Towards Fairness in Machine Learning: Balancing Racially Imbalanced Datasets Through Data Augmentation and Generative AI

Anthonie Schaap[a], Sofoklis Kitharidis[b] and Niki van Stein[c]
Leiden Institute of Advanced Computer Science, Leiden University, Einsteinweg 55, 2333 CC Leiden, Netherlands
a.l.j.schaap@issc.leidenuniv.nl, {s.kitharidis, n.van.stein}@liacs.leidenuniv.nl,

Keywords: Fairness, Fair Machine Learning, Racial Bias, Inclusive AI, GANs, Dataset Balancing, Ethnicity Classification.

Abstract: Existing AI models trained on facial images are often heavily biased towards certain ethnic groups due to training data containing unrealistic ethnicity splits. This study examines ethnic biases in facial recognition AI models, resulting from skewed dataset representations. Various data augmentation and generative AI techniques were evaluated to mitigate these biases, employing fairness metrics to measure improvements. Our methodology included balancing training datasets with synthetic data generated through Generative Adversarial Networks (GANs), targeting underrepresented ethnic groups. Experimental results indicate that these interventions effectively reduce bias, enhancing the fairness of AI models across different ethnicities. This research contributes practical approaches for adjusting dataset imbalances in AI systems, ultimately improving the reliability and ethical deployment of facial recognition technologies.

1 INTRODUCTION

As artificial intelligence (AI) continues to evolve and integrate into daily life, the need for transparent and fair AI systems becomes increasingly critical. Current AI systems can exhibit unwanted biases, often due to biases present in the datasets they are trained on (Srinivasan and Chander, 2021). These biases can significantly impact real-world applications, where human-like biases often arise due to imbalances in ethnicity, age, or gender distribution within the training data. Biases can arise from uneven distribution of attributes like age, ethnicity, and gender, as discussed in (Wang et al., 2019). It is essential to tackle the problem of model bias, which can have real-life implications, such as in AI-driven hiring processes (Bogen and Rieke, 2018) or risk assessments predicting the likelihood of a defendant reoffending (Mehrabi et al., 2022). In this paper, ethnicity bias in datasets and AI models is investigated. To detect ethnicity bias in datasets, an AI model is first trained to classify different ethnicities. This model requires mitigation of ethnicity bias to minimize its influence on classification accuracy. Datasets for ethnicity classification can contain a lot of unwanted biases, and are often unbalanced, as seen in Section 3.1. This research aims to explore the presence of ethnicity bias in datasets not specifically used for ethnicity classification but for other purposes, such as age classification and person re-identification. When heavy ethnicity biases are present in a dataset but are unknown, researchers might develop biased and racial AI models that may better classify certain ethnicities more represented in the dataset. Historically many problems such as the Google Photos misdemeanor were coupled with people of color, dehumanizing them and mislabeling them. This research addresses the following questions:

1. To what degree do existing face datasets and AI systems show biases related to ethnicity?
2. How can fairness within facial datasets be improved through diverse modifications to the training data?
3. Which method of adding facial images to the training data best mitigates unwanted bias and improves fairness?

Following the introduction of these key concepts discussed in this paper, similar areas of research related to this topic are examined, found in Section 2. After a review of several state-of-the-art papers on ethnicity bias and explainability, the methods pro-

[a] https://orcid.org/0009-0007-4961-749X
[b] https://orcid.org/0009-0005-8404-0724
[c] https://orcid.org/0000-0002-0013-7969

Schaap, A., Kitharidis, S. and van Stein, N.
Towards Fairness in Machine Learning: Balancing Racially Imbalanced Datasets Through Data Augmentation and Generative AI.
DOI: 10.5220/0013002600003837
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Joint Conference on Computational Intelligence (IJCCI 2024)*, pages 584-592
ISBN: 978-989-758-721-4; ISSN: 2184-3236
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

posed are outlined in Section 3, where the datasets used in our experiments are also described. In Section 4, the different experimental setups are described, including the training of the ethnicity classification model and the use of a generative AI model in comparison with several state-of-the-art approaches. The results of these experiments, which compare our proposed method to existing methods, are also presented within this section. Following the presentation of these results, a discussion of the findings is provided. Lastly, conclusions are drawn and some limitations and suggestions for future work are given in Section 6.

2 RELATED WORK

2.1 Bias in AI and Facial Recognition

Extensive research has highlighted the prevalence of ethnic biases in AI systems. (Angwin et al., 2016) found that algorithms used in the criminal justice system disproportionately predicted more severe outcomes for Black individuals, although (W Flores et al., 2016) disputed these findings based on statistical flaws.

In facial recognition, biases often arise from dataset imbalances (Islam, 2023), where racial stereotypes skew the representation of certain groups. Techniques such as detecting segregation patterns in datasets have been proposed to reduce disparities among different ethnic groups (Benthall and Haynes, 2019). Additionally, (Alvi et al., 2019) introduced a joint learning and unlearning algorithm that improved fairness and accuracy by addressing biases related to gender, age, and ethnicity in neural networks.

Advanced facial recognition frameworks such as DeepFace (Serengil and Ozpinar, 2021), VGG-Face2 (Cao et al., 2018), and FaceNet (Schroff et al., 2015) have been used for ethnicity classification tasks. Pretrained models like ResNet50 have also demonstrated robust performance in classifying ethnicity and gender attributes (Acien et al., 2019).

2.2 Methods for Mitigating Bias in AI

Mitigating bias in AI requires proactive measures to identify and reduce biases in models (Howard and Borenstein, 2018). (Raji et al., 2020) proposed a framework to improve AI accountability by detecting biases, while (Danks and London, 2017) emphasized that model architecture can also harbor biases, necessitating careful design and testing. Additionally, techniques like synthetic faces can reveal unnoticed biases related to appearance (Balakrishnan et al., 2020).

Generative Adversarial Networks (GANs) are a promising tool for mitigating bias by generating samples representing underrepresented groups (Maluleke et al., 2022). Advances like StyleGAN have improved the generation of photorealistic images, allowing researchers to manipulate key features such as pose and expression (Karras et al., 2019). Building on this, (Nitzan et al., 2020) introduced a method to separate facial identity from other attributes, which we utilize in our research to create diverse labeled data and improve fairness in model training.

3 PROPOSED METHODOLOGY

Three methods are proposed and evaluated to mitigate bias by altering the training data during or before model training. This section also outlines the datasets critical for our methods.

3.1 Datasets

Experiments are conducted using various datasets tailored for ethnicity classification. The first dataset, UTKFace (Zhang et al., 2017), comprises over 20,000 images labeled with age, gender, and ethnicity. This dataset is primarily used for training our ethnicity detection model, featuring diverse labels. Labels include age (0 to 116 years), gender (0 for male, 1 for female), and race (0 for White, 1 for Black, 2 for Asian, 3 for Indian, 4 for Others). Particularly, our focus is to address label imbalance, aiming for an evenly distributed dataset across all racial categories. The racial composition split is illustrated in Figure 1.

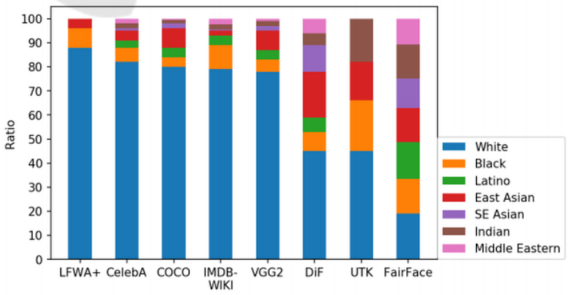

Figure 1: Different racial compositions in face datasets from (Kärkkäinen and Joo, 2019).

As depicted in Figure 1, the UTKFace dataset does not have an evenly split on racial labels, but has a far better split than other state-of-the-art face datasets, excluding the FairFace dataset (Kärkkäinen and Joo, 2019). Bias in training data, such as group imbal-

ances, typically leads to biased outcomes in AI models. Having a perfectly unbiased dataset is nearly impossible, but the goal is to mitigate the bias as much as possible.

Additionally, Figure 1 highlights that, except for FairFace, a disproportionately high number of 'White' images dominate most advanced datasets. This imbalance can cause numerous issues; for instance, when estimating the ages of individuals in a dataset, the AI model is likely to perform better at classifying images with the 'White' racial label due to the greater volume of training data available for this group, possibly resulting in poorer performance for other racial groups.

To address this issue, experiments are also conducted with the FairFace dataset (Kärkkäinen and Joo, 2019) which provides an almost even split for the racial labels White, Black, Latino, East Asian, Southeast Asian, Indian, and Middle Eastern. The FairFace dataset contains 108.501 images, each labeled with race, gender, and age. The racial compositions of the UTKFace and FairFace datasets are presented in Table 1. For clarity in presentation and analysis, the categories are grouped in the following manner in our tables: 'White', 'Black', 'Asian' (combining East Asian and Southeast Asian), 'Indian', and 'Others' (encompassing Latino and Middle Eastern).

Table 1: Ethnicity split in the UTKFace and FairFace datasets.

Ethnicity	Count	
	UTKFace	FairFace
White	8080	16527
Black	3621	12233
Asian	2759	23082
Indian	2146	12319
Others	1358	22583

3.2 Methodology

The following methods are proposed to sample or generate additional data for underrepresented ethnic groups.

Data Sampling Strategies

To use the datasets and train the model, some data engineering is needed to match the data structures of the different image datasets. Data preparation follows the UTKFace dataset structure outlined in Section 3.1. First, the FairFace dataset needs to be in the same structure as the UTKFace dataset, where the image labels are in the name of the images. After restructuring, images are sorted into ethnicity-specific folders to facilitate targeted training on each class. This splitting process is also done for the newly created images using data augmentation and generative networks. For the generative images, the StyleGAN is used to save the newly created images based on its ethnicity. After the ethnicities of the created images have been split, the name of every image is changed to match the style of UTKFace. Names are updated with a new age value and the appropriate ethnicity label. Timestamps are assigned with slight delays to ensure unique filenames for each image. Each implementation can be used as a separate dataset or used for additional training with specific ethnicity classes.

Data Augmentation

To increase both the size and quality of our datasets, data augmentation is used to generate additional training data. Data Augmentation for image data is done by adjusting the image slightly, such that the model cannot recognize the image from its original state and can use the image to improve its learning process. Image adjustments include moving, rotating, flipping, cropping, and shifting to diversify training data. Color adjustments are also applied to prevent the model from learning biases associated with specific color schemes. Our aim is to keep augmented images realistic, avoiding excessive vertical flips or rotations. Additionally, the augmented images are saved to allow for manual quality control and future reuse. Table 2 displays the data augmentation methods used in this research.

Table 2: Augmentation Methods Used to generate more realistic images.

Aug. Method	**Value**	**Description**
Shear Range	0.05	Shear image by 5%
Zoom Range	[1.0, 1.2]	Zoom in up to 20%
Rotation	5	Rotate max. 5 deg.
Horizontal Flip	[True,False]	50% flip chance

The augmentation values are randomly chosen in range as shown in Table 2. The images are not zoomed out, because that would create borders around the image which are not part of the original image. Moreover, image shifting is avoided to prevent border creation and information loss. The augmented images are labeled the same way the original images are labeled, except with a new creation timestamp.

Generating Training Data Using StyleGAN

Another method to enhance our training dataset is by generating new images using the attributes of existing ones. This is done using an implementation of

StyleGAN (Nitzan et al., 2020), where we use the inference function to create new facial images based on the identity and attributes of images from the training set. For every combination of identity and attribute image, a new image is created. For each classification label, 2000 images are generated using 10 randomly chosen attribute images and 200 identity images. This expansion of the training set improves its stability and quality, helping to prevent overfitting. When handling small datasets, the use of StyleGAN to expand the size of the training set can be very considerable, because the images created are different from the original dataset. Generated images retain the ethnicity label of the original identity image, ensuring consistency in label integrity. This research also ensure that the attribute image is from the same ethnicity label, increasing the likelihood that the generated image belongs to that particular label class.

Among the various generative models evaluated, StyleGAN was selected for its robustness and ease of use. Other GANs were considered but not used due to practical challenges such as unavailable model weights, expired resource links, and substantial implementation errors, which impeded reproducibility. StyleGAN's capacity to manipulate attribute and identity vectors allows for the generation of high-quality, diverse images, facilitating the study's focus on enhancing dataset fairness.

The StyleGAN implementation (Nitzan et al., 2020) utilizes a variety of pre-trained models developed using diverse datasets, such as FFHQ at resolutions of 256x256 and 1024x1024 with 70,000 images each, VGGFace2 with 3.31 million images, Celeb-500K with 500,000 images, and 300W with 300 images. These pre-trained models enable efficient generation of new, diverse images, examples of which are showcased in Figure 2.

3.3 Fairness Metrics for Performance Evaluation

Various fairness metrics can be used to evaluate the performance and fairness of the model. Models can be evaluated based on their prediction accuracy for each class, allowing for the determination of how many true positives were correctly classified for each class. However, in scenarios where classes are imbalanced, accuracy is not a good and informative metric. It is important to also consider the false positives, true negatives, and false negatives in the predictions. This section examines different fairness metrics used in this paper to evaluate fairness.

(a) White. (b) Black.

(c) Asian. (d) Indian.

Figure 2: Some generated images from the StyleGAN model, using UTKFace images.

F1 Score

The F1 score, a harmonic mean of Precision and Recall, assesses classification performance, providing a balance between the precision and recall for each label (Goutte and Gaussier, 2005):

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1 score ranges from 0 to 1, with 1 indicating perfect classification without error.

Equalized Odds

The Equalized Odds metric (Hardt et al., 2016) evaluates model fairness by comparing True Positive Rates (TPR) and False Positive Rates (FPR) across classes:

$$TPR = \frac{TP}{TP+FN}, \qquad FPR = \frac{FP}{FP+TN}$$

The metric focuses on minimizing the maximum difference between the TPRs or FPRs across classes, thereby enhancing fairness. It is given by:

$$\text{Eq. Odds} = max(|TPR1 - TPR2|, |FPR1 - FPR2|)$$

This research calculates Equalized Odds for all class combinations, aiming to reduce the discrepancies in recall and FPR, which could indicate bias. The average of these differences is taken to represent the model's overall fairness in handling class disparities.

4 EXPERIMENTS

4.1 Ethnicity Detection Baseline Model

Experiments were conducted using a baseline model for ethnicity detection, trained on the UTKFace dataset. During 10 training epochs, the model achieved a validation accuracy of about 91% and a loss of 0.30.

The primary goal is not just high accuracy but ensuring balanced accuracy across all ethnic groups. This is critical for the model's applicability across diverse datasets. A model may show high overall accuracy but under-perform significantly on certain ethnicity labels, like 'Indian', making it unfit for detecting ethnicity bias. It is also essential that the model avoids defaulting to a single ethnicity label when predictions are inaccurate.

Complete implementation details are available in our code repository (Anonymous, 2024).

4.2 Effect of Additional Training Methods

In this experiment, the equal FairFace dataset is used, where each label has 4000 images. The goal is to create a fair baseline, meaning that every label can be classified with a similar accuracy. The similarity of the classification accuracy is based on a difference of 2%, such that no label performs more than 2% better than any other label. Ideally, the training set should be ethnicity-balanced to achieve similar accuracy statistics. In practice, this is often not the case, and additional training is needed to balance the model. In this experiment, we investigate the effect of adding 2000 images of a certain ethnicity label to the dataset, such that the model can improve its learning process on a particular label. Table 3 presents the performance statistics of the baseline model on the UTKFace and FairFace test sets.

Table 3: Baseline statistics for the Equal FairFace dataset, tested on the FairFace and UTKFace test sets.

	FairFace		UTKFace	
	F1-score	Acc	F1-score	Acc
White	0.59	52.0%	0.66	51.0%
Black	0.75	74.4%	0.81	76.2%
Asian	0.83	86.2%	0.83	85.9%
Indian	0.55	66.1%	0.68	66.2%
Others	0.52	47.8%	0.25	65.9%

As illustrated in 3, the baseline model trained on the FairFace dataset is evaluated using two test sets. This approach mitigates biases inherent in using a test set derived from the same dataset. For UTKFace, most of the images are labeled as "White", where the same ethnicity split is used in the test set. This makes the test accuracy far higher in comparison with using a test set with a balanced ethnicity split. It also prevents the model from learning dataset-specific patterns that could unreasonably influence test set accuracy. For the FairFace test set, the F1-score is the lowest for the "Indian" label, excluding the "Others" label. To address this, 2000 data augmented "Indian" images were added to the original baseline training data, and continued training. The results are presented in Table 4.

Table 4: Additional training on "Indian" label - FairFace and UTKFace test set results.

	FairFace		UTKFace	
	F1-score	Acc	F1-score	Acc
White	0.62	84.2%	0.86	84.2%
Black	0.72	58.3%	0.70	54.4%
Asian	0.80	90.2%	0.78	88.4%
Indian	0.52	64.3%	0.66	88.2%
Others	0.11	6.1%	0.05	2.7%

When adding data, it becomes evident that the model stops trying to classify the "Others" label, to focus on the four more recognizable labels. The "Others" category consists of images that could be placed in one of the four ethnicity labels, which makes the "Others" category weak. Consequently, we exclude the "Others" category in subsequent experiments to avoid its negative impact.

Equalized Odds on Real Data

A different method to evaluate the model involves using the Equalized Odds fairness metric from Section 3.3, aiming for the lowest possible Equalized Odds value. A low Equalized Odds value signifies that True Positive Rates (TPR) and False Positive Rates (FPR) are similar across classes, enhancing fairness. After calculating the Equalized Odds value between each pair of classes, the average value is taken for each class. Classes with the highest Equalized Odds values receive additional training, as these indicate a need for improved fairness. Figure 3 displays the outcomes of the Equalized Odds experiment using real training data.

Equalized Odds on Augmented Data

In this experiment, we change the additional training data to the augmented data from the UTKFace dataset. The other hyperparameters like the number of steps and the amount of additional images each step

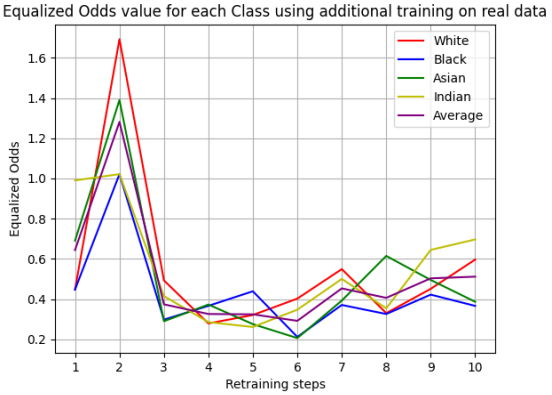

Figure 3: Retraining the model using data derived from the FairFace dataset. The additional training data is selected based on the class with the highest Equalized Odds value, aiming to enhance fairness across all ethnicity classes.

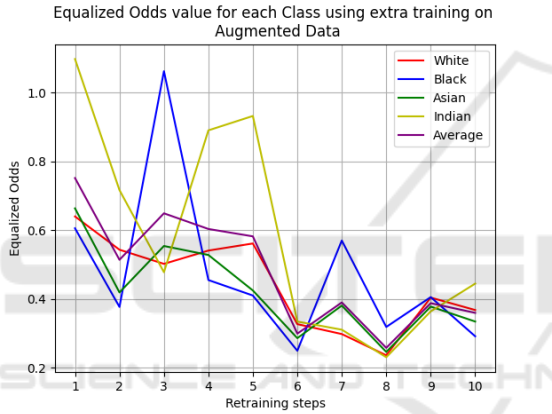

Figure 4: Retraining using augmented data derived from UTKFace based on the highest Equalized Odds value. After the initial training step, it is observed that the Indian class exhibits the highest value, which means that this class receives extra images based on data augmentation for the next training step.

stays the same. Figure 4 presents the results of this experiment.

Equalized Odds on GAN Data

In this experiment Generative Algorithms are used to create new data for the training of the model. This technique can be very helpful when dealing with very small datasets, where a limited amount of images is present. The Equalized Odds value for each ethnicity class is calculated to demonstrate the effect of adding additional training data to the baseline UTKFace training set. The results are displayed in Figure 5. Additional training is based on the highest Equalized Odds value, aiming to balance the Equalized Odds values over all classes. After several epochs,

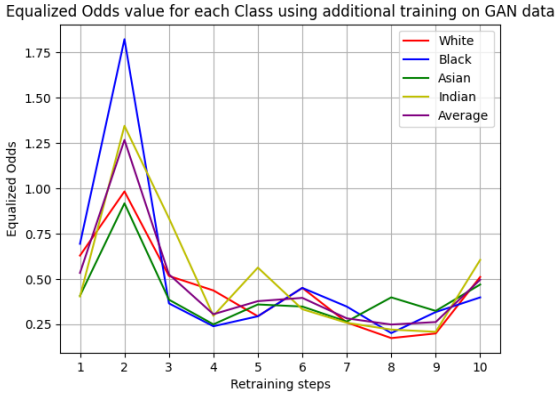

Figure 5: Model retraining with GAN-generated data from the UTKFace dataset, selected to improve fairness by targeting classes with the highest Equalized Odds values.

it is observed that the Equalized Odds values of all classes come fairly close to each other.

4.3 Balancing Fairness Using Different Input Data

This section explores the impact of various ethnic splits from different input datasets on model training. There are differences between the original UTKFace and FairFace models, with the possibility to add images using data augmentation, generative adversarial networks, or simply using data from another dataset.

Baseline UTKFace

A method to balance fairness in machine learning models is to adjust the input data the model uses for training. This experiment examines the effect of altering the ethnicity split in the input data on the fairness of the model. This is calculated using the Equalized Odds fairness metric. As shown in Figure 6, the UTKFace dataset without the 'Others' category performs reasonably well, particularly the 'Indian' class, which exhibits the highest Equalized Odds value. An examination of the ethnicity split of the UTKFace dataset in Table 1 reveals that the 'Indian' category contains only 2146 images, whereas other classes have more images. To address this imbalance, the next experiment involves equalizing the image count for each class to balance the dataset.

Equalizing Image Count in UTKFace and FairFace

For the next experiment, images were added to the Black, Asian, and Indian classes to match the number of images in the White class, which has the most

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

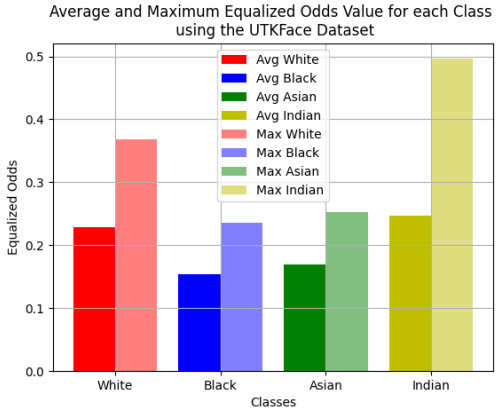

Figure 6: Equalized Odds values of UTKFace dataset without the "Others" category. It indicates that the "Black" class has the lowest average and max EQodds values, making it the most "fair" class in terms of Equalized Odds.

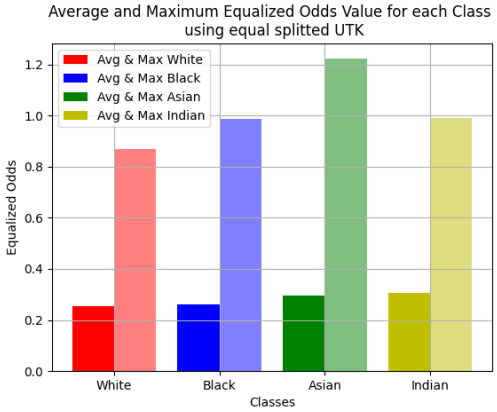

Figure 7: Equalized Odds values of the balanced UTKFace dataset. The average values are averaged over 10 runs of the model, with the maximum value found on the lighter respectable color bars.

images with some 8063 different images. By using an equalized image count for all classes, the potential for bias arising from an unbalanced ethnic split in the input dataset is mitigated. This approach also addresses the issue of the model potentially favoring classes with more images. To augment the UTKFace dataset and enable an equitable comparison with the FairFace dataset, 4000 images of each class were selected, totaling 16,000 images. This required adding images to the Black, Asian, and Indian categories, as detailed in Table 1, and removing excess images from the White category by randomly selecting 4080 images for removal. After integrating randomly chosen images from the FairFace dataset, each class was balanced to exactly 4000 images. The test set, derived from the original test set of the UTKFace dataset, was also balanced by selecting 675 images of each class, ensuring an equal number of test images for each possible classification. The results of this experiment are displayed in Figure 7.

4.4 Improving Fairness Using Additional Training by Adding (n)Generated Images

This experiment examines the effect of adding images from different datasets, augmented images or generated images using a GAN. This is done by adding 1000 images of the class with the lowest Equalized Odds value, to achieve an as balanced as possible fairness result using equalized odds. The experiment is conducted over a single epoch during which the adjusted UTKFace model is trained. After the first epoch, the performance of the model is tested and the fairness metrics are calculated. Based on the lowest

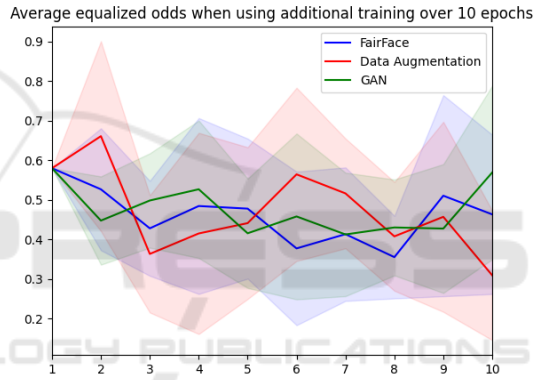

Figure 8: Equalized Odds values when performing different additional training methods to the UTKFace dataset. Each experiment is averaged over 10 runs. The area around the lines is the confidence interval calculated by taking the standard deviation over the 10 runs.

Equalized Odds value, 1000 images from three possible separate datasets are added. The first dataset comprised original FairFace images; the second featured UTKFace images augmented as described in Section 3.2. The third dataset included images generated by StyleGAN from original UTKFace images. All images from these three datasets are split over each class, allowing for images from a single class to be added to the original dataset. The experiment consists of three separate runs, each its own dataset for additional training. Each of the three experiments is averaged over 10 runs, such that randomness is lowered. The outcomes are depicted in Figure 8.

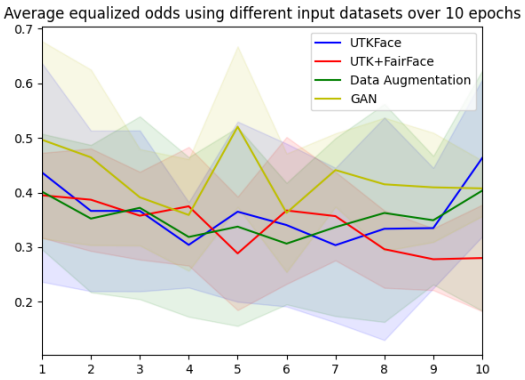

Figure 9: Equalized Odds values calculated over 10 different runs using different input training data sets. The UTKFace dataset has its original amount of images for each ethnicity. The other datasets have 4000 images for each class, totalling 16k images per dataset. This balances out the dataset but loses some of the original images corresponding to the "White" label.

4.5 Balancing Training Data by Adding Images Using Various Techniques

Section 4.4 described adding data to the model based on training performance. In this experiment, different adaptations of the UTKFace dataset are used to balance the training data and improve the equalized odds values from each ethnicity. By balancing the datasets, every class has the same amount of labeled images, such that each class gets an equal amount of training. The goal of this experiment is to create fairness in the UTKFace dataset and to see which of the methods can create fairness. In practice, extra data is not always available, so by using data augmentation or by generating new images using StyleGAN, the process of balancing datasets can still be improved. Figure 9 displays the results.

5 DISCUSSION

The experiments indicate that extending training with new image data improves Equalized Odds across classes, as shown in Figures 4 and 5. The StyleGAN-generated data performed particularly well when original images were limited. However, initial additional training on specific classes (Indian and White, as per Figure 3) led to worse Equalized Odds for other classes. Comparable Equalized Odds values were achieved for all classes when training on augmented data (Figure 4), and notably improved for the Indian and Black classes. Using an equalized image count in UTKFace and FairFace datasets (Figure 7) demonstrated similar Equalized Odds for all classes, albeit "Indian" and "White" showed consistent bias. Equalizing image count effectively curbed the model's tendency to favor labels with a higher image count.

Additional training using varying techniques (Section 4.4) initially decreased Equalized Odds, but these values increased after four epochs (Figure 9). Yet, it failed to elevate the Indian class's Equalized Odds above other classes due to the model's bias towards larger image counts. Results suggest GAN image data, FairFace, and augmented data perform similarly. To understand the effect of GAN data, different data will be added before training to achieve balanced images.

Finally, balanced training data using various techniques revealed the UTK+FairFace dataset as the worst performer, possibly owing to the StyleGAN model's training on unbalanced data. Meanwhile, data-augmented images performed similarly to the UTKFace dataset, suggesting data augmentation as a viable option for balancing and bias reduction. Further experimentations with different fairness metrics are needed to understand the impact of these techniques better.

6 CONCLUSION

This paper investigates techniques to enhance fairness in facial image datasets and models. Through our literature review, we found models often use biased image datasets, limiting their global usability. In response, the study explores adding data during training, with a focus on the class exhibiting the poorest fairness. Our results show that additional training and balanced training data effectively improve fairness. As shown in Figure 9, adding StyleGAN generated images yields the worst fairness results, likely due to the model's inherent bias. Future work should examine whether newer GAN implementations can enhance fairness. Enhancing fairness will likely increase the usage of these models in real applications where decisions should not be based on bias. This emphasizes the need for further exploration and improvements in fairness as AI continues to permeate real-world applications.

REFERENCES

Acien, A., Morales, A., Vera-Rodriguez, R., Bartolome, I., and Fierrez, J. (2019). Measuring the gender and ethnicity bias in deep models for face recognition. In Vera-Rodriguez, R., Fierrez, J., and Morales, A., edi-

tors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 584–593, Cham. Springer International Publishing.

Alvi, M., Zisserman, A., and Nellåker, C. (2019). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In Leal-Taixé, L. and Roth, S., editors, *Computer Vision – ECCV 2018 Workshops*, pages 556–572, Cham. Springer International Publishing.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anonymous (2024). Towards Fairness In Machine Learning - experiments, code and data. https://doi.org/10.5281/zenodo.12672289.

Balakrishnan, G., Xiong, Y., Xia, W., and Perona, P. (2020). Towards causal benchmarking of bias in face analysis algorithms. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 547–563, Cham. Springer International Publishing.

Benthall, S. and Haynes, B. D. (2019). Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 289–298, New York, NY, USA. Association for Computing Machinery.

Bogen, M. and Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity, and bias.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age.

Danks, D. and London, A. (2017). Algorithmic bias in autonomous systems. pages 4691–4697.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In Losada, D. E. and Fernández-Luna, J. M., editors, *Advances in Information Retrieval*, pages 345–359. Springer Berlin Heidelberg.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning.

Howard, A. and Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, 24.

Islam, A. U. (2023). *Gender and Ethnicity Bias in Deep Learning*. PhD thesis.

Kärkkäinen, K. and Joo, J. (2019). Fairface: Face attribute dataset for balanced race, gender, and age. *ArXiv*, abs/1908.04913.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks.

Maluleke, V. H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A. A., Kanazawa, A., and Guillory, D. (2022). Studying bias in gans through the lens of race.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A survey on bias and fairness in machine learning.

Nitzan, Y., Bermano, A., Li, Y., and Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG)*, 39:1 – 14.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Serengil, S. I. and Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.

Srinivasan, R. and Chander, A. (2021). Biases in ai systems. *Commun. ACM*, 64(8):44–49.

W Flores, A., Bechtel, K., and Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Federal probation*, 80.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations.

Zhang, Z., Song, Y., and Qi, H. (2017). Age progression regression by conditional adversarial autoencoder.