

Demographic Differences in Automated Facial Attractiveness Prediction: Evidence from SCUT-FBP5500 and MEBeauty

Aila Sajid Bhangoo

Operations and Analytics Department, The Kania School of Management, University of
Scranton, Scranton, PA 18501, USA

Usama Younus

Brendan Iribe Center for Computer Science and Engineering, University of Maryland, College
Park, MD 20742, USA

Haven Yi-Suh Shin

Plamica Labs, Batten Hall, 125 Western Ave., Allston, MA 02163, USA

Asma Hamid Awan (Corresponding author)

Head of the Department of Political Science, Kinnaird College for Women, Lahore, Punjab
54500, Pakistan

E-mail: asma.awan@kinnaird.edu.pk

Received: July 19, 2026 Accepted: September 28, 2026 Published: October 9, 2026

doi:10.5296/ijssr.v14i2.24083 URL: <https://doi.org/10.5296/ijssr.v14i2.24083>

Abstract

In this work, we examine automated facial beauty prediction (FBP) models to evaluate demographic variation in human attractiveness scores, facial landmark preprocessing, and cross-dataset prediction error. We perform multistage quantitative sociotechnical audits using SCUT-FBP5500 and a 1,962-image MEBeauty subset, including DLib landmark-extraction analysis, intersectional ordinary least squares (OLS) regression, observed-rating inter-rater variability analysis, and a five-seed ResNet-50 cross-dataset evaluation. DLib landmark extraction was successful for 1,929 of 1,962 MEBeauty images, with an overall failure rate of

1.7%. The highest observed subgroup failure rate was among Middle Eastern men (4.2%; 2/48). In the symmetry-adjusted OLS model ($N = 1,929$, $R^2 = .022$), no gender, ethnicity, or gender $\times$ ethnicity coefficient was statistically significant at the Benjamini–Hochberg correction. At the same time, blur was the sole significant image quality control ($p = .006$). Subgroup RMSE estimates across datasets ranged from .470 to .512. Black men had the highest point estimate (RMSE = .512), but prespecified image-level bootstrap comparisons with Asian men (difference = .002, 95% CI $[-.09, .09]$) and White men (difference = .022, 95% CI $[-.05, .09]$) both included zero. These results show how disaggregated sociotechnical auditing can be useful and also emphasize the significance of disentangling descriptive subgroup variation from statistically supported demographic disparities.

Keywords: algorithmic bias, facial beauty prediction, computational social science, fairness audit, cross-dataset evaluation

1. Introduction

1.1 Social Significance of Beauty as Stratification

Algorithmic beauty assessment matters sociologically because appearance-based evaluation can shape access to digital and physical opportunities. This article examines demographic variation in human attractiveness ratings, facial landmark preprocessing, and automated prediction errors to assess whether the evaluated pipeline exhibits demographic disparity. Physical attractiveness functions as aesthetic capital, producing documented “beauty premiums” in earnings, employment, and social treatment (Mears, 2014). Evolutionary accounts often link attractiveness to symmetry and skin consistency (Dimitrov & Kroumpouzos, 2023).

However, recent sociological research fundamentally challenges biological determinism by demonstrating that aesthetic standards vary across social and historical contexts. For instance, the aesthetic standards that defined the Renaissance drastically differ from modern normative standards of visual representation (Abid et al., 2021). Furthermore, multinational studies confirm that there is no universal, biologically predetermined aesthetic criterion; rather, evaluative repertoires depend heavily on demographic characteristics such as age, educational attainment, and geographic location (Kuipers, 2015).

In contemporary digital ecosystems, these offline biases can be captured by facial beauty prediction (FBP) systems. When FBP models are trained on datasets reflecting historical prejudices, they risk translating contingent beauty norms into apparently neutral algorithmic rankings.

1.2 What the Computer-Science Literature Says About Bias in Facial-Attractiveness Algorithms

Facial beauty prediction (FBP) is the computer-vision task of mapping facial images to numerical attractiveness scores. Regardless of the chosen architecture, these models inherit several well-documented sources of bias. To properly audit these systems, it is necessary to examine how FBP is formulated, evaluated, and deployed.

1.2.1 Task Formulation

FBP may be formulated as continuous regression, ordinal prediction, classification into score bands, pairwise ranking, or listwise ranking. In this study, FBP is formulated as a continuous regression task utilising a single-node linear output and Mean Squared Error (MSE) loss.

1.2.2 Historical Methods

Historically, FBP relied heavily on handcrafted geometric methods. These early approaches extracted specific facial ratios, landmark distances, symmetry measurements, handcrafted texture features, and facial averageness parameters, feeding them into traditional feature-engineering pipelines to estimate beauty.

1.2.3 Deep-Learning Approaches

Modern FBP predominantly employs deep Convolutional Neural Networks (CNNs). Architectures such as VGG and ResNet typically leverage transfer learning from large datasets like ImageNet, utilising either partial feature extraction or full fine-tuning. More recent literature also explores multi-task learning and attention- or transformer-based models to capture complex relationships in facial aesthetics (Danner et al., 2021).

1.2.4 Evaluation Metrics

Model performance is standardly evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Pearson correlation, Spearman correlation, pairwise accuracy, and ranking accuracy. MAE and RMSE are utilised in this study because they directly measure the magnitude of prediction error on the original rating scale, allowing for precise and interpretable subgroup fairness evaluation.

1.2.5 Within-Dataset Versus Cross-Dataset Evaluation

A critical distinction in FBP evaluation is between within-dataset testing (evaluating on random train/test splits from the same dataset) and cross-dataset testing (evaluating on an entirely different dataset). Models trained on homogeneous datasets frequently exhibit high cross-dataset prediction error due to domain shift caused by differences in lighting, pose, ethnicity composition, rating procedures, and image acquisition when applied to diverse “in-the-wild” populations.

1.2.6 Fairness Evaluation and Preprocessing Bias

Rigorous fairness evaluation requires reporting subgroup MAE, subgroup RMSE, signed residuals to capture directional bias, and error differences with uncertainty intervals. Furthermore, bias can enter the pipeline through preprocessing before predictive evaluation begins. Commercial facial analysis systems have shown intersectional accuracy disparities affecting darker-skinned people and women (Buolamwini & Gebru, 2018). Such pipeline bias can systematically exclude marginalised groups from analysis or downstream platforms entirely; landmark extraction therefore requires its own empirical audit rather than an assumed performance gap.

1.3 Research Gap

While prior work has identified demographic disparities in facial attractiveness models, this literature has not sufficiently integrated fairness diagnostics with sociological theories of aesthetic capital using a fully disaggregated, intersectional methodology. Previous studies often collapse ethnic groups into binary categories (e.g., Asian/Non-Asian), masking potential group-specific or intersectional disparities.

Although previous FBP studies have examined predictive accuracy and demographic variation, fewer studies combine fully disaggregated ethnicity-by-gender regression, subgroup cross-dataset error analysis, and an empirical audit of landmark-detection failures in a single sociotechnical evaluation. This study addresses this gap by executing a twelve-group

intersectional analysis and empirically auditing the DLib preprocessing pipeline.

1.4 Research Questions

To systematically interrogate these dynamics, this study is guided by the following research questions:

RQ1: Do demographic differences in attractiveness ratings persist after controlling for facial symmetry and image-quality covariates?

RQ2: How much agreement exists among raters, and does agreement differ across demographic groups?

RQ3: Does a model trained on a controlled-attractiveness dataset exhibit uneven, group-specific prediction errors when applied to a heterogeneous dataset, specifically across the 12 ethnic and gender intersections in MEBeauty?

1.5 Contributions

This article makes three contributions. First, it uses aesthetic capital and intersectionality to motivate a multistage audit of human ratings, preprocessing success, and prediction error, while distinguishing these measured outcomes from broader social mechanisms that the data cannot directly test. Second, it contributes empirically by applying a fully disaggregated twelve-group gender $\times$ ethnicity framework to test, rather than assume, demographic disparities in attractiveness ratings and cross-dataset prediction error. In the current reproducible analysis, no demographic main effect or interaction remained statistically significant after multiple-testing correction, and the two prespecified cross-dataset RMSE contrasts also included zero. This null pattern is substantively important because it demonstrates the value of disaggregated auditing while cautioning against inferring demographic inequity solely from subgroup point estimates. Third, it contributes methodologically by demonstrating how symmetry-controlled regression, interrater variance analysis, and a preprocessing failure audit can be combined into a computational social science audit.

2. Theoretical and Conceptual Framework

This study uses aesthetic capital and intersectionality to motivate a demographic audit of facial attractiveness assessment. These perspectives inform the choice of outcomes and demographic comparisons; the analyses do not directly test cultural hegemony, aesthetic labour, or the social consequences of appearance-based evaluation.

Aesthetic capital describes how appearance can acquire social value and contribute to advantage or disadvantage (Mears, 2014; Hanquinet et al., 2014). In this study, human attractiveness ratings are treated as recorded evaluations of appearance, rather than measures of an objective or universal standard of beauty. The regression examines whether these recorded evaluations are associated with gender, ethnicity, and their interactions after adjustment for measured symmetry and image characteristics. It does not measure the conversion of appearance into employment, income, or social status.

An intersectional perspective motivates retaining the twelve gender $\times$ ethnicity groups in MEBeauty rather than collapsing ethnicity into a binary category. Gender $\times$ ethnicity terms allow the regression to examine whether the association between gender and ratings differs across the recorded ethnicity categories. Group-specific summaries also make variation in rater disagreement, preprocessing success, and prediction error visible. These comparisons test for differences without assuming that disadvantage necessarily compounds across demographic categories.

The audit distinguishes human evaluation, technical inclusion, and predictive performance. Within-image rating dispersion describes disagreement among observed raters; DLib extraction success describes whether images pass preprocessing; and residuals, MAE, and RMSE describe prediction error relative to the available human reference scores. Cultural-hegemony perspectives motivate scrutiny of whose evaluations are represented in those scores, but the datasets do not directly measure the social mechanisms producing them. The results are therefore interpreted as evidence about specific audit outcomes, rather than a direct test of broader theories of aesthetic stratification.

3. Methodology

3.1 Research Design: A Multistage Quantitative Sociotechnical Audit

The study applies a multistage quantitative sociotechnical audit comprising dataset harmonisation, preprocessing and failure analysis, intersectional regression, interrater variability analysis, and cross-dataset predictive evaluation. It examines demographic variation in recorded human ratings, the inclusion of images after landmark extraction, and prediction errors relative to human reference scores. These outcomes provide complementary audit measures but do not directly measure cultural hegemony, aesthetic labour, or downstream social discrimination.

Prior work on facial evaluation suggests that rater agreement varies across different types of judgments, with attractiveness often showing greater subjective variation than more directly observable demographic attributes (Kramer et al., 2018). The OLS regression examines demographic associations in human attractiveness scores after adjustment for measured covariates, whereas the cross-dataset CNN evaluation examines prediction errors relative to those scores.

3.2 Data Sources and Historical Context

The research examined publicly accessible datasets that serve as foundational benchmarks in computer vision and computational aesthetics. This choice demonstrates the evolution of facial assessment, which has transitioned from manual anthropometric methods to deep-learning architectures.

3.2.1 SCUT-FBP5500 Dataset

SCUT-FBP5500 comprises 5,500 controlled frontal-face images (50/50 gender balance; primarily Asian/Caucasian), with inclusion criteria: complete demographics, no occlusion ($>80\%$ landmarks detected); final $N = 5,500$ (Female-Asian: 2,000; Male-Asian: 2,000;

Female-Non-Asian: 750; Male-Non-Asian: 750) (Liang et al., 2018). Standardised illumination and pose were used during data acquisition to minimise extrinsic variability. The dataset's ratings were produced primarily by Asian annotators (Hamel & Busch, 2025), a documented source of rater bias in the FBP literature (Danner et al., 2021), which is treated here as a limitation of the ground truth rather than of the present analysis alone.

From a sociological perspective, this dataset constructs a highly specific normative ideal. While biometric tools can precisely measure skin symmetry and evenness, early psychological studies indicate that attractiveness is frequently associated with mathematically average facial shapes rather than anomalous features (Strzalko & Kaszycka, 1992). However, the systemic exclusion of darker-skinned demographics explicitly reinforces prevailing, hegemonic beauty ideals—most notably, Eurocentric and East Asian phenotypic characteristics.

3.2.2 MEBeauty Dataset

The image-based MEBeauty analyses use a 1,962-image export from the MEBeauty Hugging Face fbp training split (Lebedeva et al., 2022). The sample spans six ethnicity categories—Asian, White, Black, Hispanic, South Asian, and Middle Eastern—and two gender categories used for the primary intersectional analyses. Gender and ethnicity metadata were joined to the exported image identifiers using the accompanying MEBeauty metadata files. MEBeauty image-level attractiveness scores were transformed from the original 1–10 scale to a 1–5 scale using:

$$y = 1 + \frac{4}{9}(x - 1)$$

Where x denotes the original score and y denotes the harmonised score. This transformation maps the original endpoints of 1 and 10 to 1 and 5, respectively. Individual rater scores used in the interrater variability analysis remained on the original 1–10 scale.

DLib landmark extraction was successful for 1,929 images and failed for 33 images; consequently, analyses requiring DLib-derived facial features and the cross-dataset fairness audit use the DLib-successful analytical sample of $N = 1,929$. The binary Asian/Non-Asian coding is retained only for SCUT-FBP5500 where required by the original dataset structure; MEBeauty analyses preserve the twelve gender $\times$ ethnicity intersections.

The interrater variability analysis is not performed on the same sample as the 1,962-image preprocessing/CNN subset, since it uses the long-format MEBeauty rater-level data, which are available separately. The rater-level sample and its demographic mapping are described separately in Section 3.4.

3.3 Variables and Operationalisation

3.3.1 Dependent Variables

For SCUT-FBP5500, attractiveness is computed as the average of the available source ratings on the 1–5 scale per image. For the image-based analyses of MEBeauty, the provider

attractiveness score was transformed as specified in Section 3.2.2. Individual MEBeauty rater scores remain on the original 1–10 scale and are used independently for the interrater variability analysis; they cannot be interchanged with the image-based CNN sample. Due to the unbalanced rater matrix, agreement is estimated directly from the observed ratings via the within-image rating standard deviation for each image, rather than by constructing a rectangular rater matrix or imputing missing ratings. Lower within-image standard deviation indicates greater agreement, and higher values indicate greater disagreement between raters.

3.3.2 Independent Variables

Demographic Attributes: The focal demographic variables are Gender (Male/Female) and Ethnicity, parameterised as six specific categorical groups rather than as a binary classification. The corresponding twelve gender $\times$ ethnicity intersections are retained for the fully disaggregated MEBeauty analyses.

Image-Quality Covariates: To characterise potentially relevant image conditions, three covariates were extracted and standardised before modelling:

(1) **Blur:** operationalized as the variance of the Laplacian computed on the grayscale image as processed by the analysis pipeline.

(2) **Brightness:** measured as the mean grayscale pixel intensity on the 0–255 scale.

(3) **Image-centring proxy:** A horizontal centring proxy was derived from the displacement of the grayscale intensity centroid from the image midline, normalised by image half-width and scaled by a constant factor of 10 for numerical convenience in the analysis pipeline. Formally,

$$\text{Centring} = 10 \times \frac{|c_x - w/2|}{w/2}$$

where c_x is the horizontal intensity centroid and w is the image width. Larger values indicate greater horizontal off-centring. Because of this scaling factor, observed values may exceed 1. The measure was z-standardised before modelling and entered as an image-level adjustment covariate. It should be interpreted as a proxy for image centring, not as a full 3D head-pose estimate. The scaling constant is applied by the preprocessing routine that generated the stored covariate values: the function `compute_pose_placeholder()` in `02_preprocess_images.py` of the reproducibility package multiplies the half-width-normalised displacement by a factor of 10 before the result is written to the stored covariate column. Because the constant is common to all images, it cancels under z-standardisation, so the covariate entered into the models is unaffected by it.

Morphological Features (Symmetry Control): Facial symmetry was derived from the DLib 68-point landmarks (King, 2009) using predefined bilateral landmark pairs. A vertical reflection axis was defined by the midpoint of landmarks 36 and 45. Left-side landmarks were reflected across this axis and compared with their corresponding right-side landmarks. Pairwise Euclidean differences were normalised by the distance between landmarks 36 and 45, and the inverse mean normalised difference formed the raw symmetry score. The score was

subsequently z-standardised, with higher values indicating greater estimated bilateral similarity.

DLib Empirical Audit: Extraction success and failure rates were summarised descriptively by gender × ethnicity group to examine extraction performance before predictive modelling. We also modelled extraction success with a binomial-logit generalised linear model using HC3 robust covariance. The model included controls for gender, five ethnicity indicators, and standardised controls for blur, brightness, and image centring. Female and White were reference groups. The Benjamini–Hochberg correction was applied across the six demographic main-effect tests, but image-quality controls were excluded from that correction family.

3.3.3 Derived Variables (bias metrics)

To evaluate directional bias, the algorithmic residual is defined as:

$$Residual_i = Observed_i - Predicted_i$$

Under this definition, positive residuals indicate that the model underpredicted the observed human rating, whereas negative residuals indicate that the model overpredicted it. Signed residuals describe the direction of prediction error within a subgroup; MAE and RMSE quantify error magnitude without indicating direction. Because subgroup point estimates can differ without being statistically distinguishable, residual, MAE, and RMSE values are interpreted together with uncertainty intervals and prespecified pairwise comparisons.

3.4 Analytical Strategy

The analysis proceeds in four phases, moving from descriptive observation to predictive evaluation.

Phase 1: Descriptive statistics were used to summarise attractiveness scores by gender × ethnicity groups and to characterise DLib landmark-extraction success and failure rates. An additional inferential binomial-logit model with HC3 robust covariance was estimated for DLib extraction success and is shown in Supplementary Table S1. No inferential ANOVA is performed on the descriptive table of the attractiveness score, as the study's primary demographic inference is conducted in the multivariable OLS model.

Phase 2: An intersectional Ordinary Least Squares (OLS) regression was estimated to maintain interpretability. The model regresses attractiveness scores on gender, five ethnicity proxies, and five gender × ethnicity interaction terms. The model controls for morphological symmetry, blur, brightness, and the image-centring proxy. Heteroscedasticity was accounted for using robust HC3 standard errors (MacKinnon & White, 1985). The family of 11 focal demographic tests (5 ethnicity main effects, 1 gender main effect, and 5 gender × ethnicity interaction terms) was corrected using the Benjamini–Hochberg (BH) false discovery rate procedure (Benjamini & Hochberg, 1995). Symmetry and image-quality controls were excluded from this family of corrections because they were adjustment covariates, not focal demographic hypotheses. To test whether the demographic coefficients were sensitive to that control, a sensitivity model without the DLib-derived symmetry measure was estimated. Additionally, a Huber robust-regression specification was estimated as a sensitivity analysis to assess robustness

against large residuals; this procedure reduces the influence of observations with large residuals and does not constrain model predictions to the bounded 1–5 range.

Phase 3: Inter-rater variability. The long-format MEBeauty rater-level dataset, available separately, is unbalanced, with varying numbers of images rated by different subsets of raters. Accordingly, the final analysis relies on observed ratings alone, without imputing missing scores. For each image with at least two observed ratings, the within-image sample standard deviation was computed. We summarised agreement at the group level using the mean within-image standard deviation and a 95% percentile confidence interval derived from 1,000 bootstrap resamples at the image level. Fifty-four rater-level image records were not classifiable into a defined gender $\times$ ethnicity category and were excluded from the main twelve-group presentation, resulting in 2,550 classified images. The standard deviation of image-level ratings by demographic group was regressed using an HC3-robust one-way OLS model as an omnibus test of group differences.

Phase 4: Evaluation across datasets: For this purpose, a convolutional neural network was trained on the SCUT-FBP5500 dataset and tested only on the DLib-successful MEBeauty image sample. We calculated group-wise MAE, RMSE, and signed residuals across the twelve gender $\times$ ethnicity intersections. We used 1,000 bootstrap resamples at the image-ID level to generate uncertainty intervals for Table 5 and for the two prespecified RMSE contrasts shown in Table 6. For each resampled image, all five seed-level predictions were retained, so predictions from different seeds for the same image were not treated as independent observations.

3.5 Ethical Considerations and Data Governance

This secondary analysis used existing datasets that are publicly or research-accessible and did not involve recruitment, intervention, or interaction with individuals. Therefore, no formal ethical review was required according to the authors' institutional policy. No effort was made to identify or re-identify individuals, to infer personality traits, or to perform biometric classification at the individual level. The analysis was conducted for critical and diagnostic research purposes to assess the demographic variation in automated facial-attractiveness assessment. The analysis is based on facial imagery and attractiveness ratings, so the authors are aware of the ethical sensitivity involved in the reuse of facial image datasets. Results are reported primarily at the level of aggregate demographic groups (Buolamwini & Gebru, 2018). Original facial images, names, and personal identifiers are not directly redistributed. Non-identifying image identifiers are employed solely as keys to the derived prediction and error records to allow verification of the reported analyses.

3.6 Software, Model Implementation, and Reproducibility

The reported analyses were implemented in Python using pandas, NumPy, scikit-learn, statsmodels where applicable, OpenCV, DLib, and PyTorch. The predictive model used a PyTorch ResNet-50 architecture (He et al., 2016) initialised with ImageNet-pretrained weights. The final classification layer was replaced with a single linear regression output. Images were resized to 224×224 pixels, converted to tensors, and normalised using the ImageNet channel

means and standard deviations. Random horizontal flipping was applied during training.

The model was optimised using mean squared error loss and Adam with a learning rate of 1×10^{-4} , weight decay of 1×10^{-5} , a batch size of 32, and a maximum of 50 epochs. Early stopping was applied after five epochs without improvement in validation RMSE. SCUT-FBP5500 was partitioned into 80% training, 10% validation, and 10% test sets using the supplied split identifiers.

Models were trained on 5 random initialisation seeds: 42, 101, 202, 303 and 404. MEBeauty was used only as an external validation dataset and was not used for hyperparameter selection, early stopping, threshold selection, or model selection. Group-specific MAE and RMSE were reported across the five model runs. We estimated cross-dataset uncertainty intervals by generating 1,000 image-level cluster bootstrap resamples, selecting all five seed-level predictions for each image.

For the image-based MEBeauty export, there were 1,962 images, with 1,929 passing DLib landmark extraction and 33 failing. Therefore, the OLS and cross-dataset fairness analyses use the DLib-successful analytical sample of $N = 1,929$. The associated reproducibility package includes the synchronised analytical datasets, executable analysis scripts, model checkpoints, prediction records, software environment specifications, figure-generation code, and tabular outputs for the analyses reported in this manuscript.

4. Results

To address the methodological limitations of binary demographic groupings, this section reports empirical findings utilising a fully disaggregated twelve-group (gender $\times$ ethnicity) intersectional framework for the MEBeauty dataset. Findings are presented via an empirical audit of the DLib preprocessing pipeline (Table 1), descriptive statistics (Tables 2A and 2B), an intersectional Ordinary Least Squares (OLS) regression with robust standard errors and image-quality covariates (Table 3), interrater variance (Table 4), and a twelve-group stratified fairness audit (Table 5).

All MEBeauty image-based analyses that require DLib-derived features and cross-dataset predictions use the DLib-successful analytical sample of $N = 1,929$ drawn from an initial 1,962-image export. To control the expected false discovery rate across multiple simultaneous hypothesis tests, the Benjamini–Hochberg (BH) correction was applied. The BH correction family consisted of the five ethnicity main effects, the male main effect, and five gender $\times$ ethnicity interaction terms, for a total of 11 tests. Symmetry and image-quality controls were not included in this correction family because they were included as adjustment covariates rather than focal demographic hypotheses. CNN performance is reported as the average across five random initialisation seeds to ensure reproducibility.

4.1 Preprocessing Bias: DLib Empirical Audit

Before the predictive modelling procedure, an empirical audit of the DLib 68-point landmark detector was performed to quantify representation bias in the preprocessing pipeline (see Table 1). Extraction success and failure rates were described by demographic group. DLib landmark

extraction succeeded for 1,929 of 1,962 MEBeauty images and failed for 33, representing an overall failure rate of 1.7%. The highest observed subgroup failure rate occurred among Middle Eastern men, for whom 2 of 48 images failed extraction (4.2%; 95% CI [1.2, 14.0]), whereas Black women had a failure rate of 0.9% (1/116) and Black men had a failure rate of 0.8% (1/122). Because this subgroup estimate rests on only two failures, its confidence interval is correspondingly wide; the descriptive point estimate does not establish an elevated population failure rate, and the low total number of failures limits the statistical precision of all subgroup comparisons.

Table 1. Empirical audit of DLib landmark-extraction failure by gender $\times$ ethnicity (MEBeauty)

Group	Original N	Successful extraction	Failed extraction	Failure rate (%)	95% CI
Female–Asian	163	159	4	2.5	[1.0, 6.1]
Female–Black	116	115	1	0.9	[0.2, 4.7]
Female–Hispanic	121	118	3	2.5	[0.8, 7.0]
Female–Middle Eastern	112	111	1	0.9	[0.2, 4.9]
Female–South Asian	121	120	1	0.8	[0.1, 4.5]
Female–White	455	446	9	2.0	[1.0, 3.7]
Male–Asian	114	113	1	0.9	[0.2, 4.8]
Male–Black	122	121	1	0.8	[0.1, 4.5]
Male–Hispanic	118	117	1	0.8	[0.1, 4.6]
Male–Middle Eastern	48	46	2	4.2	[1.2, 14.0]
Male–South Asian	122	119	3	2.5	[0.8, 7.0]
Male–White	350	344	6	1.7	[0.8, 3.7]
Total (MEBeauty)	1962	1929	33	1.7	[1.2, 2.4]

Note: Total failure rate is 1.7% (33/1,962). The highest subgroup failure rate was observed among Male–Middle Eastern subjects (4.2%), not among Black subjects. Wilson score 95% CIs are shown.

4.2 Descriptive Statistics

The present study assessed the distribution of mean attractiveness scores across demographic cohorts to characterise observed rating patterns prior to modelling (Table 2A). In the DLib-successful MEBeauty analytical sample, group means ranged from 3.12 for Black men to 3.25 for Black women, Asian women, and Middle Eastern men after rounding. SCUT-FBP5500 group means ranged from 2.87 for Asian men to 3.14 for non-Asian women.

Table 2A. Observed attractiveness ratings by dataset, gender, and ethnicity

Dataset	Group	Original N	Analytical N	Mean score	SD
SCUT-FBP5500	Female–Asian	2000	2000	3.05	0.71
SCUT-FBP5500	Female–Non-Asian	750	750	3.14	0.71
SCUT-FBP5500	Male–Asian	2000	2000	2.87	0.66
SCUT-FBP5500	Male–Non-Asian	750	750	3.01	0.62
MEBeauty	Female–Asian	163	159	3.25	0.51
MEBeauty	Female–Black	116	115	3.25	0.46
MEBeauty	Female–Hispanic	121	118	3.21	0.51
MEBeauty	Female–Middle Eastern	112	111	3.19	0.51
MEBeauty	Female–South Asian	121	120	3.21	0.51
MEBeauty	Female–White	455	446	3.22	0.53
MEBeauty	Male–Asian	114	113	3.16	0.59
MEBeauty	Male–Black	122	121	3.12	0.49
MEBeauty	Male–Hispanic	118	117	3.16	0.55
MEBeauty	Male–Middle Eastern	48	46	3.25	0.47
MEBeauty	Male–South Asian	122	119	3.19	0.53
MEBeauty	Male–White	350	344	3.24	0.52

Note: SCUT means reflect per-image averages computed directly (rating scale 1–5). MEBeauty means and SDs were calculated using the analytical sample after filtering for successful DLib extraction.

Table 2B. Descriptive statistics for raw morphological and image-quality covariates (MEBeauty, N = 1,929)

Covariate	N	Mean	SD	Min	Max
Raw symmetry index	1,929	16.983	14.165	1.759	157.924
Blur	1,929	396.148	463.061	16.182	8,045.014
Brightness	1,929	106.932	34.161	21.861	212.244
Image-centring proxy	1,929	0.786	0.669	0.000258	4.040

Note. Descriptive statistics are for the MEBeauty analytical sample with successful DLib extraction (N = 1,929). Values represent raw (unstandardised) covariates. Regression and logit models reported in Tables 3, S1, S2, and S4 used z-standardised versions of these variables. The image-centring proxy is scaled by a constant factor of 10 as defined in Section 3.3.2 and can exceed 1.

Table 2B summarises the raw distributions of the symmetry and image-quality covariates used as adjustment controls in the multivariable models.

As shown in Figure 1, these descriptive differences indicate modest variation in average attractiveness scores across groups. Still, they are not interpreted as statistically significant “beauty premiums” because no corresponding inferential ANOVA is reported. The results are therefore treated as descriptive summaries rather than as tests of systematic group-level differences.

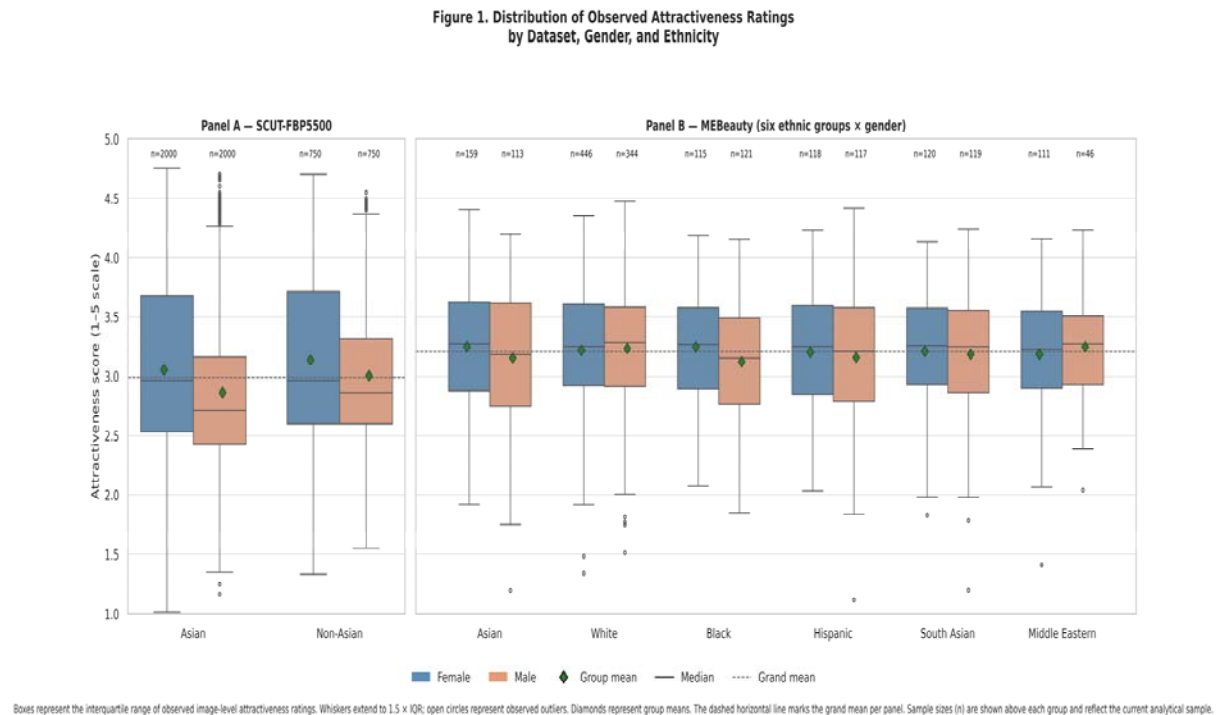


Figure 1. Observed attractiveness ratings by dataset, gender, and ethnicity

4.3 Adjusted Demographic Associations in Attractiveness Ratings

To overcome the shortcomings of a collapsed binary analysis, an intersectional OLS regression was fitted to the MEBeauty dataset (Table 3). The model regresses attractiveness scores on six-group ethnicity indicators, gender and all five gender × ethnicity interaction terms. The specification employs HC3 robust standard errors and controls for morphological symmetry as well as three standardised image-quality covariates: blur, brightness, and the image-centring proxy.

Table 3. Intersectional OLS regression predicting attractiveness scores (MEBeauty, N = 1,929, $R^2 = 0.022$)

Predictor	Coef.	Robust SE	95% CI	Raw p	BH-adj. p
Intercept	3.22	0.03	[3.17, 3.27]	<.001	–
Male	0.01	0.04	[-0.07, 0.08]	.839	.839
Asian	0.03	0.05	[-0.07, 0.12]	.556	.839
Black	0.02	0.05	[-0.08, 0.12]	.700	.839
Hispanic	-0.01	0.05	[-0.12, 0.09]	.803	.839
South Asian	-0.01	0.05	[-0.11, 0.09]	.837	.839
Middle Eastern	-0.04	0.06	[-0.15, 0.07]	.432	.839
Male × Asian	-0.11	0.08	[-0.26, 0.05]	.171	.839
Male × Black	-0.12	0.07	[-0.26, 0.02]	.093	.839
Male × Hispanic	-0.05	0.08	[-0.20, 0.11]	.542	.839
Male × South Asian	-0.04	0.08	[-0.19, 0.11]	.601	.839
Male × Middle Eastern	0.06	0.09	[-0.12, 0.24]	.512	.839
Symmetry index (control)	-0.02	0.01	[-0.05, 0.00]	.053	–
Blur (control)	-0.06	0.02	[-0.10, -0.02]	.006	–
Brightness (control)	0.02	0.01	[-0.00, 0.05]	.062	–
Image centring (control)	-0.02	0.01	[-0.04, 0.00]	.112	–

Note: Reference categories: Female, White. Robust (HC3) SEs. BH = Benjamini–Hochberg correction across 11 focal demographic tests. No demographic main effect or interaction term is significant after correction; only the blur control ($p = .006$) reaches significance.

Using the DLib-successful analytical sample (N = 1,929), the intersectional OLS model explained 2.2% of the variation in image-level attractiveness scores ($R^2 = .022$). After Benjamini–Hochberg correction across the 11 focal demographic tests, no gender main effect, ethnicity main effect, or gender × ethnicity interaction was statistically significant. With Female and White as reference categories, the male coefficient of 0.01 (BH-adjusted $p = .839$) represents the male–female difference within White subjects, and the Black coefficient of 0.02 (BH-adjusted $p = .839$) represents the Black–White difference within women. Of the adjustment covariates, the Laplacian-variance measure labelled blur was statistically significant ($\beta = -.06$, $p = .006$), while symmetry ($p = .053$), brightness ($p = .062$), and image centring ($p = .112$) did not reach the conventional .05 threshold. These results indicate that the tested demographic coefficients were not statistically significant under this specification and correction procedure; they do not establish the absence of demographic differences.

4.4 Interrater Variability: Consensus vs Dissensus

To quantify the degree of rater consensus versus dissensus in the unbalanced MEBeauty rating matrix, Table 4 reports the within-image standard deviation of the ratings actually observed for each image on the original MEBeauty 1–10 scale. Rater scores were not imputed if missing. Smaller values indicate greater agreement among raters who rated a given image; larger values indicate greater dispersion.

Table 4. Interrater variability in observed ratings by gender × ethnicity group (MEBeauty)

Group	Images	Unique raters	Mean ratings/image	Mean within-image SD	95% bootstrap CI
Female–Hispanic	151	308	17.7	1.66	[1.58, 1.73]
Female–White	551	334	22.7	1.69	[1.66, 1.72]
Female–Middle Eastern	147	305	18.2	1.75	[1.68, 1.83]
Female–Asian	199	326	18.3	1.82	[1.75, 1.88]
Female–South Asian	152	301	16.9	1.82	[1.74, 1.89]
Female–Black	151	299	18.0	1.93	[1.85, 2.00]
Male–White	448	360	34.7	1.94	[1.91, 1.98]
Male–Hispanic	149	353	26.4	1.96	[1.89, 2.02]
Male–Middle Eastern	150	353	29.2	2.00	[1.94, 2.06]
Male–Black	148	350	26.7	2.04	[1.97, 2.10]
Male–South Asian	152	348	27.1	2.07	[1.99, 2.16]
Male–Asian	152	348	23.2	2.09	[2.02, 2.17]

Note: Within-image SD is calculated from observed rater scores on the original MEBeauty 1–10 scale; no missing ratings are imputed. The table uses 2,550 image records with defined gender × ethnicity labels; 54 Unknown–Unknown records are excluded. The 95% CI is the percentile interval from 1,000 image-level bootstrap resamples. Lower mean within-image SD indicates greater agreement.

The mean within-image rating SD across the twelve defined groups ranged from 1.66 for Hispanic women to 2.09 for Asian men. An HC3-robust omnibus model revealed variation in image-level rating dispersion by demographic group (Wald $\chi^2(11) = 279.10$, $p < .001$; $R^2 = .097$; $N = 2,550$). These findings indicate a high degree of spread in observed ratings at the group level. Still, the omnibus result by itself does not inform us which individual pairwise group differences are statistically significant, nor why these differences are present.

4.5 Cross-Dataset Prediction Bias

Fairness audit: We applied the trained CNN of SCUT-FBP5500 on the DLib-successful subset of the MEBeauty dataset (Table 5). For reproducibility, the model was trained with five different seeds of random initialisation, and MAE and RMSE are reported as the mean across seeds, with seed-to-seed RMSE standard deviations. The 95% RMSE confidence intervals were obtained by resampling image IDs and keeping all 5 seed-level errors for each sampled image. The cross-dataset audit was consistent with the regression analysis and was performed on the DLib-successful MEBeauty analytical sample (N = 1,929).

Table 5. Twelve-group stratified cross-dataset fairness audit (CNN trained on SCUT, evaluated on MEBeauty)

Group	N	Mean residual	MAE	Mean RMSE	SD across seeds	95% bootstrap CI
Female–Asian	159	0.003	0.395	0.491	0.011	[0.45, 0.53]
Female–Black	115	-0.001	0.380	0.470	0.020	[0.42, 0.52]
Female–Hispanic	118	-0.044	0.385	0.481	0.023	[0.43, 0.53]
Female–Middle Eastern	111	-0.076	0.377	0.470	0.016	[0.41, 0.53]
Female–South Asian	120	-0.026	0.408	0.509	0.014	[0.45, 0.57]
Female–White	446	-0.018	0.386	0.486	0.018	[0.46, 0.52]
Male–Asian	113	-0.107	0.403	0.510	0.018	[0.44, 0.57]
Male–Black	121	-0.093	0.405	0.512	0.025	[0.46, 0.57]
Male–Hispanic	117	-0.090	0.369	0.500	0.022	[0.40, 0.62]
Male–Middle Eastern	46	-0.017	0.379	0.476	0.022	[0.39, 0.55]
Male–South Asian	119	-0.089	0.382	0.500	0.011	[0.42, 0.59]
Male–White	344	-0.008	0.390	0.490	0.014	[0.46, 0.53]

Note: The analytical sample contains 1,929 DLib-successful images; group-wise N values sum to 1,929. RMSE values are averages over five CNN seeds (42, 101, 202, 303, and 404). The 95% CI is based on 1,000 image-level cluster bootstrap resamples that include all five seed-level predictions for each sampled image. Residual = observed – predicted.

Group-specific mean RMSE estimates ranged from .470 to .512. Black men had the highest point estimate (RMSE = .512), closely followed by Asian men (.510) and South Asian women (.509). However, these subgroup point estimates are not, on their own, interpreted as demonstrating statistically significant demographic disparities; they are descriptive summaries that require direct contrasts for inferential interpretation. Signed mean residuals were predominantly negative across the male subgroups, indicating modest average overprediction (predicted scores exceeding observed ratings) under the residual convention observed –

predicted, but the magnitude and direction varied by group.

Table 6. Prespecified pairwise RMSE differences with bootstrap CIs

Comparison	RMSE difference	95% bootstrap CI
Male–Black vs Male–Asian	0.002	[-0.09, 0.09]
Male–Black vs Male–White	0.022	[-0.05, 0.09]

Note: Both prespecified contrasts are non-significant because both 95% image-level bootstrap CIs include zero.

Table 6 shows two prespecified pairwise differences in mean five-seed RMSE, using image-level bootstrap contrasts between selected male subgroups. No pre-specified pairwise RMSE contrast was statistically distinguishable from zero. The Black-male vs Asian-male RMSE difference was 0.002 (95% CI [-0.09, 0.09]). The Black-male vs White-male difference was 0.022 (95% CI [-0.05, 0.09]). Thus, despite the largest point estimate of subgroup RMSE in Table 5 for Black men, the direct bootstrap comparisons in Table 6 do not indicate significantly greater prediction error relative to the two pre-specified male reference groups.

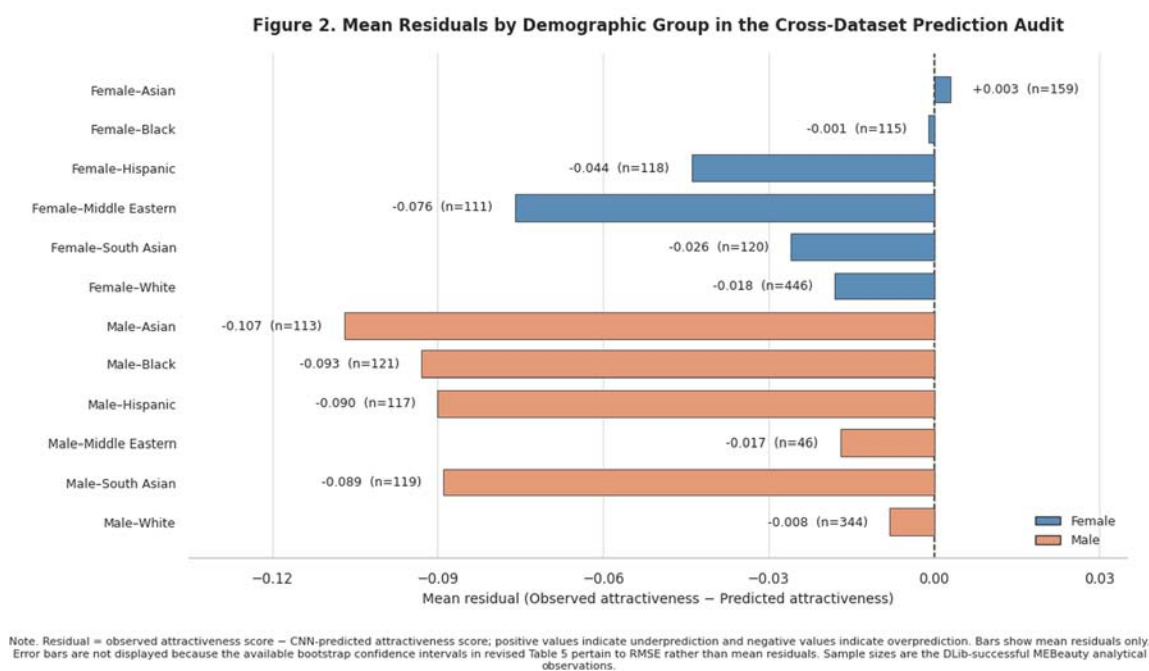


Figure 2. Mean residuals by demographic group in the cross-dataset prediction audit

Figure 2 visualises the mean residuals by demographic group in the cross-dataset prediction audit, with a horizontal reference line at zero. Points or bars represent mean residuals, defined as the observed human rating minus the model prediction; positive values indicate average

underprediction, and negative values indicate average overprediction. Figure 2 is a descriptive summary of signed mean residuals. The bootstrap comparisons in Table 6 concern RMSE differences and do not directly test differences in these signed mean residuals.

5. Discussion

The current analysis provides a qualified view of demographic variation in automated facial-attractiveness assessment. Descriptive differences in human ratings and CNN error metrics are observable across gender $\times$ ethnicity groups in SCUT-FBP5500 and MEBeauty. Still, the symmetry-adjusted OLS model does not identify statistically significant gender, ethnicity, or gender $\times$ ethnicity effects after correction for multiple testing, and the two prespecified pairwise RMSE contrasts are not statistically distinguishable from zero. The DLib audit also shows a comparatively low overall extraction failure rate of 1.7% in MEBeauty, with no concentrated failure pattern for any single demographic group. Taken together, these findings illustrate why fairness audits must distinguish descriptive subgroup variation from inferential evidence of systematic disparity.

5.1 Aesthetic Capital and AI-Mediated Social Evaluation

MEBeauty mean scores vary modestly across gender $\times$ ethnicity groups, with analytic means in the DLib-successful sample clustered in a narrow band from about 3.12 to 3.25. SCUT-FBP5500 shows a slightly larger spread but similarly suggests that average attractiveness ratings differ across binary gender and ethnic categories. However, when attractiveness scores were modelled as a function of gender, six-group ethnicity, their interactions, and measured facial symmetry and image-quality covariates, no demographic main effect or interaction term remained statistically significant after correction for multiple testing. Hence, the present analysis does not provide evidence that a stable male or ethnicity-specific attractiveness penalty remains after correction.

From a sociological perspective, theories of aesthetic capital, cultural hegemony, and intersectionality continue to be useful lenses for examining how hierarchies of attractiveness may be organised and reproduced. However, in this data set, the empirical evidence does not reach the level to confirm those mechanisms as realised demographic penalties after modelling and multiple-testing correction. The theoretical frameworks in Section 2 are therefore motivating, i.e., they indicate what should be scrutinised, rather than describing the effects of the adjusted regression.

Following BH correction, none of the gender $\times$ ethnicity interactions was statistically significant, and the demographic main effects were also nonsignificant. Thus, the current intersectional specification finds no evidence of multiplicative intersectional effects, and the main effects were not statistically significant. This avoids treating a nonsignificant estimate as a proven zero effect. Rather, its contribution is to show that a fully disaggregated framework can test intersectional hypotheses explicitly, and in this case, constrain claims about compounding disadvantage.

The interrater results complement these findings by examining disagreement rather than average attractiveness or prediction accuracy. Mean within-image rating SDs ranged from

1.66 to 2.09 across the twelve groups, and the omnibus test identified group variation in rating dispersion (Wald $\chi^2(11) = 279.10$, $p < .001$). Group differences in disagreement can coexist with nonsignificant adjusted differences in image-level mean ratings: the analyses address different features of the rating process. The omnibus result does not identify which individual group pairs differ.

This distinction also matters for the CNN evaluation. Prediction errors are measured relative to the available human reference scores, which summarise subjective judgments rather than an independently established objective standard. However, the interrater analysis uses 2,550 classified images and the original 1–10 rating scale, whereas the regression and CNN analyses use a separate analytical sample of 1,929 images and harmonised scores. We therefore do not infer that the observed rater disagreement explains the CNN error patterns or the low regression R^2 . Establishing such relationships would require linking image-level rating dispersion to the corresponding scores and prediction errors.

5.2 Algorithmic Normalisation and Cross-Dataset Error

The cross-dataset fairness audit with a CNN trained on SCUT and tested on the DLib-successful MEBeauty sample shows subgroup RMSE point estimates ranging from .470 to .512 and mean residuals of different magnitude and sign for the twelve gender $\times$ ethnicity groups. The RMSE point estimate is greatest for Black men, then Asian men, and then South Asian women. Signed mean residuals are largely negative for male groups, indicating modest average overprediction under the residual convention “observed”-“predicted”.

However, subgroup point estimates alone do not establish statistically significant differences in prediction error. The two prespecified bootstrap contrasts, between Black men and Asian men, and between Black men and White men, both have confidence intervals that include zero, indicating that the observed RMSE differences are not statistically distinguishable from chance variation at the chosen threshold. Similarly, the DLib audit shows a low overall failure rate and no strong concentration of landmark-extraction failures in any subgroup. The subgroup point estimates vary, but the two tested RMSE contrasts remain imprecise; nonsignificance should not be interpreted as evidence that all subgroup differences are small or absent.

The present findings differ from the positive disparity results reported by Hamel and Busch (2025), but the studies examine different statistical targets and pipelines. Hamel and Busch trained ResNet-152 models separately on SCUT-FBP5500 and MEBeauty, used MTCNN face preprocessing, and examined ethnicity-specific prediction and error distributions through Kruskal–Wallis tests and post hoc Dunn comparisons with Benjamini–Hochberg adjustment. The present study uses a five-seed ResNet-50 trained on SCUT-FBP5500, a DLib-successful MEBeauty subset, twelve gender $\times$ ethnicity groups, and two prespecified bootstrap contrasts in RMSE. Its adjusted OLS analysis additionally concerns human scores rather than model-generated score distributions. A distributional disparity can therefore coexist with a nonsignificant difference in the selected RMSE contrasts. These methodological differences offer possible explanations for divergent findings, but their contributions cannot be isolated without matched reanalysis.

Danner et al. (2021) similarly emphasise that strong agreement between predicted and human attractiveness ratings does not resolve the problem of bias in those annotations, and propose AestheticNet and an approach to improving fairness. Their emphasis on model development and bias mitigation complements the present diagnostic audit. Our nonsignificant adjusted rating coefficients and selected RMSE contrasts do not validate the reference ratings as unbiased or demonstrate that the predictor is fair under all criteria. Together, these studies motivate distinguishing label characteristics, prediction distributions, and prediction errors when interpreting fairness results.

Normatively, this pattern highlights the necessity of reporting disaggregated fairness metrics even in the absence of inferential results or when results are inconclusive. Descriptive variation in residuals and RMSE still leaves open questions about the generalisability of models trained on relatively homogeneous datasets to more diverse populations. Yet, the present evidence does not support treating any particular subgroup as empirically confirmed to have substantially greater prediction error.

5.3 Real-World Implications for Platform Ecosystems

While not directly analysing any production platform or measuring downstream outcomes (e.g., visibility, employment, monetisation, engagement), the analysis has implications for hypothetical or future uses of attractiveness scoring in systems such as ranking algorithms, recommendation platforms, or automated assessments. Previous work on engagement-optimised ranking and platform dynamics has indicated that visual presentation and perceived attractiveness can influence exposure and reward structures. Still, it has not considered rating distributions and model errors in research datasets.

If consequential systems include facial-attractiveness components, it would be imperative to conduct disaggregated auditing of training data and model performance, regardless of the statistical significance of the adjusted demographic effects. The relatively small but diverse residual patterns presented here illustrate how model behaviour can differ across subgroups even when conventional significance thresholds are not met. In these circumstances, it is fair to anticipate that governance frameworks should be subject to precautionary scrutiny, especially when training datasets embody narrow or culturally skewed ideas of attractiveness that may interact with other types of biases in complex platform environments.

5.4 Policy and Digital Governance

Three cautious implications can be drawn from the current findings in a sociotechnical governance perspective. First, for appearance-related systems, fairness audits should report both ratings and prediction errors at the fully disaggregated gender $\times$ ethnicity level rather than binary or aggregated categories that can mask local variation. Even when the inferential results are null, the disaggregated diagnostics help to identify where descriptive differences merit further scrutiny or focused robustness checks.

Second, audits should treat face detection and landmark extraction steps in the preprocessing pipeline as part of the model rather than as neutral upstream steps. MEBBeauty analysis shows that failures to extract landmarks are rare, but not uniform. Even low overall failure rates merit

reporting of extraction performance by subgroup before any deployment decisions.

Third, for models deployed in socially consequential domains, calibration and sensitivity analyses should explicitly examine if subgroup error patterns are robust to alternative specifications, datasets, and evaluation protocols. This study shows how sensitivity analyses, such as the symmetry-excluded OLS model and the Huber robust regression, can support the null results for the focal demographic coefficients or qualify the main findings. These designs can serve as a resource for governance frameworks in which claims about fairness must be supported by reproducible, multi-specification evidence rather than single-model point estimates.

5.5 Limitations and Future Research

These results should be interpreted with several limitations in mind. First, any attempt to quantify physical attractiveness risks participating in the same processes of bodily commodification and value assignment that sociological critiques seek to interrogate. This paper treats attractiveness scores as objects of analysis rather than as appropriate decision signals, but the normative tension remains.

Second, neither SCUT-FBP5500 nor MEBeauty have direct measures of socioeconomic status, such as proxies for income, education, or occupation, which restricts the ability to investigate how class intersects with race, gender, and attractiveness in these datasets. Third, subgroup sizes are unequal. The image export contains 48 Middle Eastern men, of whom 46 remain in the analytical sample, compared with 446 White women and 344 White men in that sample. Several other analytical groups contain approximately 110–160 images. Estimates for smaller groups are less precise, and nonsignificant contrasts involving these groups should not be interpreted as evidence that meaningful differences are absent.

Fourth, the cross-dataset fairness audit and the regression analyses are performed on the DLib-successful MEBeauty sample, conditioned on the successful landmark extraction. The image-centring covariate was taken from the stored analytical dataset, and its scaling was not independently reconstructed from the source images; it should therefore be read as a stored image-level centring measure rather than a re-derived quantity. This design guarantees the same analytical sample across modelling steps but does not directly investigate how prediction performance would behave on images that fail preprocessing. Although extraction failures occurred in only 33 of 1,962 images, predictive performance on those excluded images was not evaluated. The reported regression and CNN findings therefore apply to the DLib-successful sample and do not establish equivalent performance or the absence of disparities among failed cases.

The rating scales also constrain interpretation. Linear rescaling changes measurement units and, by itself, does not remove information or demonstrate scale compression. However, image-level aggregated ratings do not retain the full variation in individual judgments. The narrow range of demographic group means should not be confused with the range of individual image scores. Differences in rating procedures and score distributions across datasets may affect cross-dataset interpretation, and nominal scale harmonisation does not

establish that ratings have identical meanings across datasets.

The OLS model has limited explanatory power: the included demographic, symmetry, and image-quality variables account for 2.2% of the observed variation in image-level attractiveness scores. The analysis therefore provides a limited account of the factors associated with these ratings. Low explanatory power is distinct from the uncertainty of individual demographic coefficients, and neither low R^2 nor nonsignificance establishes demographic equivalence. Interpretations should remain confined to the estimated associations and confidence intervals under the specified model.

Finally, the study does not independently verify coordinate-level landmark accuracy across skin tones or facial configurations, nor does it generalise the cross-dataset evaluation to additional intersectional datasets such as FairFace or to domains where SES information is present. Future work should combine attractiveness audits with direct tests of landmark detection accuracy, expand the range of datasets and demographic configurations evaluated, and incorporate richer covariate sets including socio-economic variables and contextual information about image production where possible.

Unequal subgroup sizes and the uncertainty of the estimated contrasts limit statistical sensitivity. The 95% confidence interval for the Black-male versus Asian-male RMSE difference ranges from -0.09 to 0.09 , and the interval for Black men versus White men ranges from -0.05 to 0.09 . These intervals include zero and nonzero differences in either direction, so they do not demonstrate equivalence. Moreover, the two prespecified comparisons do not test all possible differences among the twelve groups. In the OLS analysis, the eleven focal demographic tests are subject to false-discovery-rate correction, and several interaction estimates have relatively wide confidence intervals. The study therefore cannot exclude smaller demographic associations or error differences that the available samples and analysis may not detect. Future studies should define substantively meaningful differences in advance and plan subgroup sample sizes around those targets.

6. Conclusion

This study explored automated facial beauty prediction systems based on two publicly available datasets, SCUT-FBP5500 and MEBeauty, using a combination of disaggregated descriptive statistics, symmetry-adjusted regression, observed-rating inter-rater-variability analysis, and a cross-dataset fairness audit. The analyses identify demographic variation in interrater dispersion and descriptive variation in mean ratings and model errors. Adjusted demographic coefficients in the OLS model were not statistically significant after multiple-testing correction, and the two prespecified pairwise RMSE contrasts were not statistically distinguishable from zero. These findings concern different outcomes and do not establish demographic equivalence.

The lack of statistically supported adjusted demographic penalties in this empirical setting does not lessen the significance of disaggregated sociotechnical auditing. Instead, it emphasises that fairness assessments should be able to detect disparities when they are present, but also to limit strong claims when they are not. The study offers a methodological template for the study of

appearance-related algorithms that combines intersectional modelling with explicit sensitivity analyses and a documented preprocessing audit without overextending the empirical claims.

The work does not directly evaluate real-world platforms or outcomes, nor does it claim that facial-attractiveness scoring is currently deployed in hiring, social media, or creator economy systems. However, given the potential sensitivity of such systems to cultural hegemony encoded in training data, the findings support calls for rigorous, disaggregated auditing of any future appearance-related components in consequential applications. Future work should continue to forge links between computational fairness diagnostics and sociological theory across various datasets and contexts to deepen both empirical and normative understanding of the intersection of aesthetic hierarchies and algorithmic governance.

Data Availability Statement

This study uses secondary data from the publicly available, research-accessible SCUT-FBP5500 and MEBeauty datasets. The SCUT-FBP5500 source materials are available from the Human Computer Intelligent Interaction Laboratory release repository at <https://github.com/HCIILAB/SCUT-FBP5500-Database-Release>. MEBeauty source materials are available through the MEBeauty provider repository at <https://github.com/fbplab/MEBeauty-database> and the research-access Hugging Face distribution at <https://huggingface.co/datasets/dr-irina-lebedeva/MEBeauty>, subject to the respective provider terms of use. Original third-party facial image files are not duplicated in the supplementary reproducibility package; the source images are available directly from these providers.

The supplementary reproducibility package contains the analysis-ready non-image datasets used in the reported analyses; the MEBeauty export manifest; SCUT-FBP5500 train/validation/test image identifiers; variable definitions and coding information; transformation and statistical-method documentation; preprocessing and statistical-analysis scripts; the canonical PyTorch ResNet-50 training and external-evaluation code; CSV outputs corresponding to Tables 1–6 (including the raw covariate descriptive statistics reported in Table 2B) and Supplementary Tables S1–S4; figure-generation code; environment specifications; an analysis crosswalk and run documentation; five trained model checkpoints for random seeds 42, 101, 202, 303, and 404; and the corresponding MEBeauty prediction/error records.

The image-based MEBeauty analysis uses an exported sample of 1,962 images. DLib landmark extraction succeeded for 1,929 images and failed for 33 images; analyses requiring DLib-derived facial features and the cross-dataset prediction audit therefore use the 1,929-image analytical sample. The separately available rater-level analysis is documented independently and uses only observed ratings without imputing missing scores. Derived prediction/error records are keyed only to non-identifying image identifiers and are supplied solely to permit verification of the reported analyses. The authors do not redistribute original facial images, names, and direct personal identifiers.

Ethical and Legal Declarations

Funding Declaration

The authors received no specific funding for this work. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

None.

Research ethics statement

Under the authors' institutional policy, formal ethical review was not required for this secondary analysis because it used pre-existing, publicly/research-accessible datasets and involved no recruitment, intervention, or interaction with individuals. No attempt was made to identify or re-identify individuals, infer personality characteristics, or perform individual-level biometric classification. The analysis was conducted for critical and diagnostic research purposes to evaluate demographic variation in automated facial-attractiveness assessment. The SCUT-FBP5500 and MEBeauty datasets were used in accordance with the research access and use terms specified by their respective providers. Because the analysis involves facial imagery and attractiveness ratings, the authors recognise the ethical sensitivity of repurposing facial image datasets. Findings are therefore reported principally at the aggregate demographic-group level. Original facial images, names, and direct personal identifiers are not redistributed in the reproducibility package. Derived prediction and error records keyed only to non-identifying image identifiers are provided solely to enable editors and reviewers to verify the reported analyses.

Informed consent statement

No new participants were recruited, and no new human-subject data were collected by the authors for this study. The analysis uses pre-existing secondary research datasets; therefore, no additional informed consent was obtained by the authors for the present secondary analysis.

Author contributions

Conceptualisation, A.S.B.; methodology, A.S.B. and H.Y.S.; formal analysis, U.Y. and H.Y.S.; data curation, U.Y. and H.Y.S.; writing—original draft preparation, A.S.B.; writing—review and editing, U.Y. and A.H.A.; supervision and project administration, A.H.A. All authors have read and approved the final manuscript.

Conflict of interest

The authors declare no conflict of interest.

Use of AI/LLM tools

Use of AI/LLM tools: Machine-learning tools were used in the study's methodology for model training and cross-dataset evaluation. Large Language Model tools were used for language editing and formatting support. No Large Language Model, artificial intelligence tool, or

machine-learning system is listed as an author.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CI	Confidence Interval
SD	Standard Deviation
IRB	Institutional Review Board
MAE	Mean Absolute Error
MEBeauty	Multi-Ethnic Beauty Dataset
MSE	Mean Squared Error
OLS	Ordinary Least Squares
RMSE	Root Mean Squared Error
SCUT-FBP5500	South China University of Technology Facial Beauty Prediction 5500 Dataset
SES	Socioeconomic Status

References

- Abid, S., Liaquat, N., & Malik, A. A. (2021). On being and becoming beautiful: The social construction of feminine beauty. *Pakistan Social Sciences Review*, 5(2), 403-413.
- Barnes, S. (2025). The changing nature of beauty, enhancement and harm. In *The harms of beauty* (pp. 29-50). Bristol University Press.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Berry, B. (2016). *The power of looks: Social stratification of physical appearance*. Routledge.
- Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, 27(10), 947-960. <https://doi.org/10.1016/j.tics.2023.06.008>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Danner, M., Weber, T., Peng, L., Gerlach, T., Su, X., & Rätsch, M. (2021). Ethically aligned

deep learning: Unbiased facial aesthetic prediction. arXiv.
<https://doi.org/10.48550/arXiv.2111.05149>

Dimitrov, D., & Kroumpouzos, G. (2023). Beauty perception: A historical and contemporary review. *Clinics in Dermatology*, 41(1), 33-40.
<https://doi.org/10.1016/j.clindermatol.2023.02.006>

Francis, B. (Ed.). (2006). *The nature of gender*. SAGE Publications.

Hamel, C., & Busch, M. (2025). Analysis of bias in deep learning facial beauty regressors. arXiv. <https://doi.org/10.48550/arXiv.2509.24138>

Hanquinet, L., Roose, H., & Savage, M. (2014). The eyes of the beholder: Aesthetic preferences and the remaking of cultural capital. *Sociology*, 48(1), 111-132.
<https://doi.org/10.1177/0038038513477935>

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>

Keyser-Verreault, A., & Rail, G. (2023). Gender, beauty and the future of neoliberalism: Aesthetic labour and women's (anti) aspirationalism in Taiwan. *Chinese Sociological Review*, 55(5), 564-590. <https://doi.org/10.1080/21620555.2023.2189577>

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10, 1755–1758.

Kramer, R. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLOS ONE*, 13(8), e0202655. <https://doi.org/10.1371/journal.pone.0202655>

Kuipers, G. (2015). Beauty and distinction? The evaluation of appearance and cultural capital in five European countries. *Poetics*, 53, 38-51. <https://doi.org/10.1016/j.poetic.2015.10.001>

Lakoff, R., & Scherr, R. (2022). *Face value: The politics of beauty*. Routledge.

Lebedeva, I., Guo, Y., & Ying, F. (2022). MEBeauty: A multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, 34, 14169–14183.
<https://doi.org/10.1007/s00521-021-06535-0>

Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018). SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 1598–1603). IEEE.
<https://doi.org/10.1109/ICPR.2018.8546038>

Liebelt, C. (2018). Beauty and the norm: An introduction. In *Beauty and the norm: Debating standardisation in bodily appearance* (pp. 1-19). Springer.

Lipton, B. (2021). Academics' dress: Gender and aesthetic labour in the Australian university. *Higher Education Research & Development*, 40(4), 767-780.
<https://doi.org/10.1080/07294360.2020.1773767>

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305–325. [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)

Mascia-Lees, F. E. (2011). *A companion to the anthropology of the body and embodiment*. John Wiley & Sons.

Mears, A. (2014). Aesthetic labour for the sociologies of work, gender, and beauty. *Sociology Compass*, 8(12), 1330-1343. <https://doi.org/10.1111/soc4.12211>

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), pgaf062. <https://doi.org/10.1093/pnasnexus/pgaf062>

Parameswaran, R., & Cardoza, K. (2009). Melanin on the margins: Advertising and the cultural politics of fair/light/white beauty in India. *Journalism & Communication Monographs*, 11(3), 213-274. <https://doi.org/10.1177/152263790901100302>

Pawlowski, M. D. (2023). The posthuman beauty myth [Master's thesis, University of Canterbury].

Strzalko, J., & Kaszycka, K. A. (1992). Physical attractiveness: Interpersonal and intrapersonal variability of assessments. *Social Biology*, 39(1-2), 170-176. <https://doi.org/10.1080/19485565.1992.9988813>

Supplementary Materials and Appendices

Appendix A. Facial Symmetry Computation

Facial symmetry was estimated from the 68-point facial-landmark configuration returned by the DLib shape predictor. The computation used predefined bilateral landmark pairs representing facial regions such as the eyes, eyebrows, mouth, and jawline.

A vertical reflection axis was defined using the horizontal midpoint of DLib landmarks 36 and 45. For each predefined bilateral pair, the left-side landmark was reflected across this vertical axis and compared with the corresponding right-side landmark. Pairwise Euclidean differences were normalised by the distance between landmarks 36 and 45 to account for facial scale. The raw symmetry score was defined as the inverse of the mean normalised bilateral distance:

$$S_i = \left[(1/K) \sum_{k=1}^K (\|R_{\text{mid}}(L_{ik}) - R_{ik}\|_2 / d_{i,36,45}) \right]^{-1}$$

where L_{ik} and R_{ik} denote the left and right landmarks in bilateral pair k , $R_{\text{mid}}(L_{ik})$ denotes the reflection of the left landmark across the vertical axis defined by the midpoint of landmarks 36 and 45, $d_{i,36,45}$ is the interocular reference distance, and K is the number of bilateral landmark pairs. The resulting raw score was z-standardised before inclusion in the regression model, with higher values indicating greater estimated bilateral similarity.

This procedure does not perform rotational facial alignment and does not estimate a separate anatomical nasal midline. The symmetry measure should therefore be interpreted as a landmark-based bilateral similarity index generated by the specified reflection procedure.

Table S1. Inferential logistic model of DLib landmark-extraction success

Predictor	Odds ratio	Robust SE (log-odds)	95% CI for OR	Raw p	BH-adj. p
Male	1.02	0.36	[0.51, 2.07]	0.948	0.948
Asian	1.13	0.55	[0.38, 3.34]	0.830	0.948
Black	2.28	0.77	[0.50, 10.36]	0.285	0.948
Hispanic	1.17	0.61	[0.36, 3.87]	0.792	0.948
South Asian	1.1	0.58	[0.36, 3.41]	0.865	0.948
Middle Eastern	0.92	0.65	[0.25, 3.31]	0.896	0.948
Blur (control)	0.81	0.11	[0.66, 1.00]	0.051	–
Brightness (control)	0.85	0.21	[0.57, 1.28]	0.446	–
Image centring (control)	0.53	0.11	[0.43, 0.67]	<.001	–

Table S2. Sensitivity analysis excluding the DLib-derived symmetry measure

Predictor	Coef. (no symmetry)	Robust SE	95% CI	Raw p
Male (vs Female)	0.01	0.04	[–0.06, 0.08]	.783
Black	0.02	0.05	[–0.08, 0.12]	.658

Table S3. CNN training, validation, and test performance across five seeds

Seed	Best epoch	Validation MAE	Validation RMSE	SCUT test MAE	SCUT test RMSE
42	22	0.218	0.290	0.206	0.268
101	21	0.221	0.292	0.203	0.261
202	22	0.227	0.298	0.219	0.282
303	15	0.224	0.299	0.219	0.282
404	10	0.230	0.308	0.221	0.288

Table S4. Sensitivity of demographic coefficients to Huber robust regression

Predictor	Robust-regression coefficient	Standard error	95% CI	Raw p
Male (vs. Female)	0.00	0.04	[–0.07, 0.07]	.996
Black	–0.01	0.06	[–0.11, 0.10]	.913

Supplementary interpretation

Taken together, the supplementary analyses support a cautious interpretation of the main results. The no-symmetry OLS model and the Huber robust-regression specification do not materially change the focal demographic estimates. The CNN seed-level results document reproducible model performance across five initialisations. In the inferential DLib-success model (Table S1), no demographic main effect was statistically significant after Benjamini–Hochberg correction; the standardised image-centring control was associated with lower odds of successful extraction ($OR = 0.53$, $p < .001$). Given the small total number of extraction failures, demographic odds-ratio estimates remain imprecise and should be interpreted cautiously.

The supplementary analyses also reinforce the distinction between descriptive subgroup variation and statistically supported disparity. Differences in subgroup means, residuals, or odds ratios can motivate further investigation, but they should not be described as demographic penalties, exclusion effects, or fairness failures without corresponding inferential support and a clearly defined estimate.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).