Original Research Article

Comparative Analysis of DeepFace Attribute Classification on the FairFace Validation Set

Jiaming Qian

Diamond Bar High School, 21400 Pathfinder Rd, Diamond Bar, CA 91765, United States

ABSTRACT

Automated facial analysis systems increasingly estimate demographic attributes such as age, gender, and race from images, yet performance disparities across demographic groups remain a substantial concern. This study evaluates the off-the-shelf DeepFace attribute classifier, accessed through the deepface Python library, on the FairFace validation set. FairFace labels are treated as ground truth, and DeepFace is applied without additional fine-tuning. To align the label spaces, FairFace's East Asian and Southeast Asian categories are merged into a single Asian class to match DeepFace's race output. Performance is assessed using age-range accuracy, mean signed age error (age bias), overall accuracy, balanced accuracy, per-class precision, recall, and F1 score (F1), macro-averaged F1, Cohen's kappa coefficient (κ), and chi-square tests of race-related error disparities. On 10,954 validation images, DeepFace achieved age-range accuracy of 0.28437 with a positive age bias of 3.65926 years, gender accuracy of 0.72074, and race accuracy of 0.59540. Gender results showed strong asymmetry between female recall (0.44537) and male recall (0.96616), while race results showed substantially lower F1 scores for Indian, Latino/Hispanic, and Middle Eastern groups than for Asian and Black groups. These findings show that strong face-verification performance does not necessarily translate into equitable demographic attribute prediction and underscore the need for subgroup-level evaluation and fairness-aware model development.

Keywords: DeepFace; FairFace; facial attribute classification; algorithmic bias; demographic parity; fairness evaluation; computer vision

INTRODUCTION

Deep convolutional neural networks have dramatically advanced face recognition, reaching and even surpassing human performance on popular benchmarks (1, 2). A prominent example is DeepFace, which achieved 97.35% verification accuracy on the Labeled Faces in the Wild dataset and effectively closed most of the remaining gaps to human-level performance in unconstrained face verification (1). Building on such successes, several software frameworks wrap face recognition backbones with additional heads for facial attribute analysis, enabling prediction of attributes such as age, gender, and race alongside identity (3, 4). The deepface Python library is one such framework, providing a convenient interface around models including VGG-Face, DeepFace, ArcFace, and others for both recognition and attribute

Corresponding author: Jiaming Qian, E-mail: jiamingqian420@gmail.com.

Accepted April 10, 2026
https://doi.org/10.70251/HYJR2348.42364373

estimation (3).

At the same time, there is increasing evidence that facial analysis systems do not perform equally well for all demographic groups. Studies auditing commercial gender classification APIs have revealed large gaps between the error rates for light-skinned men and darker-skinned women, in some cases exceeding 30 percentage points in difference (5, 6). Similar disparities have been reported for face recognition systems more broadly, in which error rates on non-White faces were dramatically higher than on White faces (7, 8). These patterns are often traced back to skewed training datasets in which lighter-skinned and male subjects are overrepresented, and to models that are optimized for overall accuracy without explicit fairness constraints (9-11).

To address the data imbalance problem in face datasets, the FairFace dataset was introduced as a large-scale collection of more than 100,000 face images balanced across seven race groups, two genders, and age intervals (10). Models trained on FairFace have been shown to generalize better to diverse populations and to exhibit more uniform error rates across race and gender groups than models trained on traditional, highly skewed datasets such as VGGFace2 or MS-Celeb-1M (2, 10, 12). In parallel, fairness-aware modeling approaches such as InclusiveFaceNet learn explicit race and gender representations and then transfer them to downstream attribute detectors, improving accuracy on minority subgroups while preserving or improving average performance (9, 11, 13).

This study focuses on evaluating the attribute classification performance of the DeepFace model, as exposed through the deepface library's "analyze ()" function, when applied to the FairFace validation set. Unlike the FairFace ResNet-34 model, which is trained directly on FairFace for attribute prediction (10), DeepFace was originally trained for face verification on large web-scale datasets and only later adapted for age, gender, and race prediction (1-3). Evaluating DeepFace on FairFace therefore provides insight into how well a general-purpose face recognition embedding transfers to attribute tasks on a demographically balanced dataset without further fine-tuning.

Beyond simple accuracy, rigorous assessment of such systems requires appropriate statistical metrics and hypothesis tests. For multi-class problems like race classification, per-class precision, recall, and F1 scores, macro-averaged F1, balanced accuracy, and confusion matrices are informative for understanding which groups are favored or disfavored by the model (14, 15). Agreement coefficients such as Cohen's kappa quantify the extent to which a model's predictions exceed chance-level agreement with ground truth (16). In fairness research, chi-square tests and related methods are often used to test whether error rates differ significantly across groups (10, 17).

This study has three objectives. First, it evaluates DeepFace on the FairFace validation set using a harmonized label space for age, gender, and race. Second, it quantifies performance using age-range accuracy, age bias, overall accuracy, balanced accuracy, per-class precision, recall, F1 score, Cohen's kappa, and chi-square tests of race-related error disparities. Third, it interprets the observed subgroup differences in the context of fairness-aware facial attribute modeling.

METHODS AND MATERIALS

Dataset and Ground Truth

The experiments use the FairFace face attribute dataset, which was designed to mitigate race bias by balancing images across seven race categories, two genders, and age groups (10). The full dataset contains 108,501 images collected from the YFCC-100M Flickr corpus and annotated by crowd workers with perceived race, gender, and age-group labels (10). In this study, only the official validation split is used. Because DeepFace outputs a single Asian race label, East Asian and Southeast Asian ground-truth labels are merged into one Asian category for evaluation; race results are therefore reported with macro-averaged metrics and balanced accuracy in addition to overall accuracy. The validation spreadsheet provides one row per image with the file name, race, gender, and age group, and these labels are treated as ground truth.

Sample Data

Table 1 illustrates the format of the ground-truth annotations used in evaluation. These labels constitute the reference targets against which DeepFace age, gender, and race predictions are compared throughout the analysis.

Figure 1 provides two examples of the image quality and demographic variation present in the validation set, including differences in age range, race label, and facial appearance under real-world lighting conditions.

Table 1. Sample ground-truth annotations from the FairFace validation split used in this study. Each row corresponds to one validation image and reports the file name, age group, gender label, and race label used as reference targets for evaluation.

File name	Age group	Gender	Race
val/1.jpg	3–9	Male	East Asian
val/2.jpg	50–59	Female	East Asian
val/3.jpg	30–39	Male	White
val/4.jpg	20–29	Female	Latino/Hispanic
val/5.jpg	20–29	Male	Southeast Asian
val/6.jpg	30–39	Male	Latino/Hispanic
val/7.jpg	20–29	Male	Black
val/8.jpg	3–9	Male	East Asian
val/9.jpg	20–29	Male	Southeast Asian
val/10.jpg	3–9	Male	Southeast Asian

Figure 1. *Representative images from the FairFace validation split used for qualitative illustration in this study. The left image is labeled 3–9 years, male, East Asian; the right image is labeled 50–59 years, male, Latino/Hispanic, according to the FairFace annotations.*

DeepFace Attribute Prediction Pipeline

DeepFace attribute predictions were generated using the deepface Python library, version 0.0.95 (18), which provides a unified interface to multiple face-recognition backbones and attribute-analysis heads (3). In this study, the analyze function was called with actions=['age', 'gender', 'race'] for each image in the FairFace validation folder. The library performs face detection, alignment, and attribute inference using the standard detect–align–normalize–represent pipeline employed in modern face recognition systems (1-3).

For each successfully processed image, DeepFace returns a predicted age, a dominant gender label, and a dominant race label. Gender outputs were mapped by converting 'Man' to 'male' and 'Woman' to 'female'. Because DeepFace provides a single Asian class, FairFace East Asian and Southeast Asian ground-truth labels were merged into one Asian class to enable consistent multi-class evaluation. This mapping is necessary for comparability, but it reduces demographic granularity and may obscure performance differences between East Asian and Southeast Asian subgroups; this limitation is discussed further in Section [Limitations] (3, 10). After harmonization, the DeepFace predictions were merged with the FairFace validation table by file name to create the final analysis dataset.

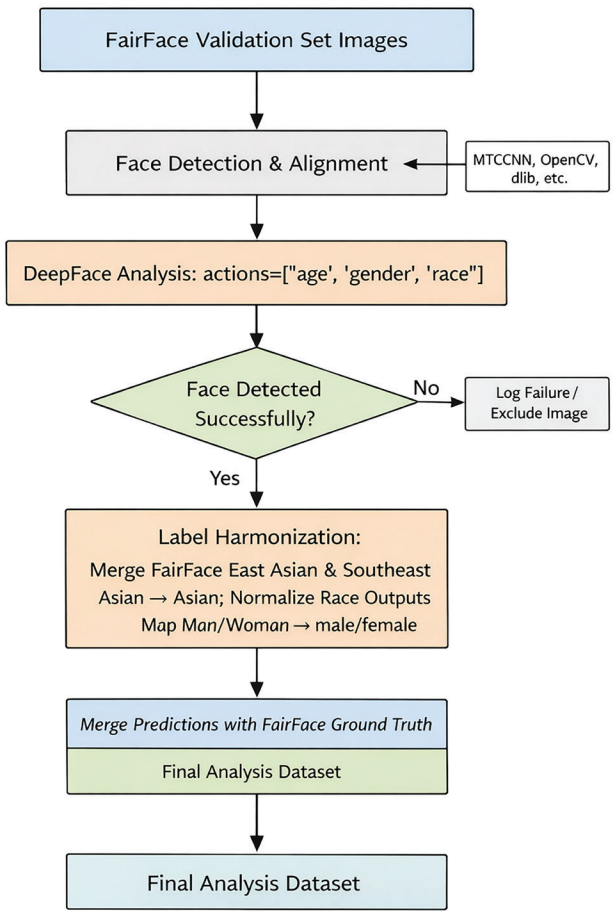

Figure 2. *Workflow of the DeepFace attribute-prediction and label-harmonization pipeline applied to the FairFace validation set, including image input, face detection and alignment, attribute prediction, label normalization, race-category harmonization, and merger with ground-truth annotations for statistical analysis.*

Statistical Models and Performance Metrics

To characterize DeepFace performance in a way that is useful for both machine learning and statistical inference, a set of complementary metrics is defined. These include age-range accuracy and age bias, overall accuracy, balanced accuracy, per-class precision, recall, and F1 score (F1), macro-averaged F1, Cohen's kappa coefficient (κ), and chi-square tests of race-related error disparities. The notation below assumes a dataset of images with ground-truth labels and model predictions; contingency counts such as true positives, false positives, and false negatives are used to derive the reported metrics (14,15).

Age-Range Accuracy and Age Bias

FairFace provides age labels as discrete intervals R_i (e.g., 3–9, 10–19, 20–29). Let denote the numerical age predicted by DeepFace for image i, and let $\mathbb{1}(\cdot)$ be the indicator function. Age-range accuracy counts a prediction as correct when the predicted age falls inside the ground-truth interval. It is defined as:

$$\text{AgeAccuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\hat{a}_i \in R_i\right)$$

Because the validation labels are age ranges, it is also informative to quantify systematic over- or underestimation. To this end, each age interval R_i is mapped to its midpoint m_i (for example, 3–9 becomes 6). The signed age error for image i is $\hat{a}_i - m_i$, and the mean signed error (age bias) is:

$$\text{AgeBias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{a}_i - m_i)$$

A negative AgeBias indicates a tendency to underestimate age, whereas a positive value indicates overestimation. These two measures together provide a concise summary of age prediction performance, combining a categorical notion of correctness with a continuous notion of bias (4, 19).

Accuracy and Balanced Accuracy

For each attribute treated as a classification problem (age group, gender, race), overall accuracy is defined as the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\hat{y}_i = y_i\right)$$

In multi-class settings with potentially imbalanced class frequencies, accuracy can obscure poor performance on less common classes. Balanced accuracy addresses this by averaging the recall (true positive rate) over all classes (20). Let C be the set of classes and let TP_c and FN_c denote the number of true positives and false negatives for class c, respectively. The recall for class c is:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

Balanced accuracy is then defined as:

$$\text{BalancedAccuracy} = \frac{1}{|C|}\sum_{c\in C} \text{Recall}_c$$

Because FairFace is approximately balanced across race groups by design, balanced accuracy and standard accuracy are expected to be similar for race, but balanced accuracy still serves as a useful fairness-oriented summary and is more robust to any residual imbalance (10, 17).

Per-Class Precision, Recall, F1, and Macro-F1

To analyze which classes DeepFace predicts well or poorly, per-class precision, recall, and F1 scores are computed. For a fixed class c, precision and recall are defined in terms of true positives (TP_c), false positives (FP_c), and false negatives (FN_c):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

The F1 score for class is the harmonic mean of precision and recall:

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

To summarize performance across all classes, the macro-averaged F1 score is used:

$$\text{MacroF1} = \frac{1}{|C|}\sum_{c\in C} \text{F1}_c$$

Macro-F1 weights all classes equally and is particularly informative for the race classification task, where equal performance across demographic groups is desirable from a fairness perspective (5, 10, 13).

Cohen's Kappa Coefficient

Cohen's kappa (κ) quantifies the agreement between DeepFace predictions and ground truth beyond what would be expected by chance (16, 21). Let P_o denote the observed agreement (i.e., accuracy) and P_e the expected agreement under random guessing based on the empirical marginal label distributions. Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Values of κ near 1 indicate almost perfect agreement, values near 0 indicate agreement close to chance, and negative values indicate systematic disagreement (16).

Chi-Square Tests for Race Parity

To assess whether DeepFace error rates differ significantly across race groups, chi-square tests of independence are performed on contingency tables derived from the confusion matrix (15). For example, to test parity of overall correctness across races, a $2 \times K$ table is constructed where rows correspond to correct vs. incorrect predictions and columns correspond to the K race groups. Let O_{jk} be the observed count in row j and column k, and E_{jk} the expected count under the null hypothesis of independence, computed as:

$$E_{jk} = \frac{(\text{row total})_j \cdot (\text{column total})_k}{\text{grand total}}$$

The chi-square statistics are then:

$$\chi^2 = \sum_j \sum_k \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

In addition to the chi-square statistic, Cramér's V is reported as an effect-size measure for the strength of association in the contingency table (22).

$$V = \sqrt{\chi^2 / N}$$

A small p-value suggests that prediction correctness is not evenly distributed across race groups. Similar tests can also be applied to false-positive or false-negative patterns when more specific disparity analyses are required (10, 17).

RESULTS

This section reports the empirical performance of DeepFace on the harmonized FairFace validation set. Results are presented in four parts: descriptive statistics, age-prediction results, gender-classification results, and race-classification results.

Descriptive Overview

The merged DeepFace–FairFace evaluation set contains 10,954 validation images with aligned ground-truth labels and DeepFace attribute predictions. Table 2 summarizes the demographic composition of this evaluation set after label harmonization. Because the DeepFace race head outputs a single Asian category, FairFace East Asian and Southeast Asian labels were merged into one consolidated Asian group for race evaluation.

The consolidated Asian category is the largest group (2,965 images), followed by White (2,085), Latino/Hispanic (1,623), Black (1,556), Indian (1,516), and Middle Eastern (1,209). Because this six-class mapping is not perfectly balanced, later sections emphasize balanced accuracy and macro-averaged F1 alongside overall accuracy when interpreting race performance.

The gender distribution is relatively even, with 5,792 images labeled male and 5,162 labeled female. The age-group distribution is more uneven, with the 20–29 bin being the largest (3,300) and the oldest categories

Table 2. *Distribution of images by race, gender, and age group in the harmonized FairFace validation set (n = 10,954).*

Attribute	Category	Count
Race	Asian	2,965
	White	2,085
	Latino/Hispanic	1,623
	Black	1,556
	Indian	1,516
	Middle Eastern	1,209
	Total	10,954
Gender	Male	5,792
	Female	5,162
	Total	10,954
Age Group	0–2	199
	3–9	1,356
	10–19	1,181
	20–29	3,300
	30–39	2,330
	40–49	1,353
	50–59	796
	60–69	321
	>70	118
	Total	10,954

(60–69: 321; >70: 118) containing substantially fewer samples.

Table 3 reports overall performance across the three attribute tasks, with 95% confidence intervals (CI). Among the attributes, gender classification is the strongest, with accuracy of 0.72074 (95% CI: 0.71234–0.72914). Race classification is lower, with accuracy of 0.59540 (95% CI: 0.58621–0.60459). Age-range prediction is the most challenging task, with accuracy of 0.28437 (95% CI: 0.27592–0.29282).

Table 3. *Overall accuracy for age-range, gender, and race prediction on the harmonized FairFace validation set (n = 10,954), with 95% confidence intervals (CI).*

Metric	Accuracy	95% CI (Lower)	95% CI (Upper)
Age-Range Accuracy	0.28437	0.27592	0.29282
Gender Accuracy	0.72074	0.71234	0.72914
Race Accuracy	0.59540	0.58621	0.60459

Age Prediction Results

Table 4 reports age-range accuracy, mean signed error (Age Bias), and the standard deviation (SD) of signed age errors. DeepFace achieves an age-range accuracy of 0.28437, meaning that fewer than one-third of predictions fall within the ground-truth FairFace age intervals. The mean signed error is +3.66 years, indicating average age overestimation, and the SD of 13.55 years shows substantial dispersion in age errors across samples.

Table 4. Age-prediction performance on the FairFace validation set, including age-range accuracy, mean signed error (Age Bias, in years), and the standard deviation (SD) of signed age errors.

Metric	Value
Age-Range Accuracy	0.28437
Age Bias (years)	3.65926
Age Bias SD (years)	13.54673

Figure 3 shows a broad error distribution that is shifted toward positive values. Figure 4 shows that age-range accuracy is highest in the younger adult categories and lower in the older groups, especially above age 50.

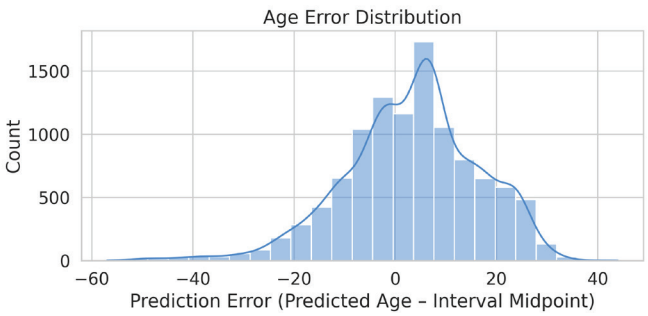

Figure 3. *Distribution of age-prediction errors on the FairFace validation set, where positive values indicate age overestimation and negative values indicate age underestimation.*

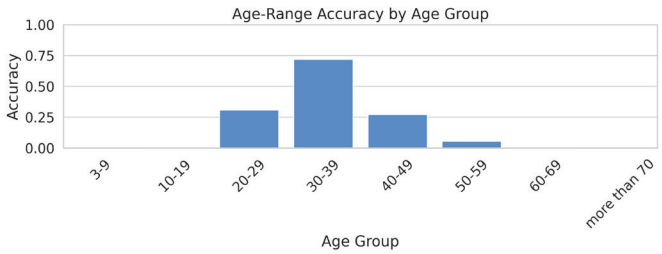

Figure 4. *Age-range accuracy across FairFace age groups, showing the proportion of predictions that fell inside each ground-truth age interval.*

Gender Classification Results

DeepFace achieves a gender-classification accuracy of 0.72074 and a balanced accuracy of 0.70577, indicating moderately strong performance on this binary attribute but substantial asymmetry between classes. Cohen's kappa coefficient (κ) for gender prediction is 0.4234, indicating moderate agreement beyond chance (16).

Per-class precision, recall, and F1 score are summarized in Table 5. The model attains high precision for the female class (0.92144) but low recall (0.44537), meaning that more than half of all true female faces are misclassified. In contrast, the male class shows recall of 0.96616 but lower precision of 0.66154, indicating that the model predicts male much more frequently than female.

Figure 5 makes this imbalance visually clear. Of the 5,162 true female faces, 2,863 are predicted as male, whereas only 196 of the 5,792 true male faces are predicted as female.

These counts indicate a substantially higher false-negative rate for the female class than for the male class.

Table 5. *Gender-classification precision, recall, F1 score, and support on the FairFace validation set. Support denotes the number of ground-truth samples in each class.*

Gender	Precision	Recall	F1 score	Support
Female	0.92144	0.44537	0.6005	5,162
Male	0.66154	0.96616	0.78535	5,792
Accuracy	—	—	0.72074	10,954
Macro average	0.79149	0.70577	0.69292	10,954
Weighted average	0.78402	0.72074	0.69824	10,954

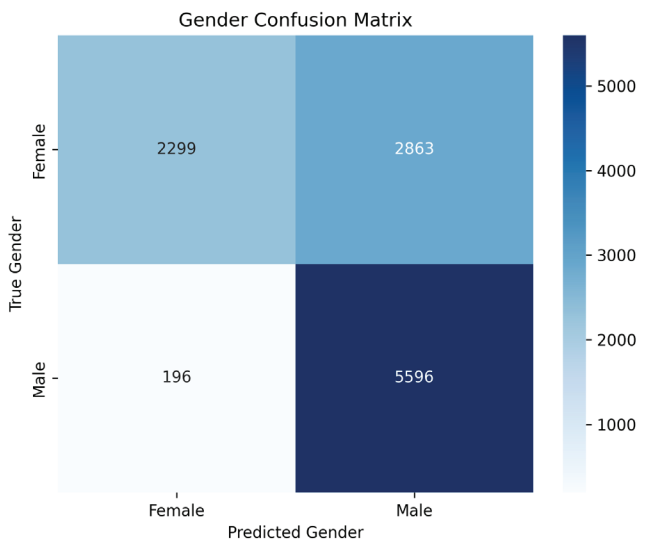

Figure 5. *Gender confusion matrix for DeepFace predictions on the FairFace validation set. Rows represent ground-truth labels and columns represent predicted labels; diagonal cells indicate correct classifications.*

Race Classification Results

DeepFace achieves overall race-classification accuracy of 0.59540 and balanced accuracy of 0.53254, indicating substantial variation in performance among the six consolidated race groups. Cohen's kappa coefficient (κ) for race prediction is 0.49474, indicating moderate agreement beyond chance.

Table 6 reports per-class precision, recall, and F1 score. Performance is strongest for Asian (F1 = 0.775) and Black (F1 = 0.720) faces. In contrast, Indian (F1 = 0.339), Latino/Hispanic (F1 = 0.308), and Middle Eastern (F1 = 0.393) faces show markedly lower recall and F1 values. The macro-averaged F1 score of 0.52423 highlights the uneven performance across classes.

Table 6. *Race-classification precision, recall, F1 score, and support for the six harmonized race categories used in this study (n = 10,954).*

Race	Precision	Recall	F1 score	Support
Asian	0.67474	0.91096	0.77526	2,965
Black	0.6871	0.75643	0.7201	1,556
Indian	0.73246	0.22032	0.33874	1,516
Latino/Hispanic	0.36348	0.26741	0.30813	1,623
Middle Eastern	0.47811	0.33416	0.39338	1,209
White	0.53664	0.706	0.60978	2,085
Accuracy	—	—	0.5954	10,954
Macro average	0.57875	0.53254	0.52423	10,954
Weighted average	0.59038	0.5954	0.56415	10,954

Figure 6 shows systematic misclassification patterns. Indian, Latino/Hispanic, and Middle Eastern faces are frequently predicted as Asian or White, whereas Asian and White faces show stronger diagonal dominance in the confusion matrix.

The chi-square test of independence reported in Table 7 confirms that race and prediction correctness are strongly associated ($\chi^2 = 3451.63$, df = 5, $p < 0.00001$). Cramér's V of 0.561 indicates a large association, supporting the observed variation in per-class performance.

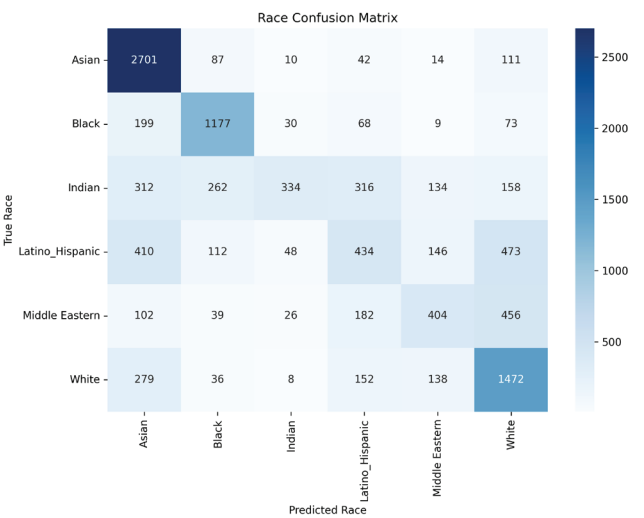

Figure 6. *Race confusion matrix for the six-class harmonized FairFace taxonomy. Rows represent ground-truth race labels and columns represent DeepFace-predicted race labels after East Asian and Southeast Asian were merged into a single Asian category.*

Table 7. *Chi-square test of independence between race group and prediction correctness. χ^2 denotes the chi-square statistic, df the degrees of freedom, and Cramér's V the effect-size measure for the association.*

Statistic	Value
Chi-Square (χ^2)	3,451.63
Degrees of Freedom	5
p-value	< 0.00001
Cramér's V	0.561

DISCUSSION

Interpretation of subgroup disparities

The results indicate that DeepFace transfers unevenly from face verification to demographic attribute classification. Age estimation was the weakest task, with low age-range accuracy and a positive mean signed error, indicating limited reliability when predictions are compared with interval-based ground-truth labels. Gender performance was higher in aggregate, but the confusion matrix shows a pronounced asymmetry in which female faces were much more often predicted as male than the reverse. Race performance was also uneven, with substantially lower recall and F1 values for Indian, Latino/Hispanic, and Middle Eastern groups than for Asian and Black groups. Taken together, these patterns show that overall accuracy alone would understate subgroup-specific errors and that balanced accuracy, macro-averaged F1, confusion matrices, and chi-square testing are necessary for a fairness-oriented evaluation (5, 10, 17).

Comparison with fairness-aware baselines

Comparison with published baselines clarifies the size of the observed gaps. A FairFace-trained ResNet-34 baseline, trained on balanced FairFace labels, reported substantially higher and more uniform subgroup performance than models trained on older, skewed datasets (10). By contrast, DeepFace achieved lower overall gender and race accuracy and larger subgroup disparities, especially in female recall and in the F1 scores for Indian, Latino/Hispanic, and Middle Eastern groups. Fairness-aware architectures provide a second comparison point: InclusiveFaceNet reported high gender accuracy and strong race separability while reducing minority-group false rates (13). Relative to these baselines, the present results suggest that balanced training data and demographic-aware objectives are better suited to equitable attribute prediction than a generic face-verification backbone that was not optimized for demographic parity.

Limitations

A key limitation of this evaluation is the mismatch between the FairFace race taxonomy and the DeepFace race output space. FairFace provides seven race categories, including East Asian and Southeast Asian as distinct labels, whereas DeepFace returns a single Asian label. To enable a like-for-like multi-class evaluation and confusion-matrix reporting, this study merged FairFace's East Asian and Southeast Asian ground-truth labels into one Asian class prior to computing race metrics.

Although this harmonization step is necessary for comparability, it reduces demographic granularity and can mask within-group disparities. The aggregated Asian metric may conceal materially different error rates for East Asian and Southeast Asian faces, and the merge also changes the effective class distribution of the evaluation set. In addition, some published FairFace baselines report results under the original seven-class taxonomy, so direct comparisons may require re-scoring under the same merged mapping. Future work should therefore preserve finer-grained Asian subgroup labels whenever the model output space allows it, or report additional subgroup-specific analyses when harmonization is unavoidable.

CONCLUSION

This study evaluated DeepFace on the FairFace validation set using harmonized labels and a fairness-focused set of metrics. On 10,954 images, the model performed weakest on age prediction, achieved moderate gender accuracy with strong asymmetry between female and male recall, and produced uneven race performance across the six harmonized groups. The statistically significant association between race and prediction correctness further indicates that errors were not evenly distributed across demographic categories.

These findings indicate that high face-verification performance does not guarantee accurate or equitable demographic attribute prediction. Off-the-shelf systems used for age, gender, or race inference should therefore be audited on balanced datasets and reported with subgroup metrics, not only aggregate accuracy. Future work should extend the analysis to intersectional subgroups, calibration, and fairness-aware retraining while preserving more granular race labels whenever possible.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGMENTS

The author thanks the creators of the FairFace dataset for providing a demographically balanced benchmark that enabled rigorous subgroup-level evaluation. The author also acknowledges the developers of the DeepFace framework and the deepface Python library for making their implementation publicly accessible. Constructive comments from anonymous reviewers are sincerely appreciated.

REFERENCES

1. Taigman Y, Yang M, Ranzato M & Wolf L. DeepFace: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014; pp. 1701-1708. https://doi.org/10.1109/CVPR.2014.220
2. Wang M & Deng W. Deep face recognition: A survey. Neurocomputing. 2018; 429: 215-244. https://doi.org/10.1016/j.neucom.2020.10.081
3. Serengil SI & Ozpinar A. LightFace: A hybrid deep face recognition framework. In 2020 Innovations in Intelligent Systems and Applications Conference. 2020; pp. 23-27. IEEE. https://doi.org/10.1109/ASYU50717.2020.9259802
4. Savchenko AV. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. arXiv preprint arXiv:2103.17107. 2021. https://doi.org/10.1109/SISY52375.2021.9582508
5. Buolamwini J & Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018; pp. 77-91).
6. Raji ID & Buolamwini J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2019; pp. 429-435. https://doi.org/10.1145/3306618.3314244
7. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW & Jain AK. Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security. 2012; 7 (6): 1789-1801. https://doi.org/10.1109/TIFS.2012.2214212
8. Merler M, Ratha N, Feris RS & Smith JR. Diversity in faces. arXiv preprint arXiv:1901.10436. 2019.
9. Beutel A, Chen J, Zhao Z & Chi EH. Data decisions and theoretical implications when adversarially learning fair representations. Proceedings of the 2017 NIPS Workshop on Machine Learning and Computer Security. 2017.
10. Karkkainen K & Joo J. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021; pp. 1548-1558. https://doi.org/10.1109/WACV48630.2021.00159
11. Morales A, Fierrez J & Vera-Rodriguez R. SensitiveNets: Learning agnostic representations with application to face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.
12. Cao Q, Shen L, Xie W, Parkhi OM & Zisserman A. VGGFace2: A dataset for recognising faces across pose and age. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. 2018; pp. 67-74. https://doi.org/10.1109/FG.2018.00020
13. Ryu HJ, Adam H & Mitchell M. InclusiveFaceNet: Improving face attribute detection with race and gender diversity. arXiv preprint arXiv:1712.00193. 2018.
14. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27 (8): 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
15. Agresti A. Categorical Data Analysis (3rd ed.). Wiley. 2013.
16. Landis JR & Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33 (1): 159-174. https://doi.org/10.2307/2529310
17. Hardt M, Price E & Srebro N. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems. 2016; pp. 3315-3323.
18. Serengil SI. deepface (Version 0.0.95) [Computer software]. 2024. https://github.com/serengil/deepface.
19. Niu Z, Zhou M, Wang L, Gao X & Hua G. Ordinal regression with multiple output CNN for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; pp. 4920-4928. https://doi.org/10.1109/CVPR.2016.532
20. Brodersen KH, Ong CS, Stephan KE & Buhmann JM. The balanced accuracy and its posterior distribution. In Proceedings of the 20th International Conference on Pattern Recognition. 2010; pp. 3121-3124. IEEE. https://doi.org/10.1109/ICPR.2010.764
21. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological

Measurement. 1960; 20 (1): 37-46. https://doi.org/10.1177/001316446002000104
22. McHugh ML. The chi-square test of independence. Biochemia Medica. 2013; 23 (2): 143-149. https://doi.org/10.11613/BM.2013.018
23. Dwork C, Hardt M, Pitassi T, Reingold O & Zemel R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012; pp. 214-226. https://doi.org/10.1145/2090236.2090255