# Uncovering Gender Biases in Gender Identification Models for Japanese Data Analysis

Ana Manzano Rodríguez[1,3] * Camille Guinaudeau[2] Shin'ichi Satoh[3]
[1]Universitat Politècnica de Catalunya, Spain
[2]Japanese French Laboratory for Informatics, CNRS / University Paris Saclay, France
[3]National Institute of Informatics, Japan

## Abstract

*Gender bias in facial recognition systems is a critical issue that affects the accuracy and fairness of these technologies. This paper investigates the performance of gender classification models across diverse datasets, focusing on the CNN-based Face-Gender-Classification-PyTorch model. We evaluate the model on three key datasets: Kaggle Gender Classification, UTKFace and All-Age-Faces. Initial experiments reveal that while the model performs well on Kaggle dataset, its accuracy drops on Asian data with notable gender performance disparities. To address this, we apply fine-tuning, among other strategies, using the FairFace dataset and a mixed dataset approach. While FairFace alone improves overall accuracy, combining it with the mixed dataset produces more balanced results—reducing the gender gap from 30 to 6 percentage points and achieving near-optimal accuracy. The findings provide evidence of racial and gender bias and show it can be mitigated through approachable data balancing techniques. We further analyze model behavior by evaluating racial group performance using UTKFace and applying Grad-CAM to interpret decision-making. Finally, we test the best-performing model on Japanese TV data demonstrating its potential for large-scale gender fairness monitoring.*

## 1. Introduction

Gender bias in Artificial Intelligence (AI) refers to the tendency of AI systems to produce outcomes that disproportionately favor one gender over another. As AI becomes increasingly integrated into critical areas of society, this bias raises growing concerns. Its root causes typically stem from two primary sources: the data used to train these systems and the design of the algorithms themselves.

Data bias occurs when training datasets underrepresent certain genders—for instance, a model trained predominantly on male images may struggle to accurately classify female faces. Algorithmic bias, on the other hand, can arise from design choices that unequally emphasize features, further reinforcing these disparities.

The consequences of gender bias in AI are far-reaching. Differences in performance across genders can result in unequal treatment in domains such as facial recognition, recruitment, or healthcare [9]. In the context of media monitoring—particularly in news—fair gender identification models are crucial. While automated approaches using speech [11], facial recognition [10], or voice [4] enable large-scale analysis, their performance often drops significantly on Japanese data, even when they are claimed to be race- and gender-unbiased. This is partly due to imbalanced training datasets that lack sufficient racial diversity. Gender-related performance gaps persist as well, with lower accuracy for female predictions. When uncertain, the model often defaults to male, further compounding the bias.

This paper presents a threefold contribution. First, we systematically evaluate the biases of two widely-used models across diverse datasets encompassing several racial groups. Second, we assess the effectiveness of two potential solutions—fine-tuning and pseudo-labeling—to improve the fairness of gender identification models. Finally, we evaluate the performance of the best model on a manually labeled Japanese dataset. This evaluation aims to determine the feasibility of automatically analyzing large-scale TV archives from multiple Japanese TV channels.

## 2. Related Work

The presence of gender and racial bias in AI systems, particularly within facial recognition, has been widely acknowledged in recent studies. For instance, seminal work by Buolamwini and Gebru [2] revealed that commercial facial recognition systems perform worse for women and darker-skinned individuals, underscoring the need for more diverse and inclusive training datasets. Zhao et al. [14] also revealed that even seemingly neutral datasets can reinforce

*The author is currently affiliated with the University of Amsterdam

stereotypes if not carefully curated, emphasizing the need for rigorous dataset auditing and mitigation techniques.

While much of the literature has focused on Western contexts, there is increasing recognition of the challenges associated with applying AI models to non-Western populations. Studies such as those by Drozdowski et al. [3] have shown that facial recognition systems trained predominantly on Western datasets often underperform on Asian faces.

Methods such as transfer learning [13] demonstrated improvements in predictive performance for underrepresented groups. However, despite these advances, biases are not fully eliminated, particularly when there are significant differences between source and target datasets. Unsupervised domain adaptation, as shown by Hoffman et al. [6], offers a promising solution to align model performance across domains, yet its application to non-Western gender identification—especially in Japanese contexts—remains underexplored. This paper aims to fill this gap by focusing on the performance of face-based gender identification models when applied to Japanese data.

## 3. Face-based Gender Identification

### 3.1. Models

This research employs a CNN based on the ResNet18 architecture [5], hereafter the *PyTorch model*[1]. Its robustness stems from its deep architecture and the integration of a residual learning framework, making it well-suited for applications in biometrics, security, and demographic analysis. Specifically adapted to achieve high accuracy in identifying gender from facial images, the model is trained on the Kaggle classification dataset, described in the next Section.

The second model is the open-source *inaFaceAnalyzer*[2], built on a ResNet50 architecture and trained on the FairFace dataset [8] (see Section 4 for more details about this dataset). It uses OpenCV's CNN back-end for face detection with optimized parameters for gender classification. The model is tuned to minimize demographic bias—defined as error disparities across groups such as male/female, younger/older adults, and white/non-white individuals.

### 3.2. Datasets

The experiments are conducted using three large datasets: the Kaggle Gender Classification Dataset [7], the UTKFace Dataset [12], and the All-Age-Faces (AAF) Dataset [1], following standard splits for training, validation, and testing.

The **Kaggle Gender Classification Dataset (Kaggle)** contains cropped facial images labeled as male or female. The training set includes approximately 23,000 images per gender, and the validation set contains around 5,500 per gender. The dataset is demographically diverse but lacks detailed information on race and age.

The **UTKFace Dataset (UTKFace)** comprises over 20,000 facial images annotated with age, gender and ethnicity. Subjects range in age from 0 to 116 years, and the gender distribution is balanced, with 52.3% male and 47.7% female. The dataset shows wide variability in pose, expression, illumination, occlusion, and resolution. It was selected not only for its gender balance but also for its detailed ethnicity and age labels, which provide rich insights.

The **All-Age-Faces Dataset (AAF)** consists of 13,322 facial images, primarily of individuals of Asian descent, aged between 2 and 80 years. It includes 7,381 female and 5,941 male images. This dataset is particularly valuable for evaluating performance on underrepresented Asian demographics.

| Validation Dataset | PyTorch model | inaFaceAnalyzer |
|---|---|---|
| **Kaggle** | 97.21<br>**95.60** - *98.81* | 94.39<br>**89.02** - *99.6* |
| **UTKFace** | 90.44<br>**95.44** - *84.79* | 94.6<br>**90.4** - *99.56* |
| **AAF** | 70.40<br>**97.93** - *48.46* | 95.94<br>**97.04** - *95.02* |

Table 1. Performance comparison on different validation datasets with the pre-trained PyTorch and *InaFaceAnalyzer* models, including general accuracy and accuracies separated by gender. Accuracy values for male (resp. female) are in **bold** (resp. *italic*). Best results are underlined.

### 3.3. Models Comparison

In this initial experiment, the goal is to evaluate the performance of the selected model across the three datasets. This assessment is conducted on the validation set without additional training, using the pre-trained PyTorch model and inaFaceAnalyzer.

As shown in Table 1, the PyTorch model underperforms compared to inaFaceAnalyzer on all datasets except Kaggle, which was used for its original training. Its performance drops particularly on the Asian data, with accuracy on the AAF dataset falling to 70.40%, indicating potential racial bias. Moreover, inaFaceAnalyzer shows a smaller accuracy gap between male and female predictions, suggesting improved fairness. In contrast, the PyTorch model shows growing gender disparity as validation data diverges from the training set. This highlights the advantage of inaFaceAnalyzer's training on the diverse FairFace dataset.

However, a closer examination reveals that inaFaceAnalyzer fails to detect a significant number of faces—especially female faces. In the AAF dataset, for instance, 26% of faces are not detected, most of which are asian women. This indicates that while its classification accuracy may appear superior, the model still exhibits bias at the face detection stage.

[1] https://github.com/ndb796/Face-Gender-Classification-PyTorch
[2] https://github.com/ina-foss/inaFaceAnalyzer

| Validation Dataset | No Fine-tuning | FairFace | FairFace + Mixed Dataset |
|---|---|---|---|
| **Kaggle** | 97.21<br>**95.60** - *98.81* | 96.10<br>**93.75** - *98.44* | 97.34<br>**96.29** - *98.39* |
| **UTKFace** | 90.44<br>**95.44** - *84.79* | 96.74<br>**97.38** - *96.02* | 94.37<br>**96.61** - *91.86* |
| **AAF** | 70.40<br>**97.93** - *48.46* | 86.41<br>**95.36** - *79.28* | 96.27<br>**96.27** - *96.33* |

Table 2. Impact of fine-tuning with FairFace and FairFace + Mixed datasets, including general accuracy and accuracies separated by gender. Accuracy values for male (resp. female) are in **bold** (resp. *italic*). Best results are underlined.

For the remainder of the paper, we focus on improving the PyTorch model by balancing the training data through fine-tuning and pseudo-labeling to improve performance across all racial groups, with a particular focus on Japanese data. The task of improving fairness in face detection used in the inaFaceAnalyzer tool will be reserved for future work.

## 4. Model Fine-tuning

To enhance the fairness of the PyTorch model, ensuring it achieves consistent results for both males and females across all racial groups, we will fine-tune the pre-trained model (originally based on the Kaggle dataset) using two additional datasets: the FairFace dataset [8] and a combined dataset that merges Kaggle data with All-Age-Faces data.

The FairFace dataset contains 108,501 face images balanced across seven racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. It also includes labels and metrics for balanced accuracy across gender, race and age groups. The Mixed dataset consists of the full AAF dataset (predominantly Asian faces) and a 20% sample from the Kaggle dataset, added to mitigate the AAF dataset's heavy Asian bias and to create a more balanced training set with racial distribution under our control.

Fine-tuning was performed in two stages to improve both generalization and demographic fairness. First, the model was trained for 5 epochs on the FairFace dataset, with early stopping to prevent overfitting while capturing broad demographic diversity. Then, it was further fine-tuned for 5 epochs on a Mixed dataset, specifically curated to improve performance on Japanese facial data by enriching representation of this group. This sequential approach balances overall accuracy with targeted fairness improvements.

As shown in Table 2, fine-tuning significantly improves results on the UTKFace and AAF datasets, with no negative impact on Kaggle performance. FairFace-only fine-tuning yields the most balanced results for UTKFace, with male and female accuracy nearly aligned. However, adding the Mixed dataset in the second phase slightly lowers accuracy and increases the gender gap for UTKFace. For AAF, the mixed dataset yields the best performance, boosting female accuracy by nearly 50 percentage points—though this may be optimistic, as some AAF training samples were included in the fine-tuning data. In Section 5, we will demonstrate that, although the performance achieved through fine-tuning with both the FairFace and Mixed datasets is not matched, the fine-tuning process can still be enhanced by using only the FairFace dataset in conjunction with a pseudo-labeling strategy applied to data from Japanese TV programs.

| Ethnicities | No Fine-tuning | FairFace | FairFace + Mixed Dataset |
|---|---|---|---|
| **White** | 89.51<br>**89.74** - *89.21* | 96.64<br>**97.29** - *95.80* | 94.40<br>**95.58** - *92.87* |
| **Black** | 90.49<br>**92.22** - *88.68* | 95.86<br>**95.81** - *95.91* | 94.63<br>**94.61** - *94.65* |
| **Asian** | 90.02<br>**90.96** - *89.32* | 97.08<br>**97.74** - *96.58* | 97.08<br>**98.31** - *94.15* |
| **Indian** | 91.32<br>**91.64** - *90.91* | 97.55<br>**98.33** - *96.54* | 94.53<br>**94.65** - *94.37* |
| **Others** | 88.17<br>**85.54** - *90.29* | 96.24<br>**95.18** - *97.09* | 95.16<br>**98.80** - *92.23* |

Table 3. Performance of the PyTorch model, with and without fine-tuning on each racial group of the UTKFace dataset, including general accuracy and accuracies separated by gender. Accuracy values for male (resp. female) are in **bold** (resp. *italic*). Best results are underlined.

Finally, we used the UTKFace dataset, which includes ethnicity labels, to assess how model performance varies across racial groups. As shown in Table 3, FairFace fine-tuning improves both accuracy and gender balance across all groups. However, further fine-tuning with the Mixed dataset offers no additional benefits, and for Asian faces, it slightly worsens the gender gap, even though overall accuracy remains stable.

## 5. Pseudo-labeling using Japanese TV Archive

We have access to a TV archive spanning over 20 years of the *NHK News 7* on the public TV channel *The Japan Broadcasting Corporation*. This extensive dataset provides a valuable opportunity to improve the PyTorch model through pseudo-labeling, a semi-supervised technique where the model assigns labels to unlabeled data. These predicted labels, known as pseudo-labels, are treated as ground truth for retraining the model, thus leveraging the additional data to potentially enhance its performance.

To this end, we apply the PyTorch model (with FairFace fine-tuning) to images extracted from *NHK News 7* broadcasts from 2001. Since the model provides confidence scores with its predictions, we selected only samples exceeding a given confidence threshold. To ensure gender balance, we used equal numbers of male and female images, though this reduced the dataset size due to fewer available

| **Confidence threshold** | - | 0.3 | 0.6 | 0.9 |
|---|---|---|---|---|
| **AAF** | 86.41<br>**95.36** - *79.28* | 91.21<br>**93.91** - *88.92* | 91.02<br>**92.54** - *89.80* | 89.22<br>**86.21** - *91.78* |

Table 4. Performance of pseudo-labeling on the AAF dataset, varying confidence thresholds. Accuracy values for male (resp. female) are in **bold** (resp. *italic*). Best results are underlined.

female samples.

As shown in Table 4, the best overall accuracy is achieved with a confidence threshold of 0.3, suggesting that including more data—even with lower confidence—is more beneficial than using only high-confidence samples. However, using thresholds of 0.5 or higher reduces the accuracy gap between male and female predictions. At the 0.9 threshold, overall accuracy declines further, but female identification reaches its highest level, surpassing male performance.

## 6. Gender Equality in Japanese TV News

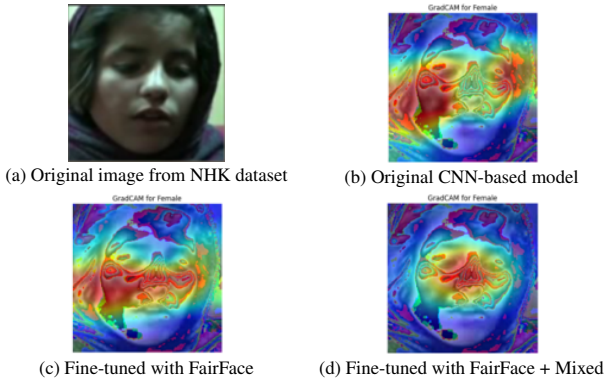
To evaluate the feasibility of gender equality analysis on Japanese TV content, we tested the PyTorch model—both with and without fine-tuning—on a manually annotated dataset of over 1,200 facial images from the broadcast of *NHK News 7*, aired by *The Japan Broadcasting Corporation*. This dataset is gender-imbalanced, with 70% male and 30% female images.

| | No Fine-tuning | FairFace | FairFace + Mixed Dataset |
|---|---|---|---|
| **NHK Dataset** | 87.36<br>**95.62** - *64.83* | 94.83<br>**97.53** - *87.46* | 94.58<br>**96.07** - *90.51* |

Table 5. Performance of the PyTorch model on the NHK dataset using different fine-tuning strategies, including general accuracy and accuracies separated by gender. Accuracy values for male (resp. female) are in **bold** (resp. *italic*). Best results are underlined.

Table 5 shows that the model's performance on Japanese TV data is notably low, particularly for females, where accuracy falls below 65%. The gender performance gap exceeds 30 percentage points in the baseline model. Fine-tuning with the FairFace dataset yields the best overall accuracy, but the gender gap remains around 10 points. The most balanced performance is achieved by fine-tuning on both FairFace and the Mixed dataset, reducing the gap to 6 points while maintaining near-peak accuracy.

To better understand model behavior, we applied GradCam (Gradient-weighted Class Activation Mapping) to visualize which image regions the CNN focuses on during predictions. As illustrated in Figure 1, the base model lacks consistent focus, while the first fine-tuning improves attention to facial regions, and the second further sharpens this focus—indicating more reliable and interpretable decision-making.




(a) Original image from NHK dataset (b) Original CNN-based model

(c) Fine-tuned with FairFace (d) Fine-tuned with FairFace + Mixed

Figure 1. Gradient-weighted Class Activation Mapping for the PyTorch model, without and with fine-tuning on a sample from the NHK Dataset.

In conclusion, despite a remaining 6-point gender accuracy gap, the model achieves over 90% accuracy for both genders and is suitable for large-scale analysis of Japanese TV news archives. Nonetheless, some residual bias should be considered during manual review of the results.

## 7. Conclusion

This paper investigates the performance and fairness of gender classification models across diverse racial and gender categories using a CNN evaluated on three major datasets. We establish baseline performance and improve it through fine-tuning with the FairFace and Mixed datasets. FairFace alone yields the highest accuracy but retains a notable gender gap, while combining it with the Mixed dataset achieves more balanced results.

Our experiments confirm the presence of racial and gender bias—particularly towards Asian individuals—and show it can be mitigated with relatively simple techniques. This highlights the importance of data curation: when we prioritize the quality and balance of the training data, fairness can be achieved without sacrificing accuracy. We emphasize that balanced datasets are essential for ensuring the responsible deployment of these tools in society.

To better understand model behavior, we further analyzed predictions across racial groups and applied GradCAM to interpret the model's decision-making process. Some limitations emerged from the data itself: manual inspection revealed mislabeled samples, likely due to semi-supervised annotation, and ambiguous cases where even human annotators struggled to determine gender.

Finally, we tested the best-performing model on a Japanese news dataset, achieving over 94% accuracy with a remaining 6-point gender gap, demonstrating the model's potential for large-scale applications. Future work will extend these improvements to other models to further promote fairness in face analysis.

### Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

### Acknowledgments

This work was carried out during the first author's internship at the National Institute of Informatics (NII) in Japan. The authors gratefully acknowledge NII's financial support.

## References

[1] All-Age-Faces Dataset. All-age-faces (aaf) mostly asian dataset, 2024. Last accessed 2024/06/17. 2

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*, 2018. 1

[3] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. 2

[4] Hadi Harb and Liming Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 24:179–198, 2005. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint*, 2015. 2

[6] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 2

[7] Kaggle Gender Classification Dataset. Kaggle gender classification dataset, 2024. Last accessed 2024/05/10. 2

[8] Kimmo Karkkainen and Jungseock Joo. 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*, pages 1548–1558, 2021. 2, 3

[9] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (23):12592–12594, 2020. 1

[10] Doughlas Pardede, Wanayumini Wanayumini, and Rika Rosnelly. A combination of support vector machine and inception-v3 in face-based gender classification. In *Proceeding of International Conference on Information Science and Technology Innovation (ICoSTEC)*, pages 34–39, 2023. 1

[11] Esther Ramdinmawii and Vinay Kumar Mittal. Gender identification from speech signal by examining the speech production characteristics. In *2016 International conference on signal processing and communication (ICSC)*, pages 244–249. IEEE, 2016. 1

[12] UTKFace Dataset. Utkface dataset, 2024. Last accessed 2024/05/04. 2

[13] Keyao Zhan, Xin Xiong, Zijian Guo, Tianxi Cai, and Molei Liu. Transfer learning targeting mixed population: A distributional robust perspective. *arXiv preprint arXiv:2407.20073*, 2024. 2

[14] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 1