Median
Asian	0.5929	0.5938	0.0722	0.1004
Caucasian	0.5883	0.5859	0.0380	0.0582
MWU Statistic	2,869,019		2,562,921.5	
MWU P-value	0.012		$< 10^{-3}$	
KS Statistic	0.0702		0.146	
KS P-value	$< 10^{-3}$		$< 10^{-3}$	

These results demonstrate that the MEBeauty-trained model fails to satisfy both Distributional Parity and Error Parity criteria when applied to the SCUT-FBP5500 dataset, providing compelling evidence of ethnic bias in cross-dataset applications of beauty prediction models.

FairFace Data Analysis

To assess bias generalization to a more diverse population, both trained models were evaluated on the demographically balanced FairFace dataset. Figure 8 presents a pairplot visualization comparing prediction distributions across seven ethnic groups. Visual inspection reveals slight patterns of differential treatment across ethnicities by both models. Though distribution overlap is significant given the volume of samples in such a narrow prediction window, statistical tests reveal robust disparities in predictions across groups for both models.

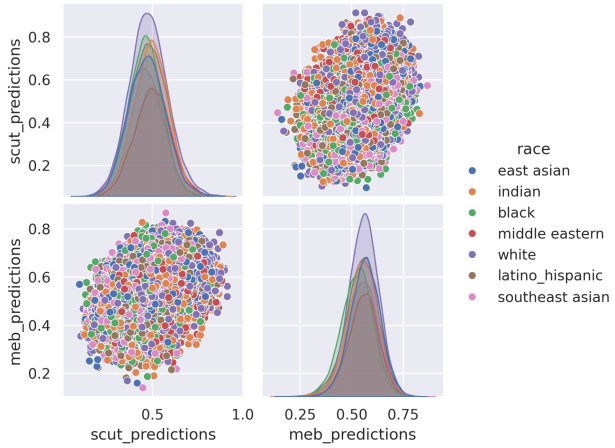

Figure 8

Pairplot of beauty score predictions from SCUT-trained and MEBeauty-trained models applied to the FairFace dataset, segmented by ethnic group. Differences in distribution shapes and centers reveal systematic prediction disparities across ethnicities.

Table 4 quantifies these disparities through descriptive statistics. Kruskal-Wallis tests strongly reject the null hypothesis of equal prediction distributions across ethnic groups for both the SCUT-trained model ($H = 1675.7$, $p < 10^{-3}$) and MEBeauty-trained model ($H = 1716.8$, $p < 10^{-3}$). These significant results confirm systematic differential treatment by race in both

models.

Table 4

Comparison of beauty score predictions from SCUT-trained and MEBeauty-trained models on the FairFace dataset across seven ethnic groups. Lower rows display Kruskal-Wallis test results assessing equality of prediction distributions.

	SCUT Model		MEBeauty Model	
Ethnic Group	Mean	Median	Mean	Median
Black	0.460	0.461	0.525	0.527
East Asian	0.470	0.469	0.563	0.566
Indian	0.493	0.494	0.550	0.555
Hispanic	0.479	0.479	0.548	0.551
Middle Eastern	0.500	0.500	0.555	0.559
Southeast Asian	0.455	0.455	0.547	0.551
White	0.479	0.477	0.556	0.559
KW Test Statistic	1675.7		1716.8	
KW Test P-value	$< 10^{-3}$		$< 10^{-3}$	

Post Hoc Dunn tests with Benjamini-Hochberg corrections reveal the specific inter-group disparities. Figure 9 shows that for the SCUT-trained model, only 1 out of 21 pairwise comparisons (4.8%) satisfies Distributional Parity at $\alpha = 0.05$. Similarly, Figure 10 demonstrates that the MEBeauty-trained model upholds Distributional Parity in only 2 of 21 comparisons (9.5%). This analysis on a balanced dataset provides evidence that beauty prediction models systematically encode and reproduce ethnic biases.

Discussion

This study reveals ethnic bias in beauty prediction models, challenging an initial hypothesis that MEBeauty's broader ethnic representation would yield fairer models than

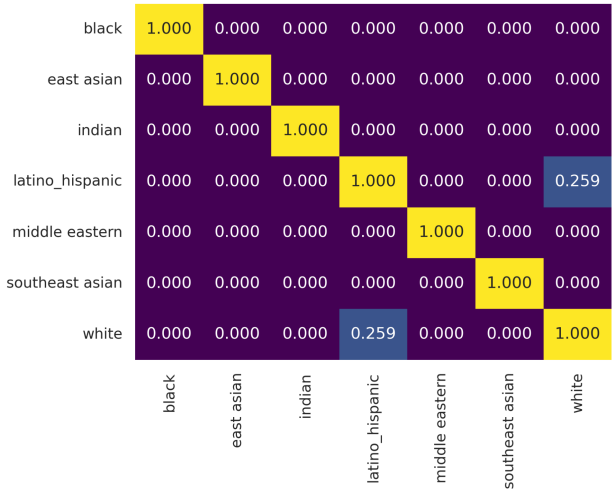

Figure 9

Heatmap of pairwise post hoc Dunn test results for SCUT-trained model predictions across ethnic groups on FairFace data. Each cell displays the adjusted p-value after Benjamini-Hochberg correction.

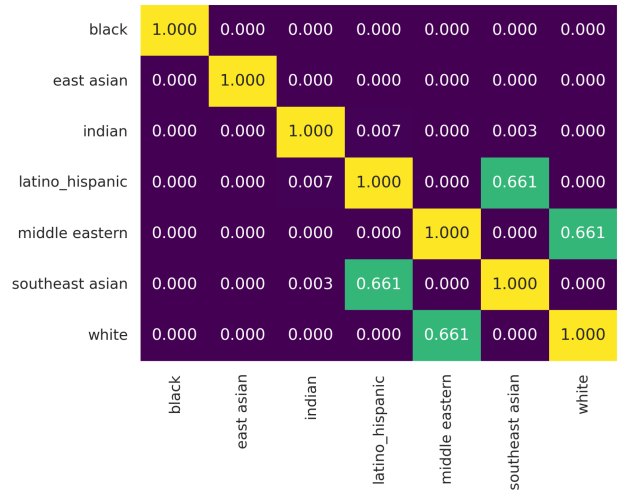

Figure 10

Heatmap of pairwise post hoc Dunn test results for MEBeauty-trained model predictions across ethnic groups on FairFace data. Each cell displays the adjusted p-value after Benjamini-Hochberg correction.

SCUT-FBP5500. This finding aligns with Chekanov et al. (2023)'s framework for algorithmic bias propagation, where multiple systemic factors compound to produce discriminatory outcomes.

Sources of Observed Bias

- **Sampling Bias:** As visualized in Figures 4 and 7, both training datasets exhibit skewed label distributions across ethnic groups. SCUT-FBP5500's Asian/Caucasian dichotomy and MEBeauty's uneven score distributions suggest underlying sampling imbalances in original data collection.

- **Labeling Bias:** The consistent error disparities (Tables 2-3) point to systemic labeling inconsistencies. Even with multi-rater annotations, the lack of demographic parity and transparency among annotators likely introduced preference patterns into the training labels.

Technical and Ethical Implications

These findings carry significant consequences for real-world deployment:

- **Model Generalization:** Both models exhibited exacerbated bias on the balanced FairFace dataset (Table 4), suggesting that beauty prediction systems potentially amplify rather than mitigate existing societal biases when applied to diverse populations.

- **Fairness-Aware Production:** While the fairness criteria—requiring distributional parity across all ethnic comparisons—may seem stringent, this standard is necessary because even slight disparities can have outsized negative impacts on marginalized groups and should be the goal for responsible AI deployment.

Paths Forward

- **Algorithmic Mitigation:** Integration of fairness constraints during training, such as the frameworks proposed by Yik and Silva (2024) and Yazdani-Jahromi et al. (2024), could help promote fairness along with model performance.

- **Data Curation:** Future datasets should:

- Enforce demographic balance through stratified sampling
- Implement annotator diversity quotas mirroring target populations
- Include metrics to capture cultural relativity and representation

- **Validation Protocols:**

- Bias testing as part of model validation
- Transparent reporting of sampling, labeling, and disparity metrics across groups

The road to equitable beauty AI remains challenging, but these results demonstrate that current approaches risk cementing harmful stereotypes rather than advancing inclusivity. The normalization of algorithmic beauty standards threatens to eclipse humanity's rich diversity of self-expression.

Conclusion

This study demonstrates that facial beauty prediction models have the potential to systematically encode ethnic biases, with both SCUT- and MEBeauty-trained models showing significant disparities across groups ($p < 0.001$). The cross-dataset validation on FairFace revealed exacerbated bias in balanced populations, indicating amplification rather than mitigation of societal prejudices. These biases stem from compounded representation and annotation limitations in beauty datasets, where sampling imbalances and biased ratings propagate through model training.

Beauty AI systems require fairness-constrained architectures (Yazdani-Jahromi et al., 2024), demographic equity validation protocols, and ethically curated datasets with diverse annotators. Unchecked deployment risks cementing algorithmic beauty standards that erase cultural diversity. Future work must prioritize decoupling aesthetic assessment from ethnic features while preserving model utility. This will be a critical step toward equitable AI in a socially impactful application.

References

Bogdanova, D., Haritos, A. V., & Kieschnick, L. (2024). Assessing diversity in ai-generated images for beauty campaigns. In C. Baumgarth (Ed.), *Special issue innovative brand management iv* (pp. 1–28). Berlin School of Economics; Law (HWR Berlin).

Chekanov, K., Georgievskaya, A., Tlyachev, T., Danko, D., & Corstjens, H. (2023). How artificial intelligence adopts human biases: The case of cosmetic skincare industry. *npj Digital Medicine*, *6*(1), 194. https://link.springer.com/article/10.1007/s43681-023-00378-2

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Feldman, T., & Peake, A. (2021). End-to-end bias mitigation: Removing gender bias in deep learning. *arXiv preprint arXiv:2104.02532*. https://arxiv.org/abs/2104.02532

Gengler, E. J. (2024). Sexism, racism, and classism: Social biases in text-to-image generative ai in the context of power, success, and beauty. *Wirtschaftsinformatik 2024 Proceedings*, 48. https://aisel.aisnet.org/wi2024/48

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*. https://arxiv.org/abs/1512.03385

Ilyas, Q. (2024). *Ai in beauty content: A theoretical study of ethical and psychological considerations surrounding ai-generated beauty content* [Bachelor's Thesis]. Vaasa University of Applied Sciences [Degree Program of Beauty and Cosmetics - Beauty Care (UAS)]. https://www.theseus.fi/handle/10024/872723

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*, 1548–1558.

Lebedeva, I., Guo, Y., & Ying, F. (2021). Mebeauty: A multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, 1–15.

Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018). Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. *ICPR*.

Quer, D., Via, A., Serra-Vidal, M., Nadal, L., & Fortuny, D. (2024). Analyzing bias and discrimination in an algorithmic hiring use case. *AIMMES 2024 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies, co-located with EU Fairness Cluster Conference 2024*.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73. https://doi.org/10.1145/2812802

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. https://arxiv.org/abs/1611.05431

Yang, Y. (2025). Racial bias in ai-generated images [Open Access, Creative Commons Attribution 4.0 International License]. *AI & Society*. https://doi.org/10.1007/s00146-025-02282-1

Yazdani-Jahromi, M., Yalabadi, A. K., Rajabi, A., Tayebi, A., Garibay, I., & Garibay, O. (2024). Fair bilevel neural network (fairbinn): On balancing fairness and accuracy via stackelberg equilibrium. *arXiv preprint arXiv:2410.16432*. https://arxiv.org/abs/2410.16432

Yik, W., & Silva, S. J. (2024). Enforcing equity in neural climate emulators. *arXiv preprint arXiv:2406.19636*. https://arxiv.org/abs/2406.19636

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/lsp.2016.2603342

Zhou, B. (2024). *Ai and human beauty standards* [Manuscript]. https://philarchive.org/rec/ZHOAAH