Universiteit Leiden

Master Computer Science

Towards Fairness in Machine Learning: Balancing Racially Imbalanced Datasets through Data Augmentation, GANs, and Fairness Metrics

Name:	Anthonie Schaap
Student ID:	s2058081
Date:	26/07/2024
Specialisation:	Data Science
1st supervisor:	Niki van Stein
2nd supervisor:	Michael Lew

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Existing AI models trained on facial images are often heavily biased towards certain ethical groups due to training data containing unrealistic ethnicity splits. This bias leads to significant problems such as inaccuracies and unfair predictions in facial recognition systems. In this research, we investigate methods to recognize and reduce this potential bias. This approach involves the usage of fairness metrics to evaluate fairness and bias. By implementing more balanced AI models, we improve the fairness and accuracy of facial recognition models. The results show a need for fairness metrics to recognize bias, where bias can be prevented by balancing AI models based on the fairness metrics. This research contributes to the practical approaches of balancing datasets in AI systems.

Ethic statement

This research involves using facial images to train and test classification models. These images are mostly taken from real people, which requires privacy and ethical responsibility. We used three well-known facial image datasets: the UTKFace dataset, the FairFace dataset, and the VGGFace2 dataset. The UTKFace dataset is available for non-commercial purposes only. The FairFace dataset is derived from the Yahoo YFCC100m dataset with Creative Common Licences, which permits both academic and commercial usage. All code used for this research is open-source. The images of the UTKFace dataset are collected from the internet and the copyright belongs to the owners of the images. The labels of the images in UTKFace are estimated using the DEX algorithm and double-checked by a human annotator. Inaccuracies and labeling mistakes in the datasets are present. The VGGFace2 dataset is available for non-commercial research purposes only. When discussing ethnicities in this paper, we acknowledge the importance of fairness in AI in a manner that respects and values diversity. Our study represents only a fraction of the diversity of ethnicities present throughout the world.

Contents

Chapter 1

Introduction

1.1 Racially imbalanced datasets

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 often have unwanted biases for several reasons. An AI system can be biased when the dataset it's trained on has one or multiple biases [34, 13, 11]. These biases can have a very negative impact when used in practice, where most human-like biases are present due to an unfair ethnicity, age, or gender split. Dataset bias can vary from a broad range of uneven aspects, like age distribution, ethnicity distribution, image consistency, and even some gender stereotypes which can label a person as a "woman" when the background contains a kitchen [36]. It is essential to tackle the problem of model bias, which can impact lives when used in practice, for example when hiring people for a job using AI [8] or when AI is used to calculate the chance a defendant recommits a crime [25]. In this paper, we are going to look at ethnicity bias in datasets and AI models. To recognize ethnicity bias in datasets, first, we need to train an AI model to classify different ethnicities. To train this AI model, we need to mitigate the ethnicity bias as low as possible, such that we have a small bias toward classifying different ethnicities. Datasets for ethnicity classification can contain a lot of unwanted biases, and are often unbalanced, as seen in Section 3.1. In this research, we want to investigate the ethnicity bias in datasets not used for ethnicity classification but for different 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. This can cause several problems, such as the blunder of Google in 2015 when their image classifier in Google Photos classified two Black individuals as "Gorillas" [3]. Historically, many problems such as the Google Photos incident were coupled with people of color, dehumanizing them and mislabeling them.

1.2 Research questions

In this research, we address the following research questions:

1. To what degree do existing face datasets and AI systems show biases related to ethnicity?
2. How can we improve fairness within facial datasets 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?

1.3 Thesis overview

Following the introduction of these key concepts discussed in this paper, we discuss similar areas of research related to this topic, found in Section 2. After reviewing several state-of-the-art papers related to ethnicity bias and explainability, we propose our methods in Chapter 3. Here we look at the datasets that are used for this research together with the methods used in this project. In Chapter 4 we look at the different experiments performed in this paper, like the training of the ethnicity classification model and the usage of the AI model on several state-of-the-art researches. In Chapter 5 we look at the results of the experimentation, comparing our model to state-of-the-art models and experimenting with the model on several projects. After showcasing the results of the project, we discuss the results and interpret the results in Chapter 6 In this chapter, we discuss the results of the paper and investigate the differences and comparisons with other state-of-the-art papers. We also investigate the impact this paper can make on new research to tackle the problem of ethnicity bias in artificial intelligence models. Lastly, we conclude this paper with some limitations and suggestions for future work in Chapter 7.

Chapter 2

Related Work

This chapter summarizes the events of recent years in the fields of ethnicity detection, ethnicity bias, explainable AI, and different explainability methods. This is done through an extensive literature study, which reviews the state-of-the-art progressions that have been made in these fields. Additionally, we are going to look at some state-of-the-art progressions that have been made in the field of deep neural networks.

2.1 Ethnicity Bias in AI

Numerous studies explore the different ethnic biases present in AI systems. In Angwin et al. [4], the authors discuss that computer algorithms that are used as risk assessments in the criminal justice system can be biased towards black individuals, where the systems predict on average more severe crimes for people with black skin and less severe crimes for people with lighter skin. This is done by evaluating the risk scores of 7000 arrested people, where at first glance already a big bias is present. In the predicted group that would commit violent crimes, only for 20% of the arrested people, this would become true. Furthermore, the authors describe that when the same crime is committed, the prediction system gives higher risk scores to black individuals than to white individuals. However, in Flores et al. [35], the authors discuss the research of Angwin et al. to be based on faulty statistics and that it failed to show racial bias.

2.2 Ethnicity Bias in Face Recognition

The primary driver of the progress of an AI model is its dataset. In the case of facial images, datasets can reflect humans. When using racial categories like white and black, the complexity of the human race suggests that in between different datasets, there can be different perceptions of racial classification [23, 18]. There are many cases where racial background is from multiple countries, which makes it quite hard to label in different datasets. If different datasets have different concepts of race, then making one dataset less biased might not work for the other databases. In the paper of Khan et. al [23], we can see that racial categories rely mostly on stereotypes, instead of a true sample of different societies. Eliminating bias in a dataset is a difficult task, especially because each individual is different than the other, but by having an equal amount of pictures of different ethnicities, maybe the AI models will be less biased when making a classification. To battle the lack of ethnicity diversity in databases, name-ethnicity classifiers (NECs) can help to identify the diversity of a dataset [15]. NECs have a bias, which

is that not every name from a person determines their ethnicity. Still, the authors of the paper believe the contribution of NECs to discovering ethnicity bias in datasets outweighs the bias of NECs themselves. Another method to mitigate bias is by detecting segregation patterns in data sets, AI models can learn to mitigate social segregation, such that there are no group disparities between different ethnicities (Benthall et al. [7])
Datasets can be biased on multiple aspects, for example in datasets where the age distribution is unbalanced, the AI model might learn from these existing biases. To eliminate or mitigate these biases, Alvi et. al [2] introduce a method to eliminate bias from trained models on biased datasets. When a model is trained on a biased dataset, eg. with an imbalance in age distribution, the model will perform a lot worse on unbalanced test data. To improve the model, we need to unlearn the age bias from the model, such that the model will not make its decisions based on age for gender classification.
AI can stimulate growth and efficiency in various fields, but because AI's are written by humans and are trained on datasets made by humans, bias is common. Even when these biases are small, many small biases in various fields can lead to inequality between different people based on age, gender, and ethnicity. Although there exist mitigation measures such as nondiscrimination and data protection laws, these measures are still non-binding [9]. Sector-specific mitigation of bias is essential in challenging the discriminatory part of AI. A very important task in mitigating bias in datasets is to classify each ethnicity and gender with around the same accuracy, such that each person is treated fairly. A study that highlights these concerns from Buolamwini et al. [10] shows that on two different facial analysis benchmarks (IJB-A and Adience), the datasets are heavily skewed towards light-skinned individuals, where the worst error rate for classification is found for dark-skinned females. For creating facial analysis models that can be used for commercial companies, it is crucial that the datasets the models are trained on should have a fair age, gender, and ethnicity split. This fair split should help with a more equal classification for people with different backgrounds. Multiple ethnicities are possible for a single person, or a different gender than only "male" or "female", where very little research is done. Databases like the UTKFace and FairFace datasets 3.1 only have binary genders included, where there is already a huge bias towards trans people. In the research of Keyes et al. [22], alternatives for Automatic Gender Recognition systems are discussed, with the goal of trans-inclusive treatment of gender, which is still very lacking.

2.3 Ethnicity Classification in Face Recognition

For ethnicity classification, much data is needed. A lot of the time models are made capable of multiple classification and regression tasks, like age estimation, gender classification, ethnicity classification, and facial attribute analysis. A state-of-the-art facial recognition framework based on many state-of-the-art models like VGG-Face and Google FaceNet is the DeepFace framework [30]. This framework can perform numerous tasks on facial images, but for this research, we focus on the ethnicity detection of the framework. Another high-performance model for ethnicity classification is the Resnet50 pre-trained model, which is used in the work of Acien et al. [1]. The importance of ethnicity attributes is discussed by using VGGFace and ResNet50 pre-trained models and training a classification layer that is connected to the embedding layers of these models. The work of Acien et al. suggests that these pre-trained models can perform very well for gender and ethnicity classification in face recognition.

2.4 Mitigating Bias in AI models

To mitigate AI bias in general, identification methods are needed to identify the potential biases in AI models [16]. In the work of Raji et al., [28], a framework is presented to identify potential bias in AI models which can greatly improve the accountability of AI models. The framework works by first setting a system scope, where a specific model is determined by its potential impact on their user base. Then, the system is tested in the Pre-launch Audit where the model is tested for potential biases. In the Post-launch Audit, the model is evaluated when it is already in use after the model is deployed. Lastly, the Feedback Loop where the discovered biases in the previous steps are addressed such that future models are informed and the current model is adjusted accordingly. The framework is meant to promote AI models to enhance accountability for their user base. Bias in autonomous systems is not always due to the datasets containing bias, the structure of the AI model can also induce potential bias (Danks et al. [12]). Increased attention to algorithmic design is needed next to robust testing to mitigate the potential bias in an AI model.

To detect bias in machine learning models, many techniques are possible. For example, algorithmic bias can be detected using synthetic faces, which can indicate some of the biases present in a machine learning model [6]. For example, when a male tends to have longer hair, often they also have more facial hair, or when females have earrings they get classified more accurately. These aspects can be learned by machine learning models which are not often visible without detection methods.

2.5 Mitigating Bias using Generative Adversarial Networks

We can try to add training data to underrepresented groups to mitigate bias in an AI model. This paper looks at the underrepresented ethnic groups that cause the AI models to learn bias. A method to add training data to the underrepresented groups can be done using Generative Adversarial Networks (GANs), which create new data that resembles the training data it's trained on. The Generative Adversarial Network was first proposed by Goodfellow et al. [14]. In the implementation of the GAN, two models are created that compete against each other, being the generative model and the discriminative model. The goal of the generative model is to optimize the probability that the discriminative model classifies wrongly. The task of the discriminative model is to classify if a sample was from the training data or the generative model. This framework made significant contributions, such as image generation, data augmentation, and face editing [32, 31]. A modern approach to the GAN is the Progressive Growing of GANs, proposed by Karras et al. [21]. In this implementation the generator and discriminator are trained simultaneously, using progressively higher-resolution images. This new technique proposed by Nvidia researchers improves on the classic GAN, with better image quality, better stability, and more usage possibilities.
New research shows that Generative Adversarial Networks can also be heavily biased, which can cause results to contain biases that are present in the training data [24]. In theory, GANs can be of great importance in expanding training data and providing high-quality image data, but when bias is present, the GANs might not be able to generate high-quality images from underrepresented ethnic groups.

Recent research by Karras et al. [20] proposes the StyleGAN network, a generative adversarial network that excels at creating photorealistic images. StyleGAN introduces a new generator approach, which splits the high-level attributes of an image from the lower attributes. StyleGAN uses a new concept of "styles", where the model can approach and adjust different attributes from several possible details, for example, the pose and smile of a person. StyleGAN also improves the quality of state-of-the-art image generators, using its unique model architecture. Building on this research, Nitzan et al. [27] propose a method based on StyleGAN where the face identity of an image can be disentangled without altering the image attributes. The method of Nitzan et al. takes an identity image set and an attribute set, in which the identity of the image set is kept, but with attributes from the attribute set such as pose, expression, and illumination. For this research, we make use of the StyleGAN implementation of Nitzan et al., to create more labeled data examples for each ethnicity group.

2.6 Explainable AI

Neural network models can achieve great things in all sorts of topics, sometimes achieving very accurate predictions on data it's trained on. It has great potential in improving processes or techniques. The downside of neural networks is that they don't explain **how** a decision is made. We can see the accuracy of the model on certain test data, but we cannot know how the network has come up with its conclusion, making a model hard to trust. Especially in topics like legal issues or in medical studies or diagnoses, the need to validate and justify the decision that is made by the network is crucial. This problem is what is called a "black-box" problem, we know the inputs and outputs of a neural network but we don't exactly know how the network learns and on what features the output is based on. To tackle this issue, the need for understanding these networks is growing. This is called XAI (Explainable AI), which is a research topic to study and research the characteristics of a neural network and to improve our understanding of interpreting the decision-making ability of a neural network [26].

2.6.1 XAI in image classification

In image classification, the problem of explainability looks to be less of an issue at first glance, because if we look at a picture of a cat and the neural network classifies the picture as a cat, we understand that the network made the right decision. However, we do not know where the neural network based its decision. It might be that the training data has coincidental similarities, like background colour or certain shapes in the background. Explainability for AI methods enabled transparency, and believability and helps in validating the model. For ethnicity detection, it is very important to have an explainable model, such that we create transparency in the decision process of the model, to investigate if possible biases are present in the decision process. The research of Schorr et al. [29] proposes a toolbox that can be used to visualize every layer in a neural network, such that we can understand the key features of the model and where the model bases its decisions.

2.6.2 Saliency Networks

In this paper, we use an open-source XAI toolkit by Brian Hu et al. [17] that can be used for saliency interpretation. This toolkit can be used to help understand the database and model for image-based machine-learning tasks. The toolkit can provide a great insight into the explainability of the black box features of the machine learning tasks. The toolkit can be used on the input and output data of a model, resulting in a visual interpretable saliency map. Saliency networks can also be used for video analysis, creating dynamic saliency [5]. Dynamic saliency prediction can give a great insight into which parts of a picture or a video are important to detect the main attention point. In this paper, the authors propose a spatio-temporal saliency network using deep learning that is used for picture and video analysis, with great results. Especially in the video analysis, the model outperforms the state-of-the-art and gives a better understanding of important aspects of a picture.

Chapter 3

Proposed Methods

3.1 Datasets

In this paper, we experiment with different datasets suited for ethnicity classification. The first dataset is the UTKFace dataset, introduced by Zhang et al. [37]. The UTKFace dataset contains over 20,000 images, with age, gender, and ethnicity labels. This dataset can be used for facial image detection, age estimation, age progression, and also ethnicity detection. We use this dataset mainly for training the ethnicity detection model. Some of the important characteristics of the UTKFace model are the labels.

- Age: integer between 0 and 116, representing the age of the person
- Gender: either 0 for male and 1 for female
- Race: Integer from 0 to 4, being White, Black, Asian, Indian, and Others respectively.
- timestamp: time in the format of yyyymmddHHMMSSFF, showing the timestamp of an image when it was added to the UTKFace dataset

The most vital aspect of the dataset to have for this research is the bias. The bias should be as low as possible, meaning an evenly split dataset in all labels, but especially in the racial composition split. We can see the racial composition split in Figure 3.1.

As we can see from Figure 3.1, 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. When creating an AI model, if there is bias present in the training data, the AI model will have biased characteristics. It is nearly impossible to have a perfectly unbiased dataset, but the goal is to mitigate the bias as much as possible.
Another interesting observation from Figure 3.1 is how biased most of the state-of-the-art datasets are. We can see that in all datasets except the FairFace dataset, there is a high ratio of "white" facial images in the dataset. This can cause numerous problems, for example, if we want to estimate the ages of the facial images in a dataset, the AI learner might be better at classifying the images with the "white" racial label, because there is more training data available for that specific racial label.

To tackle this problem, we also experiment with the FairFace dataset [19] which gives 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 and has race, gender, and age labels. The racial composition split of the FairFace dataset can be seen in Figure 3.1.

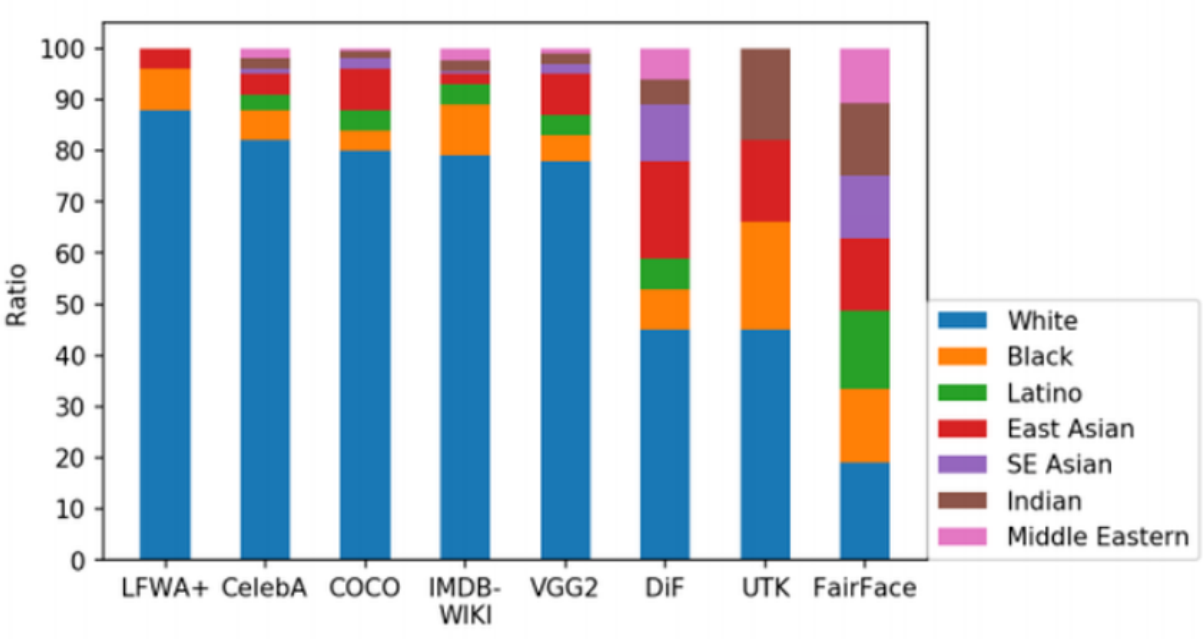

Figure 3.1: Different racial compositions in face datasets from Kärkkäinen et al. [19]

Table 3.1: Ethnicity Split in UTKFace Dataset

Ethnicity	Count
White	8080
Black	3621
Asian	2759
Indian	2146
Others	1358

3.2 Data Preprocessing

In the UTKFace dataset, we have found some incorrectly labeled images, where the ethnicity is not correct. These images have been deleted from the dataset to enforce the quality of the learning process. We also found some errors in the names of the images, where six images with a missing ethnicity label have been removed. Some images also were of very poor quality or did not contain a face at all, with some images of a single eye present in the dataset, or some handwritten faces on paper. It is a manual task to discover imperfections in the image dataset, which still leaves room for errors to be present in the dataset. We also found that the age

Table 3.2: Ethnicity Split in Fairface Dataset

Ethnicity	Count
White	16527
Black	12233
Asian	23082
Indian	12319
Others	22583

split of the UTKFace image dataset is highly biased in terms of age distribution, with most samples being in the range of 0–4 years old. This skewness might cause the model to learn features from that specific age group, potentially leading to unintentional biases in ethnicity classification. For this research, we did not remove any correctly labeled images, to keep the integrity of the UTKFace dataset.

3.3 Model Architecture

The model is based on a pre-trained FaceNet model, which is an advanced CNN (Convolutional Neural Network) model designed for facial recognition tasks. The model takes as input any amount of pixels in height and width, such that it can be used on different models. It also takes a color channel, based on three values red, green, and blue. Using transformer learning, the output of the FaceNet model is transferred to a new model, which starts by flattening the output of the FaceNet model to convert the 3D outputs into a 1D vector. This vector is used for three dense layers, with 256, 128, and 4 neurons respectively. The first and second layer uses the 'ReLu' activation function. After these layers, we use a batch normalization layer to normalize the activations of the previous layers at each batch, which can help in improving performance and stability. The last layer consists of 4 neurons and a 'softmax' activation function, used for multiclass classification. The 4 neurons correspond to the 4 ethnicity groups we use for this research.
The model is compiled using Adam optimizer, with a small learning rate of 0.005. The loss function is a "categorical cross-entropy", which is suitable for our multiclass classification task. The model is trained in batches, to ensure stability and usability on devices with limited hardware resources. The (validation) loss and (validation) accuracy are evaluated to ensure the training of the model. For the layout of our model architecture, a flowchart is made to illustrate the training and evaluation of our model in Figure 3.2.

3.4 Training of the Ethnicity Classifier

For the training of the ethnicity classifier, it is important to make GPU training available for the program to use. This greatly reduces training time when dealing with large datasets of images. In this experiment, we made use of an NVIDIA RTX 3060 graphics card.

3.4.1 Training on the UTKFace dataset

For the UTKFace dataset, the labels of the images are available in the name of each file, annotated with 0 for White, 1 for Black, 2 for Asian, 3 for Indian, and 4 for Others. The labels are present in the filenames of the images, where the file names are formatted as $[\text{AGE}]\_[\text{GENDER}]\_[\text{RACE}]\_[\text{DATE\&TIME}].jpg$. To list the ethnicity labels, we split the filename on the "race" label and placed it in a list. After the list has been made, we create a Pandas data frame to match the images and their respective label together. For the images, we load the images and scale them to 128×128 in "RGB" colour mode, which is converted to a NumPy array such that we get an array of shape $(18964, 128, 128, 3)$, with 18964 images of size 128 by 128, and 3 for the color pallet of each pixel. For the training of the model, we make use of a learning rate schedule function, which adjusts the learning rate to be smaller after some epochs, such that the model can perform exploration of the best parameter values, and after

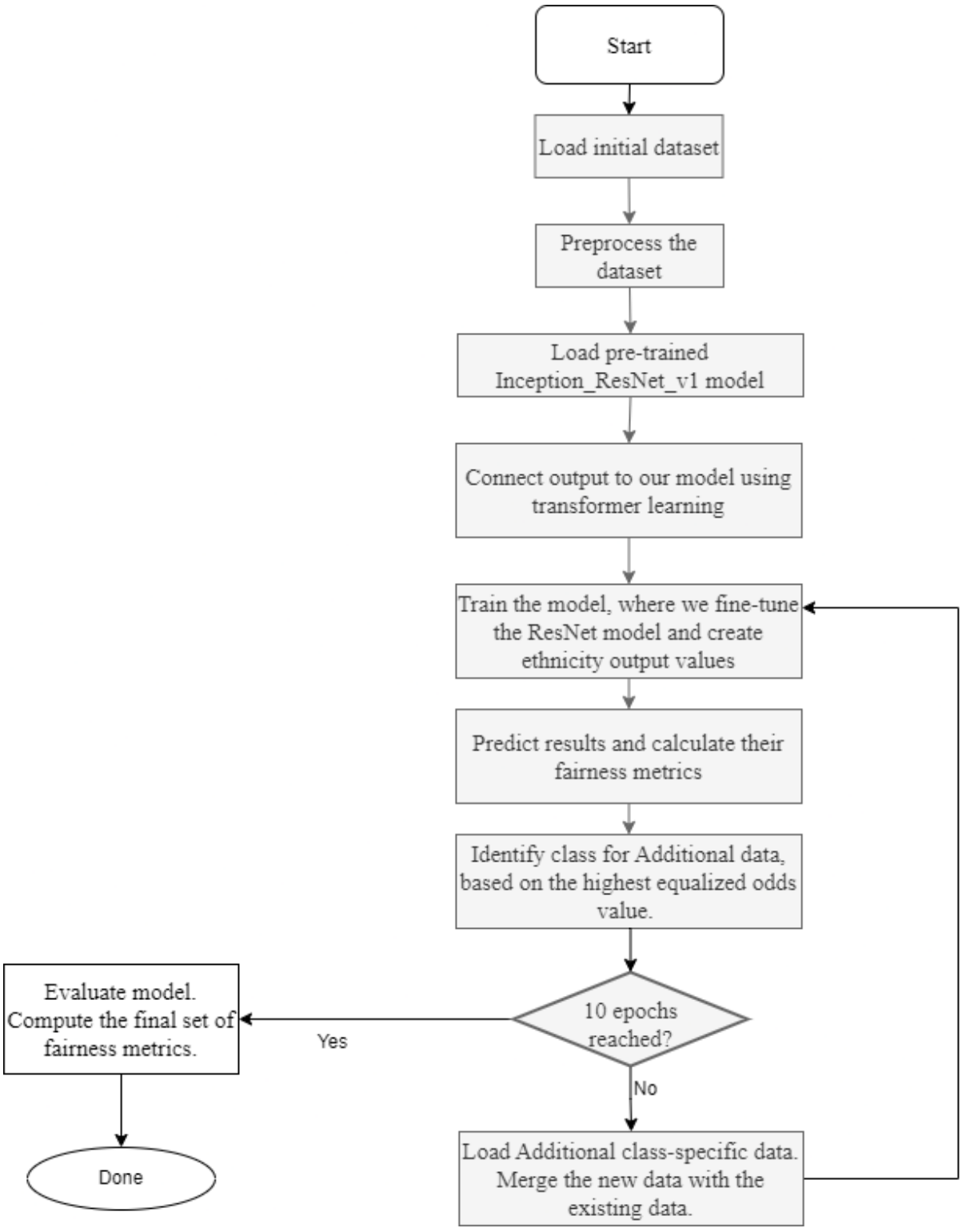

Figure 3.2: Flowchart of the Model Architecture.

some epochs use a smaller learning rate such that we perform exploitation and try to optimize maximally. To increase the training data, a data generator is used to generate more images for the model to train on. This is done by taking an image from the training set and moving the picture around horizontally and vertically, such that we get a slight alteration of the original

image. We also rotate the images and flip horizontally, to create a lot more training data for the model to work with.
The model is built by first using the FaceNet pre-trained model, where we freeze the layers and connect the output of the model to a second model. This model uses three layers, from which one is 512 nodes using the Relu activation function, the second is 256 nodes using the Relu activation function and the last layer is a Softmax activation layer for 5 possible outputs, such that we get values for each possible classification option. The optimizer used is Adam, with a small learning rate of 0.005. The loss is computed using the categorical cross-entropy, which is best suited for categorical classification tasks. Training is done in batches of 16 images, with 20% of the training data reserved for validation. If the model does not improve over 5 epochs, we use early stopping to stop the model. In Figure 3.3 the training process over 10 epochs is shown. In Figure 3.4 we can see the loss of the model using the UTKFace dataset.

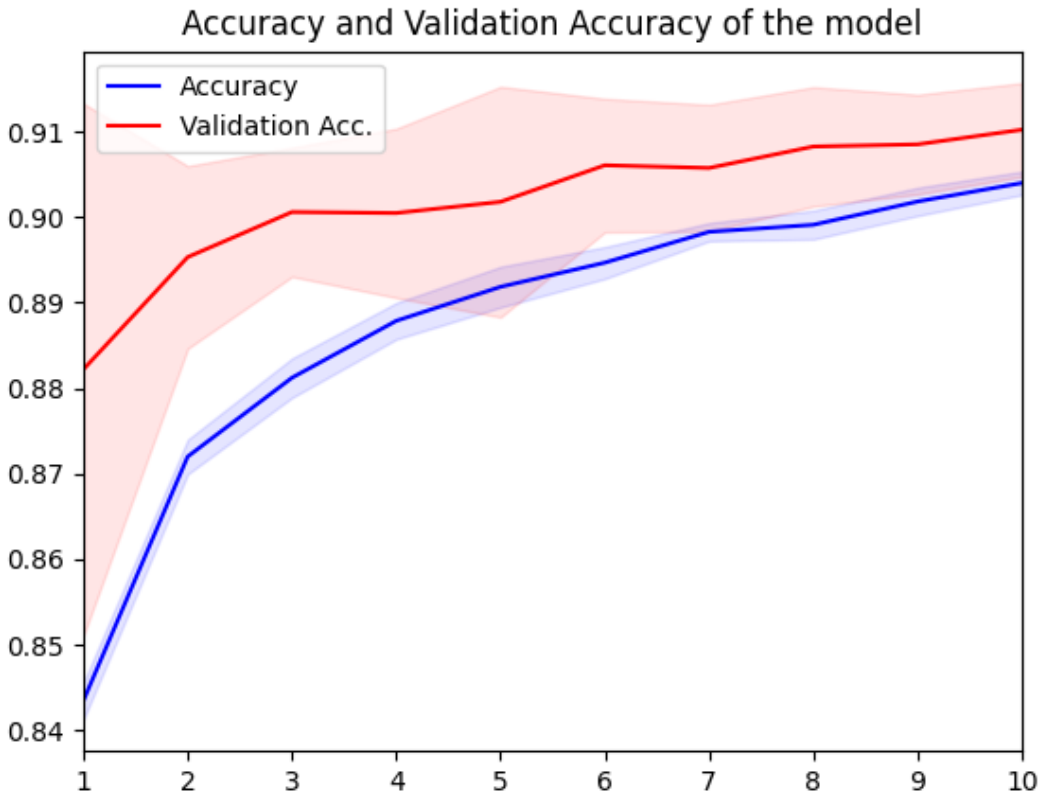

Figure 3.3: Model trained on the UTKFace dataset using 10 epochs

3.4.2 Training on the FairFace dataset

The FairFace dataset contains two comma-separated files (.csv) which contain the name of the file, the age, gender, race, and service test. One file is for the training set and one file is for the validation set. The FairFace dataset contains a total of 7 different ethnicities, annotated with 0 for White, 1 for Black, 2 for Southeast Asian, 3 for East Asian, 4 for Indian, 5 for Latino Hispanic, and 6 for Middle Eastern. Because we want to compare both datasets with each

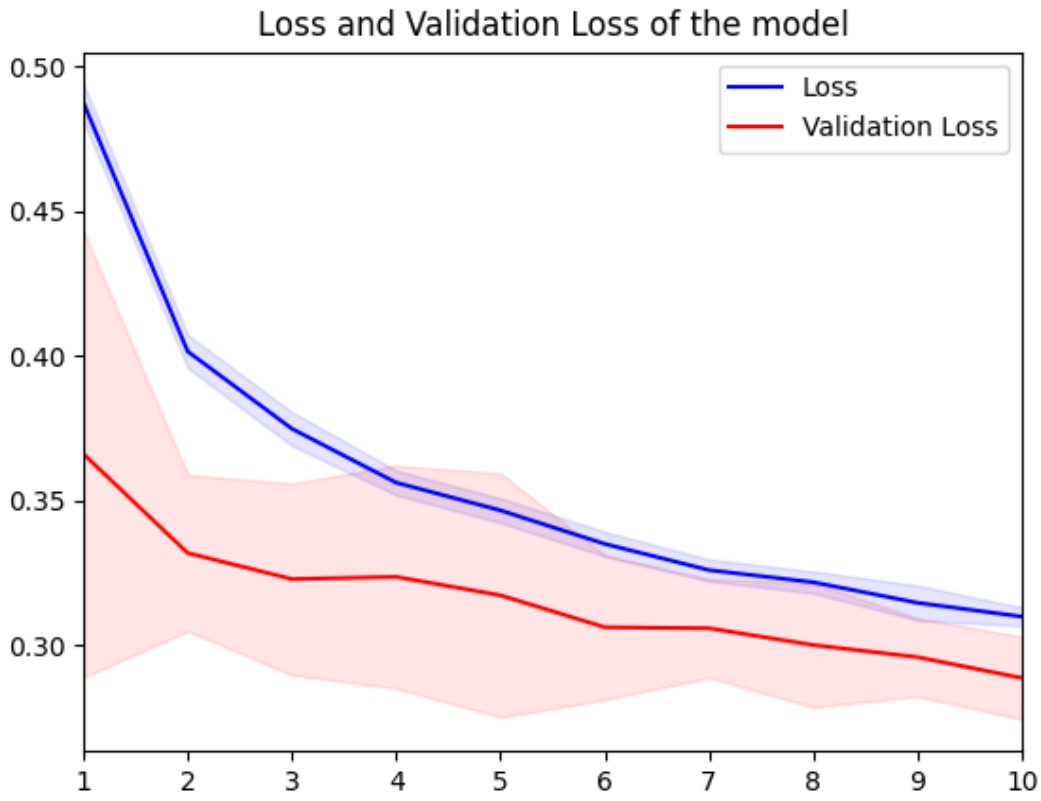

Figure 3.4: Loss of the model using 10 epochs

other, we combine the Southeast Asian and East Asian category as one ethnicity, "Asian". We also combine the Latino Hispanic and Middle Eastern group with the "Others" group, to match the labels of the UTKFace dataset. To work with both datasets and use training data for the other dataset we have changed the labeling structure of the FairFace model to match the structure of the UTKFace model. This makes it easier to train and test on both datasets without the need for different data loaders.

3.5 Data Augmentation

To enhance the size and quality of our datasets, we use data augmentation to generate more training data [33]. 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. Adjusting the images can be done in various aspects, like moving the image to any direction, rotating the image, flipping the image, cropping, and shifting the image to some direction. For colored images, there are also methods to adjust the color scheme of the image, such that the model does not learn based on a specific color style that may be present in the original dataset. For this research, we want to keep the augmented images as realistic as possible, e.g., not flipping the image horizontally or rotating it too much. Additionally, we aim to save the augmented images to allow for manual quality control and future reuse. In Table 3.3 the data augmentation methods for this research are shown.

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

Table 3.3: Augmentation Methods Used to generate more realistic images

Dataset	Image Count
FFHQ 256x256	70.000
FFHQ 1024x1024	70.000
VGGFace2	3.31 million
Celeb-500K	500.000
300W	300

Table 3.4: Datasets used to train the StyleGAN model [27]

The augmentation values in Table 3.3 are randomly chosen, which makes it possible that the same image is generated from the original image. The chance of this happening is extremely small, and the impact would be neglectful. The images are not zoomed out, because that would create borders around the image which are not part of the original image. There is also no image shifting, because of the same problem of creating borders around the image and losing too much information. The augmented images are labeled the same way the original images are labeled, except with a new creation timestamp and a random age value, which is not needed for this research.

3.6 Creating training data using StyleGAN

Another method the increase training data is by generating new images that are based on the attributes of the original images. This is done using an implementation of StyleGAN [27], 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. These images can be used to expand the size of the training data, which can lead to an improvement in stability and quality and 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. The generated images are labeled according to the original identity image, the ethnicity should be the same for the generated image when it has been generated from a particular labeled image. In this research we also make sure the attribute image is from the same ethnicity label, to increase the odds the generated image belongs to that particular label class.
The StyleGAN implementation by Nitzan et al. [27] is trained on several image datasets, which are used to train the model. In Table 3.4 the datasets are shown for which the model is trained.

This StyleGAN implementation uses a mixture of different pre-trained models, which are trained using one of the datasets in Table 3.4. Because the models are already trained, we can

(a) “White” generation

(b) “Black” generation

(c) “Asian” generation

(d) “Indian” generation

Figure 3.5: Some generated images from the StyleGAN model, using UTKFace images.

directly use the StyleGAN implementation to generate more images. In Figure 3.5 some good quality images are shown which are used in this research, one for each label.

3.7 Fairness Metrics for performance evaluation

To be able to evaluate the performance and fairness of the model, different fairness metrics can be used. A simple metric to evaluate a model is the prediction accuracy of the model for each class. This gives an insight into the amount of correctly classified true positives for each class. However, in a scenario where the model is predicting a single class 100% of the time for all possible classes, the accuracy would be 100% of that particular class but the model might still be very bad in classifying that particular class. This makes it also very important to look at the false positives, true negatives, and false negatives of the predictions. In this section, we look at different fairness metrics used in this paper.

3.7.1 F1 score

The F1 score is a metric that can be used to evaluate the performance of the model. For each label, we calculate an F1 score, based on the predictions of the model. The F1 score is a harmonic mean of Precision and Recall, such that the score gives a balance between precision and recall. To calculate the F1 score, we first need to know how the precision and recall are calculated. The precision is given by:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP is the true positive and FP is the false positive. The recall is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN is the false negative rate. Using Precision as P and Recall as R, the F1 score is given by:

$$\text{F1-score} = 2 \times \frac{P \times R}{P + R}$$

The F1 score can be read as a value between 0 and 1 with 0 being a random model on average, and 1 for a perfect model which predicts each classification without error.

3.7.2 ROC AUC

The Receiver Operating Characteristic Area Under the Curve (ROC AUC) is a fairness metric often used in classification tasks. The ROC AUC metric gives information about the capability of the model to distinguish between classes. As a fairness metric, ROC AUC is useful because a high ROC AUC value tells us the model works equally well for different classes, which is needed in this research. The ROC AUC formula is given by:

$$AUC = \sum_{i=2}^{n} (FPR_i - FPR_{i-1}) * TPR_i$$

In this formula, the FPR is the False Positive Rate and the TPR is the True Positive Rate. The ROC AUC is a useful fairness metric but should be considered in a group of different fairness metrics. Most of the time one fairness metric is not enough to sketch the real fairness of a model.

3.7.3 Equalized Odds

The Equalized Odds fairness metric is used to examine if a model provides equal treatment to all classes. It examines both the true positive rate and the false positive rate, where it compares the false positive ratios of all different classes.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where TPR and FPR are the True Positive Rate and the False Positive Rate, respectively. The True Positive Rate is also known as the Recall. In binary classification, to calculate the Equalized Odds metric we can calculate the TPR and FPR of both classes, eg. class 1 and class 2. Then we can calculate the Equlized Odds difference by using:

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

In this formula, we take the biggest difference between the true Positive rates (TPR) or false positive rates (FPR) between the two classes. As a fairness metric, we want the different classes to have as small as possible differences in their true positive or false positive rates. Therefore, the equal odds criterion is more satisfied the lower the equal odds difference becomes.
In this research, we make use of multiple classes. For calculating the Equalized Odds metric, we take the equalized odds difference of all class combinations. Then we take the average value between the classes, to get the average equalized odds difference of the whole model. The equalized odds metrics punish models in which the classes have big differences in terms of recall and the false positive rate.

3.7.4 Disparate Impact

The disparate impact is a fairness metric that takes into account that AI models can be biased toward certain groups of people. As a result, the models might discriminate against them. The disparate impact is defined by the ratio of the positive predictions for one group divided by the positive predictions of a different group. For this research, we consider the 'Indian' group as the protected group, where we calculate the ratio of positive predictions of the Indian group compared to the ratio of positive predictions of the other classes combined. A disparate impact value smaller than one means that there is a potential bias toward the protected group, whereas a value bigger than one means that there might be bias in favor of the protected class. The lower the disparate impact value, the more potential bias there is for the protected class. Ideally, we want the disparate impact to be 1, meaning that the potential bias is as low as possible for the protected group. The formula for the disparate impact is given by:

$$DI = \frac{\text{Unfavorable outcomes} \div \text{Protected Group}}{\text{Unfavorable outcomes} \div \text{Unprotected Group}}$$

Where the unfavorable outcomes are the false predictions of the classifier for a given group. Using the disparate impact fairness metric, we try to increase the fairness of the model by trying to balance the distribution of positive outcomes across all classes. The disparate impact gives insight into the bias there is towards certain classes, which shows the need for balancing fairness throughout the model.

Chapter 4

Experimental setup

In this chapter, we discuss the experimental setup of this research. We discuss the preparation of the different datasets used for this research, together with the baseline performance of the model used for experimentation.

4.1 Data Preparation

To use the datasets and train the model, some data engineering is needed to match the data structures of the different image datasets. For the data preparation, the structure of the UTKFace dataset is used, which can be seen 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 this has been done, we need to split the images based on the ethnicity class in separate maps, so that additional training can be performed using specific ethnicity classes. This splitting process is also done for the newly created images using data augmentation and generative networks. For the generative images, we use the StyleGAN implementation to save the newly created images based on its ethnicity. After the ethnicities of the created images have been split, we change the name of every image in the correct style of UTKFace. This has been done by giving a random age value, which does not matter for this research, and the correct ethnicity label. For the timestamp, the current time is used with a small delay between the creation of each image such that images cannot get the same name when images are created in milliseconds. This gives a good UTKFace structure where images are randomly sorted throughout the map due to the random age value. Each implementation can be used as a separate dataset or used for additional training with specific ethnicity classes.

4.2 Ethnicity detection using UTKFace dataset

The initial experiments utilize an ethnicity detection model trained on the UTKFace dataset. As illustrated in Figure 3.3, the training process results in a validation accuracy of approximately 91%.

The loss of the training process is displayed in Figure 3.4. We can see that the validation loss is steadily decreasing, where we find the best validation loss with a value of around 0.30. In the case of ethnicity bias mitigation, high accuracy is not the goal. More importantly, we want to have an even split on accuracy on all ethnicity labels used in the models. When the accuracy of the model is high, it might still be the case that for example the "White" label is

predicted very accurately, but that some other labels like "Indian" are predicted very poorly. If this is the case, the model is not suited for ethnicity bias detection in other datasets, if it is not capable of detecting each ethnicity with a comparable accuracy. To visualize in a single image the accuracy of each ethnicity label, we make use of a confusion matrix, which can be seen in Figure 4.1. This visualization method works great to get an understanding of what the model predicts if it is wrong. It is important for the model to not default to a single ethnicity label in case the model predicts wrong.

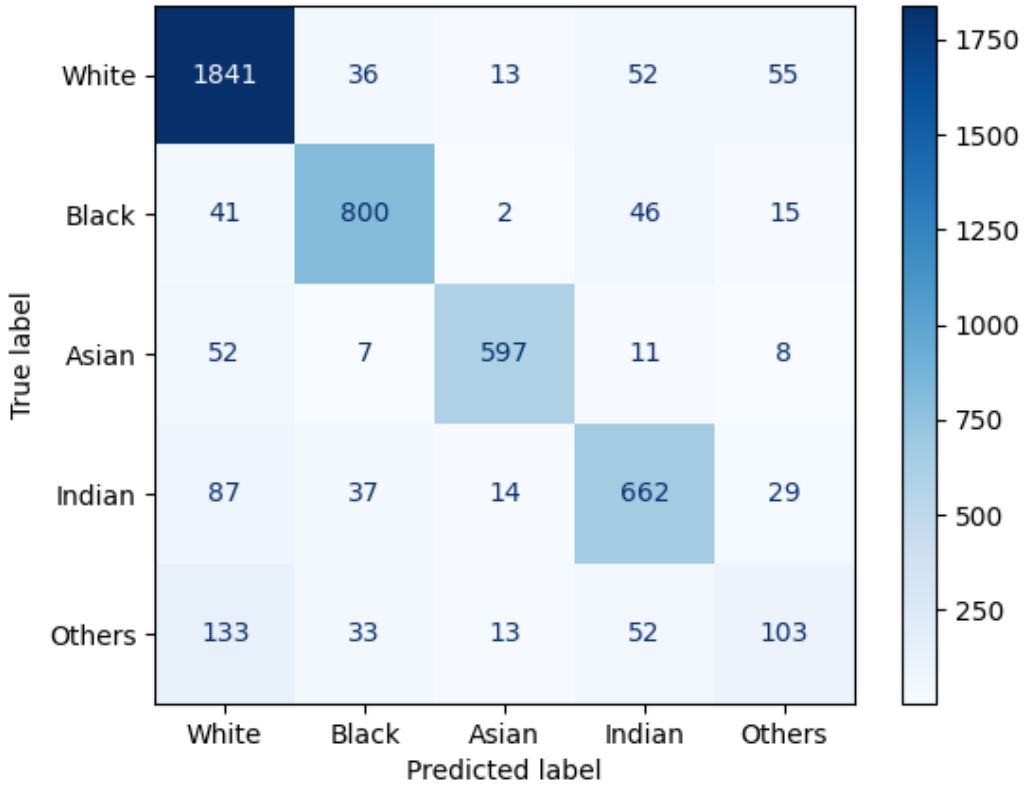

Figure 4.1: Confusion Matrix of the model trained on the UTKFace dataset

In the confusion matrix in Figure 4.1 we can see that the most common label in the UTKFace dataset is the "White" label, with 1841 cases to be predicted correctly. We also see that the label that performs the worst is the "Others" label, which can be argued to be the least important because many possible ethnicities can be present in the "Others" label, making it almost impossible for the ethnicity classification model to classify these images correctly. We can also see for example that when the "Black" label is predicted wrongly, the most frequent choice is the "Indian" label, which is probably the closest match based on skin color alone. For the "White" label, it was expected that "Asian" to have the second highest probability, given the similarities in skin color. The downside of the UTKFace dataset is also clearly visible in Figure 4.1, because of the uneven ethnicity split the model is more likely to classify "White" in case an image is unclear because the "White" label is the most common. It is also interesting and helpful to get a visual representation of individual cases, to see how images are labeled in the dataset and how they are predicted. This gives an insight into why a model might predict a

certain way and in which cases the classifier makes mistakes. For example, in a black-and-white image, we lost important data with the color not present, so it might be interesting to see if the model is still capable of classifying correctly. It is almost impossible to look at every single image in the dataset to see how it was predicted, but a sample can already be very helpful. Datasets are keen to have some instances that might not be labeled correctly or have a bad quality image, which all can impact the results of the model. In Figure 4.2 we can see a batch of 25 images with their true label, taken from the UTKFace dataset.

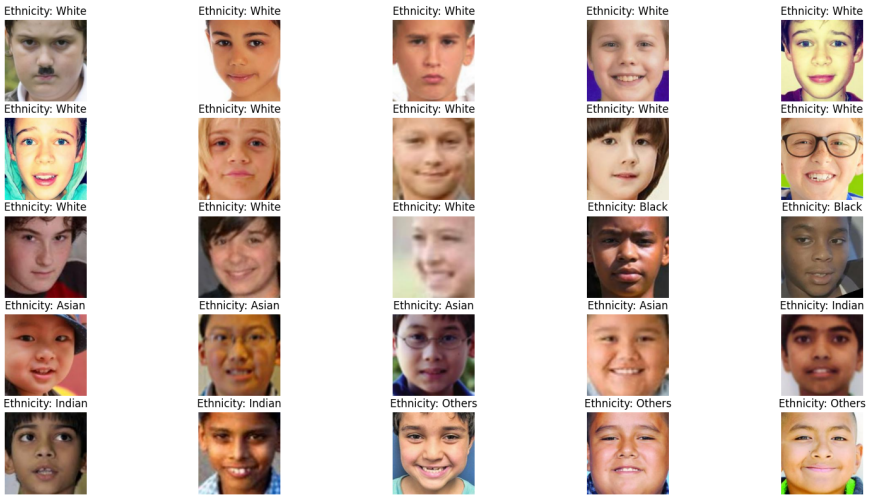

Figure 4.2: Some images from the UTKFace dataset with their label

Chapter 5

Results

5.1 Performance comparison with DeepFace

To compare the performance of the trained model from Section 4.2, we examine the capabilities of the DeepFace framework. We import the DeepFace library and use it to classify the ethnicities in the UTKFace test set. The DeepFace library makes use of the following ethnicity labels: White, Black, Southeast Asian, East Asian, Indian, Latino Hispanic, and Middle Eastern. To match the labels, we combine South East Asian and East Asian to be "Asian" and we group the Latino Hispanic and the Middle Eastern as the "Others" group, such that the ethnicity representations are the same for our model and the DeepFace classifier. To visualize the results, we use the confusion matrix representation, which can be seen in Figure 5.1.

In Figure 5.1, we can see that the DeepFace model is less accurate than all other labels except "Others".

5.2 Effect of Additional Training Methods

For 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, we want the training set to be ethnicity-balanced, such that we can get 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. In Table **??** we can see the statistics of the baseline model on the FairFace test set.

For the baseline model trained on the FairFace dataset in Table 5.2, two test sets are used. This is done because of the bias a test set can have when used 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 helps prevent the model from learning from dataset imperfections, where the model might learn certain patterns shown in the whole dataset, which massively impacts the accuracy of the test set. For the FairFace test set the F1-score is the lowest for the "Indian" label, where we exclude the "Others" label. We added 2000 data augmented "Indian" images to the original baseline training data, and continued training.

Figure 5.1: Comparison of the DeepFace model with our classification model. The top confusion matrix displays the performance of the DeepFace model, while the bottom confusion matrix shows the results of our model.

Table 5.1: Baseline statistics for Equal FairFace dataset

	F1-score	Accuracy
White	0.59	52.0%
Black	0.75	74,4%
Asian	0.83	86,2%
Indian	0.55	66,1%
Others	0.52	47,8%

Table 5.2: FairFace test set

	F1-score	Accuracy
White	0.66	51.0%
Black	0.81	76,2%
Asian	0.83	85,9%
Indian	0.68	66,2%
Others	0.25	65,9%

Table 5.3: UTKFace test set

Table 5.4: Additional training on "Indian" label

	F1-score	Accuracy
White	0.62	84.2%
Black	0.72	58,3%
Asian	0.80	90,2%
Indian	0.52	64,3%
Others	0.11	6,1%

Table 5.5: FairFace test set

	F1-score	Accuracy
White	0.86	84.2%
Black	0.70	54,4%
Asian	0.78	88,4%
Indian	0.66	88,2%
Others	0.05	2,7%

Table 5.6: UTKFace test set

Results can be seen in Table 5.5 and Figure 5.4.

When adding data, we can see 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. For this reason, in following experiments we discard the "Others" category to prevent this category from negatively impacting the experiments.

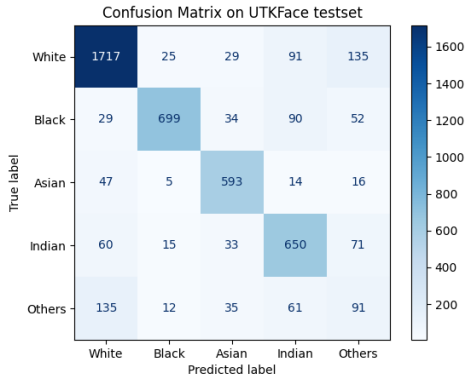

Figure 5.2: Baseline

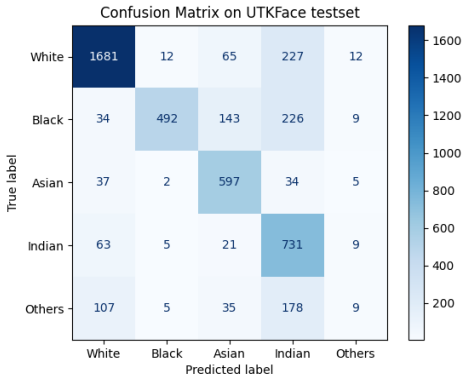

Figure 5.3: Extra

Figure 5.4: Model behaviour before and after adding 2000 Indian augmented training images. The top image are the initial results and the bottom figure is after adding indian augmented images.

5.2.1 F1 score

In this experiment, we use a limited amount of data (1000 images) as a starting point, in which additional training is done using real data to try to improve the class with the lowest F1 score. The baseline results shown in Figure 5.2, show that the Indian class has the lowest F1 score. The model is trained using 5000 images from the UTKFace dataset 3.1, taken randomly. Then after training, we compare the calculated F1 scores of each class, where the class with the lowest F1 score is chosen to get additional training. We then take 1000 real images from the UTKFace dataset with only the class with the lowest F1 score. Because one class may be harder to classify than other classes, we reduce the number of images that are used for retraining, such that we do not have an overload of images of a single class. For this experiment, we lower the number of images that are used for extra training using:

$$\#Images = 1000 - (i \times 100)$$

here i is the iterator variable for how many times we have retrained the model, such that in each training phase we use 100 images less than the previous phase. In Figure 5.5 we can see how to retraining process affects the F1 score of the different classes.

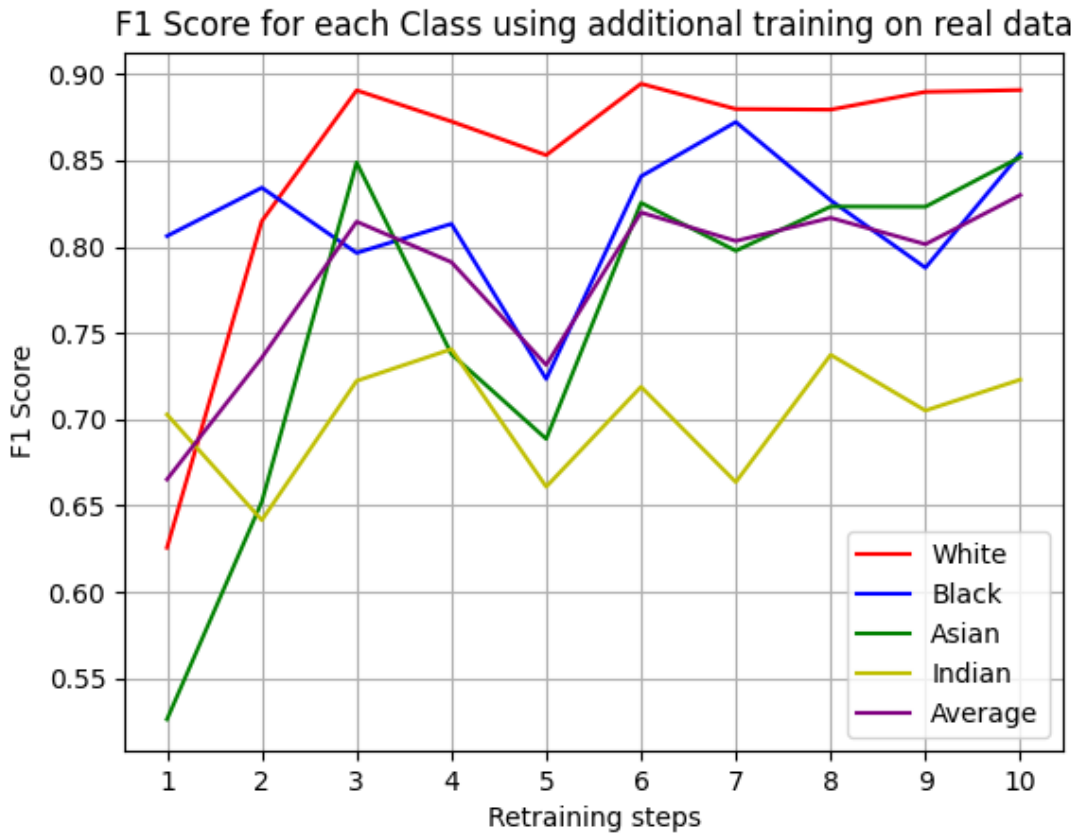

Figure 5.5: Retraining using real data from UTK based on the lowest F1 score

From Figure 5.5 we can see that at the first step, the Asian class had the lowest F1 score before the retraining phase. Using additional images from the Asian class, the class takes quite a big leap from a low F1 score of 0.526 to an F1 score of 0.849 in two training steps. Also,

the White class seems to improve quite noticeably, without receiving additional images. A hypothesis for this can be that when receiving more Asian images, the White class also becomes more distinguishable from the Asian class, which probably has the most similar skin color of the four different classes. It can also be caused due to the additional training epoch that comes when retraining the model, also adjusting the weights of the model. The average improvement can be seen in the purple line labeled "Average", where we see an overall improvement using the real data to perform additional training. We want to maximize the average F1 score of the model, to reduce the bias as much as possible.

5.2.2 Equalized Odds on Real data

A different method to evaluate the model is using the Equalized Odds fairness metric from Section 3.7.3, where we want the model to have an as low as possible Equalized Odds value. This means that the True Positive Rates and the False Positive Rates would be fairly similar between each of the classes, making the model more fair when dealing with the different classes. After calculating the Equalized Odds value between each pair of classes, the average value is taken for each class. The class with the highest Equalized Odds value is the class that might need more training, which is why this class is chosen to receive additional training data. The results of the Equalized Odds experiment using real training data can be seen in Figure 5.6.

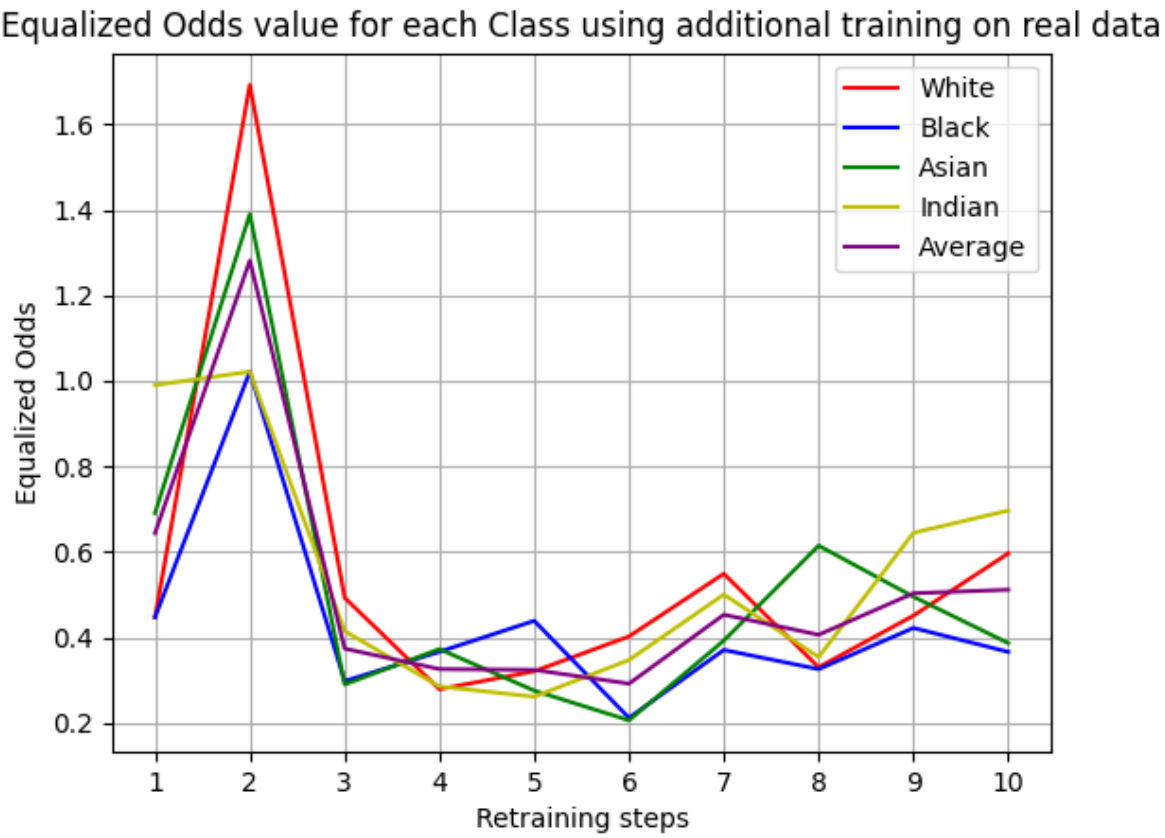

Figure 5.6: Retraining the mode using data derived from the Fairface dataset. The additional training data is selected based on the class with the highest Equalized Odds value, aiming to enhance and fairness across all ethnicity classes.

From Figure 5.6 we can see that after performing additional training to the Indian class in the first retraining step, all other classes get a worse Equalized Odds value. Then after performing additional training on the White class, all classes seem to drop quite heavily. The average trend of the classes seems to be going down, but most of the time the classes have a similar Equalized Odds score.

5.2.3 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 stays the same. Figure 5.7 shows this experiment's results. In contrast to Figure 5.6, the Equalized Odds values seem to differ a lot more initially. The only two classes that received additional training are the Indian and Black classes, which are more than often the two classes with worse results in comparison to White and Asian. At the 9th retraining step, the four classes seem to be very close to each other regarding their Equalized Odds value, with the worst value for Black being 0.41 and the best value for Indian being 0.36. The experiment of helping the outlier classes to improve their Equalized Odds value seems to work quite well, mainly by adjusting the outlier values of the Indian and Black classes.

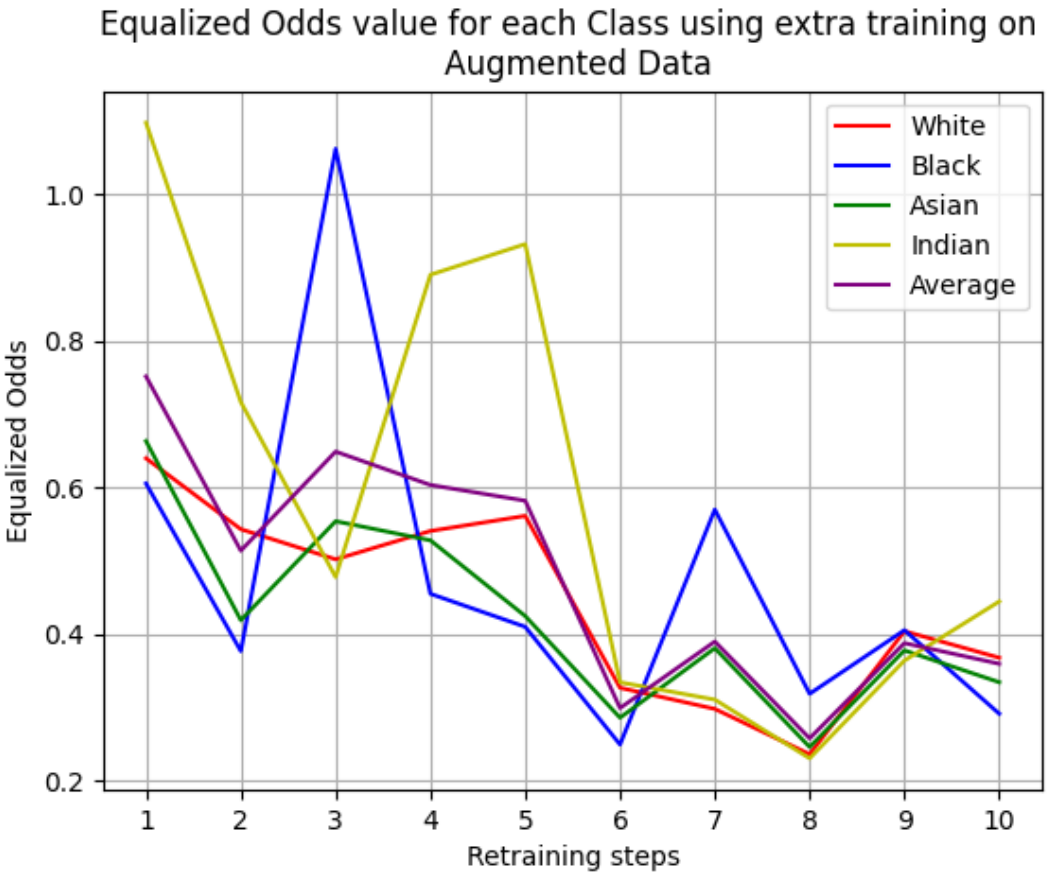

Figure 5.7: Retraining using augmented data derived from UTKFace based on the highest Equalized Odds value. After the first training step we see that the Indian class has the highest value, which means that this class receives extra images based on data augmentation for the next training step.

5.2.4 Equalized Odds on GAN Data

In this experiment we use Generative Algorithms 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. For this experiment we calculate the Equalized Odds value for each ethnicity class, to showcase the effect of adding additional training data to the baseline UTKFace training set. The results are shown in Figure 5.8. Additional training is based on the highest Equalized Odds value, to try to balance the Equalized Odds values over all classes. After some epochs, we can see that the Equalized Odds values of all classes are fairly close to each other.

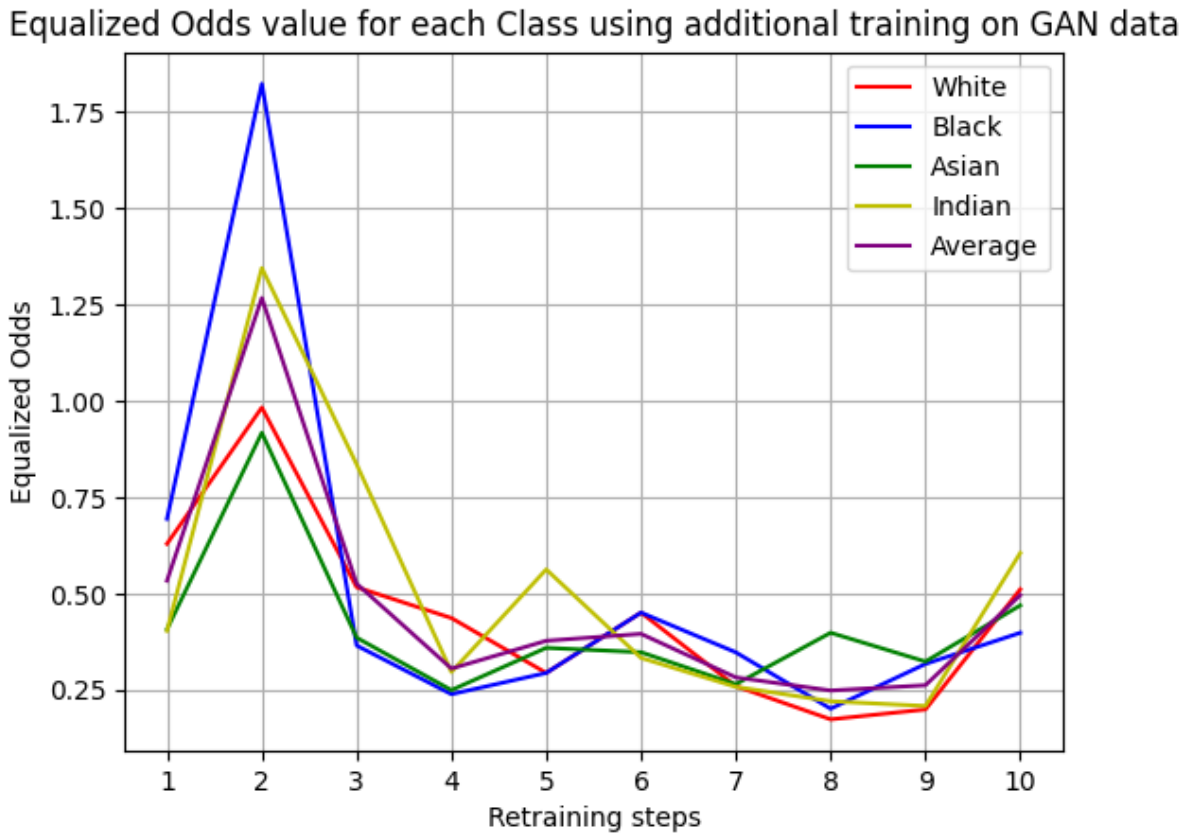

Figure 5.8: Retraining the mode using data generated by the GAN dervised from the UTKFace dataset. The additional training data is selected based on the class with the highest Equalized Odds value, aiming to enhance and fairness across all ethnictiy classes.

5.3 Balancing Fairness using different Input Data

In this section of the experiments, we try different input datasets, to experiment with different ethnic splits and the effect the splits have on the training of the model. 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.

5.3.1 Baseline UTKFace

A method to balance fairness in machine learning models is to adjust the input data the model uses for training. In this experiment, we look at 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. In Figure 5.9 we can see that the UTKFace dataset without the "Others" category performs quite decent, with the "Indian" class with the highest Equalized Odds value. Let's look at the ethnicity split of the UTKFace dataset in Table 3.1. We see that the "Indian" category has only 2146 images, while the other classes have more images. To tackle this problem, in the next experiment, we add Indian image data from the FairFace dataset to the UTKFace dataset. Using more Indian image data, we want the model to get a better understanding of classifying Indian images and to be able to tell the difference between the Indian class and the other three classes.

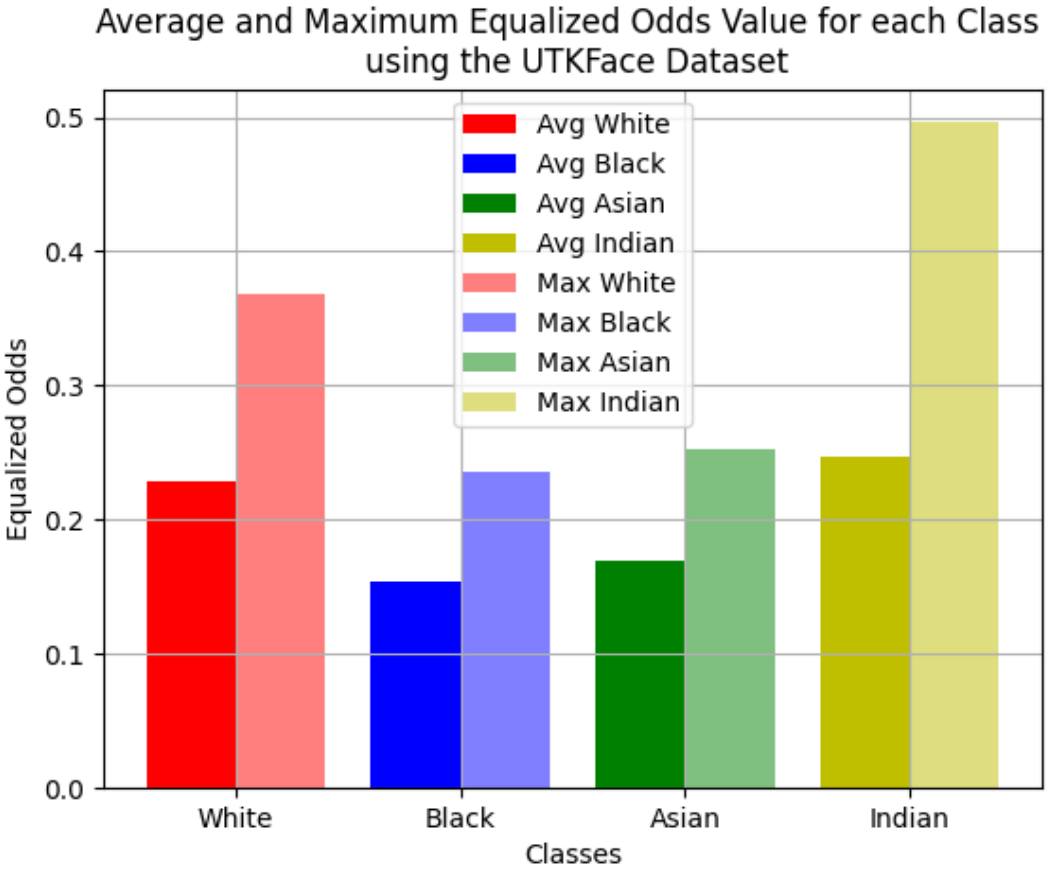

Figure 5.9: Equalized Odds values of UTKFace dataset without the "Others" category

5.3.2 Adding Indian images to UTKFace

To improve the Equalized Odds value that we get in Figure 5.9, we experiment with adding Indian images to the UTKFace dataset such that the model might improve its learning process of deciding if an image contains a person with the Indian ethnicity. We add 4924 Indian images from the FairFace dataset to match the amount of images with the White classification (8063). The results of this experiment can be seen in Figure 5.10. This experiment shows that adding

Indian images to the dataset did not improve the model compared to the baseline UTKFace dataset. There are many possibilities why this might happen, such as the quality of the images, the difference between images from UTKFace or FairFace, or because the model might favor guessing that an image has the Indian class when in doubt since there are now more Indian images. To tackle this problem, we experiment with adding images to all classes, such that every class contains exactly 8063 images.

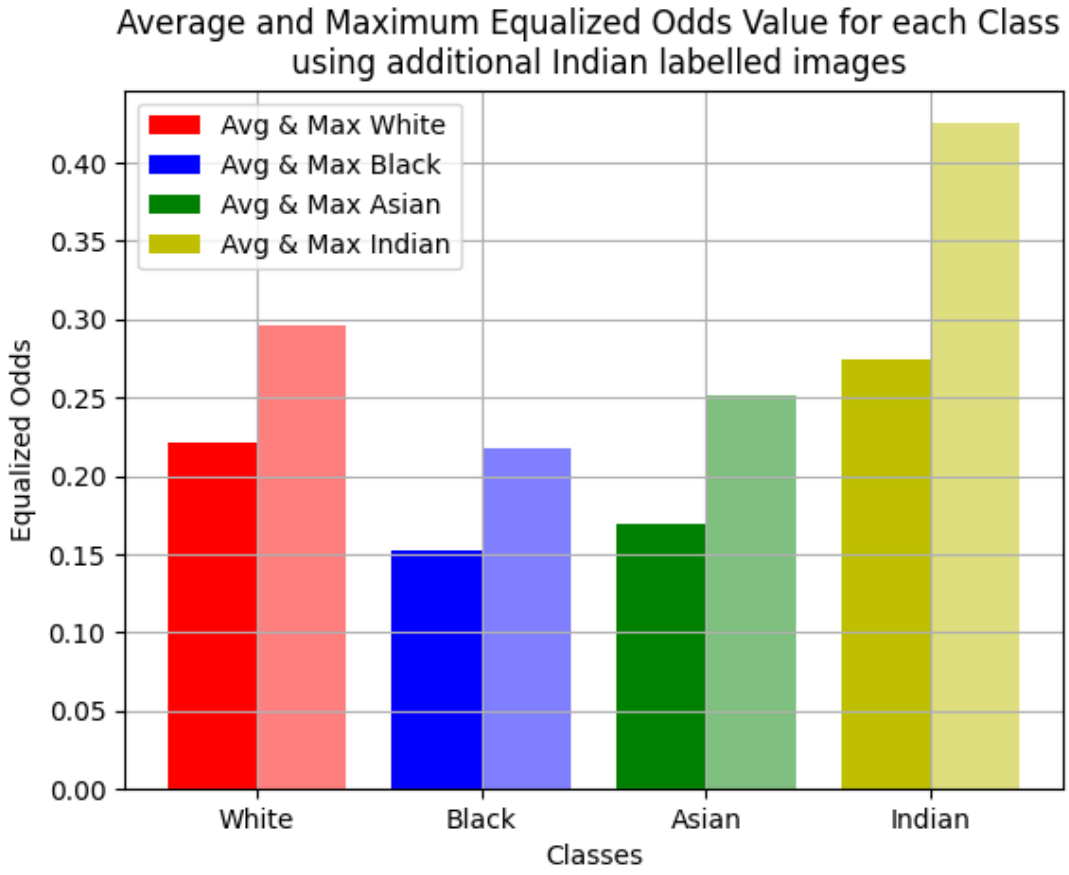

Figure 5.10: Equalized Odds values of UTKFace dataset when adding 4924 Indian images

5.3.3 Equalizing image count in UTKFace and FairFace

For the next experiment, we added images to the Black, Asian, and Indian classes to match the number of images in the White class, which has the most images with some 8063 different images. Using an equalized image count for all classes, we mitigate the chance of bias which can come from an unbalanced ethnic split in the input dataset. When in doubt, we also address the issue of the model favoring classes with more images. To add images to the UTKFace dataset and also be able to equally compare the dataset to the FairFace dataset, we take 4000 images of each class, totaling 16.000 images. For this, we need to add images to the Black, Asian, and Indian ethnicity 3.1. We also remove the excess images of the White category, by randomly selecting the 4080 images to remove. After adding randomly chosen images from the FairFace dataset, we add each class to exactly 4000 images. The test set used is a part of the original test set of the UTKFace dataset which is also balanced by taking 675 images of each

class such that there is an equal amount of test images for each possible classification. The results of this experiment are shown in Figure 5.14.

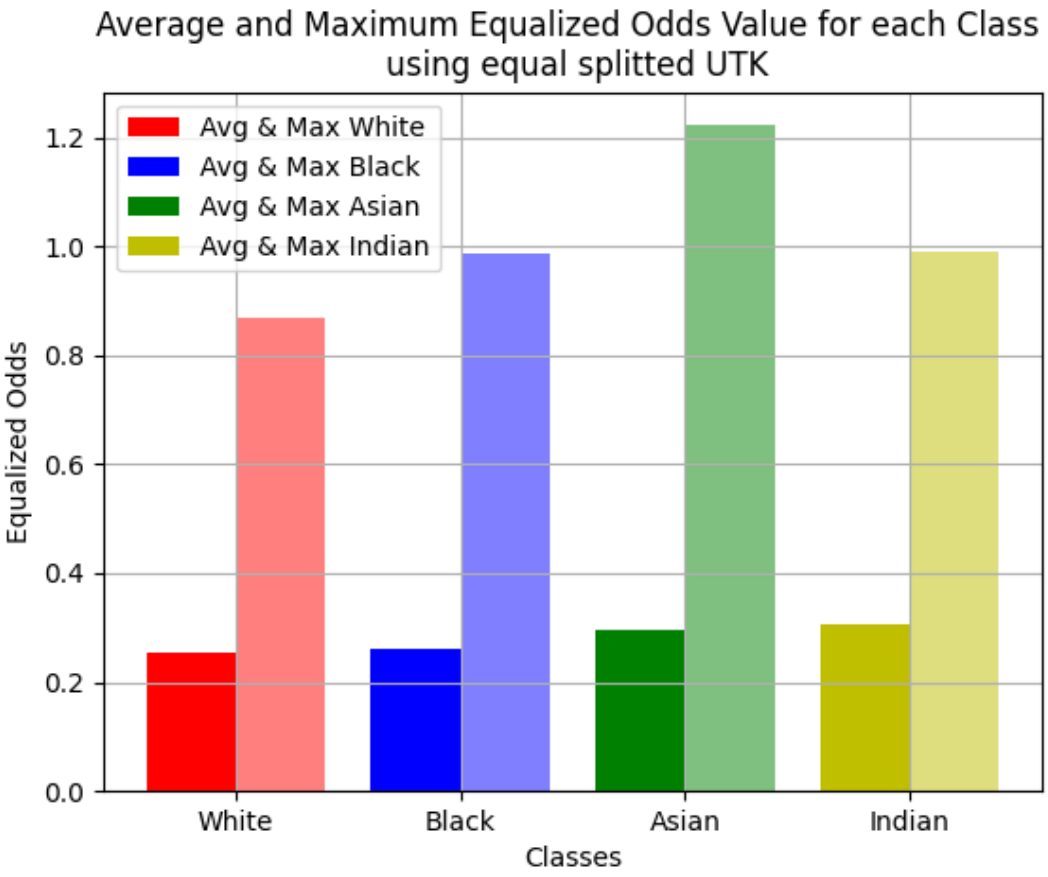

Figure 5.11: Equalized Odds values of balanced UTKFace dataset

From the results in Figure 5.14 we can see that the Equalized Odds values for all classes are quite similar to each other. We can see that the model still under performs for the "Indian" class, in comparison with the other classes. When we compare this to the other results, we see a clear pattern that the model keeps its bias towards images with the "Indian" label, but in this experiment, the differences are not substantial. Another interesting observation is that the Equalized Odds values for the "White" class are worse than expected, and also worse than in other experiments where the image count of the "White" class is in favour. Removing images from the "White" class looks to affect the model in labelling images with a low confidence. It might be the case that because of the high amount of images in the "White" class in the original UTKFace dataset, the uncertain images are more often than not classified with the "White" label, causing the accuracy to also increase. When using an equal amount of images, we remove this characteristic of the AI model to guess more often than not labels with a higher amount of images, which causes the model to make decisions more based on the actual attributes of the images.

5.4 Improving Fairness using Additional Training by adding (un)generated images

In this experiment we look at the effect of adding images from different datasets, augmented images or generated images using a GAN. We do this 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. This experiment is performed using a single epoch where we train the adjusted UTKFace model. After the first epoch we test the performance of the model and calculate the fairness metrics. Based on the lowest Equalized Odds value, we add 1000 images from three possible separate datasets. The first dataset contains original FairFace images, the second dataset contains data augmented images from the UTKFace dataset using the data augmentation methods from Section 3.5. The third and last dataset consists of generated images using StyleGAN, where the input consists of original images from the UTKFace dataset, where new images are generated. All images from these three datasets are split over each class, such that we can add images from a single class to our original dataset. The experiment consists of three separate runs, with each its own dataset for additional training. Each of the three experiments is averaged over 10 runs, such that randomness is lowered. The results are shown in Figure 5.12.

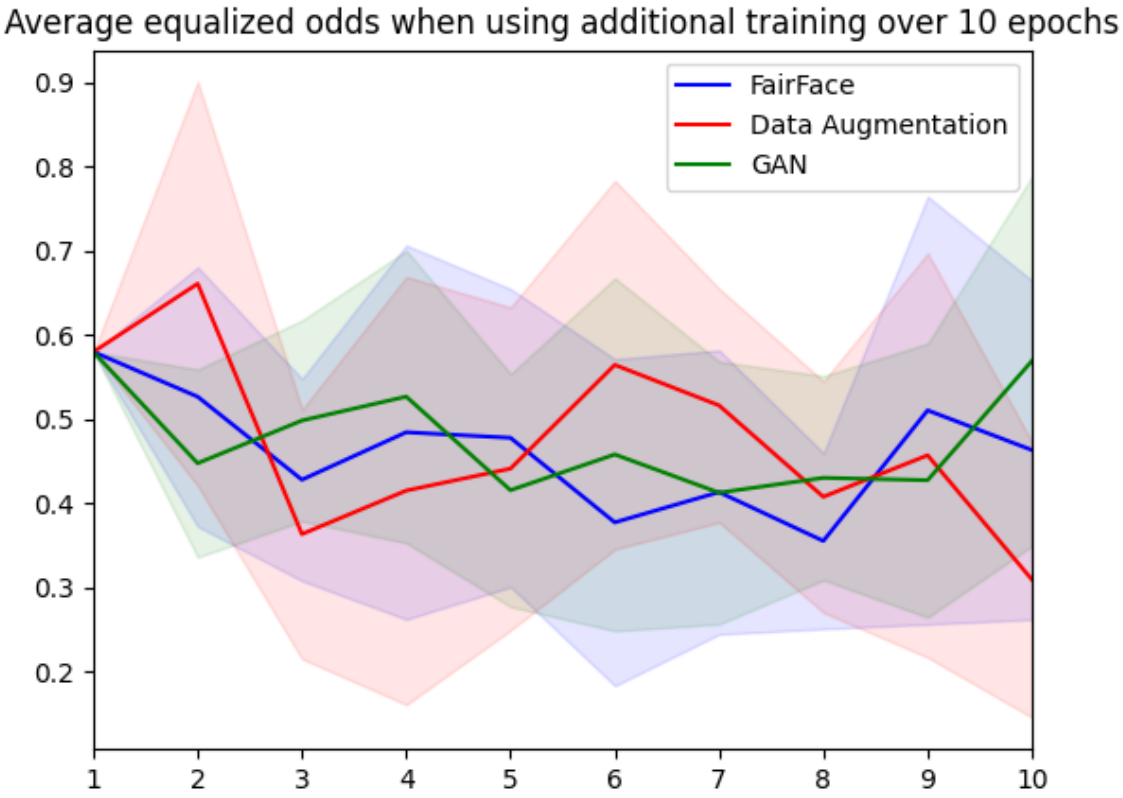

Figure 5.12: Equalized Odds values when performing different additional training methods to the UTKFace dataset. Each experiment is averaged over 10 runs. The area around the lines is the confidence interval.

From the results we can see that the three different lines seem to decrease in the first few epochs, but after the fourth epoch, the three lines all increase their Equalized Odds value. From this point, the model does not seem to optimize any further. The goal of this experiment is to balance the amount of images across the different classes, where classes with lower equalized odds values increase their image count, such that the model has more images to learn from. In this experiment almost all images were added to the Indian class, the model never seemed to get the equalized odds values for the Indian class to increase above any of the other classes. After the fourth epoch, the Indian class has the most images of all classes, which causes the model to guess more based on the fact that there are more Indian images than other classes, instead of classifying based on characteristics. This causes the model to have a high amount of false positives on the Indian class, which negatively impacts the Averaged Equalized Odds values for all classes. Another interesting result is that it does not look like the differences between the FairFace data, Augmented data, and GAN data affect the results much, with all types of image data performing quite similar. Initially, we thought that the GAN image data would perform poorly, but in the experiment, it is only slightly worse than the other two methods. To get a better understanding of the effect of the GAN data on the training process of the model, in the next experiment, we will add the different additional data before training, such that we get equally balanced images over all classes.

5.5 Balancing training data by adding images using various techniques

In Section 5.4 we have seen an experiment where we add data to the model based on its performance during training. In this experiment, we use different adaptations of the UTKFace dataset to balance the training data to 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. The results are shown in Figure 5.13.

From the results in Figure 5.13 we can see that the UTK+FairFace dataset achieves the lowest equalized odds value when averaged over 10 runs. We can also see that the training data with the worst results is the dataset where generated images are used, created using StyleGAN. This is likely due to the fact that the StyleGAN model is trained on unbalanced ethnicity training data, resulting in its lacking performance for the "Indian" class for example. We can also see that using data augmentation is a viable option to add data, which can be used to balance the number of images each class contains, which can also remove potential bias. The dataset with data-augmented images looks to perform similarly to the UTKFace dataset. There are still a lot of popular image datasets that have a greater imbalance in images per class, where using data augmentation techniques can really improve the stability of the training set and to remove potential bias. However, experimentation using different fairness metrics is needed to get an understanding of the effect of adding images to the original dataset. In this case, we have two datasets with the same characteristics and labels (UTKFace and FairFace), where merging the datasets might be a good idea to improve fairness. However, in a lot of cases, there is no other dataset available that can be used to improve fairness. In those

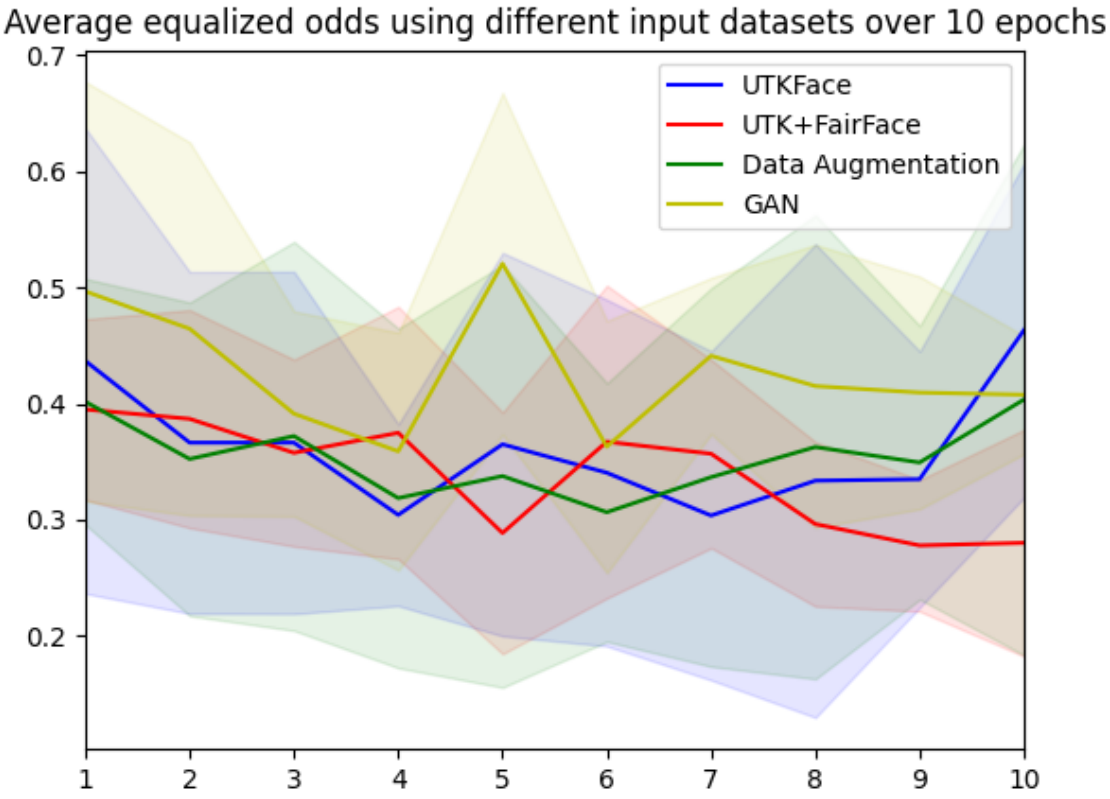

Figure 5.13: Equalized Odds values calculated over 10 different runs using different input training data sets. The UTKFace dataset has its original amount of images for each ethnicity, whereas the other datasets have 4000 images for each class, containing original images from the UTKFace dataset, supplemented with additional training data.

cases it might benefit to use other techniques such as generating new images using various GAN implementations, or by using data augmentation to get more training data and balance the classes of the training data.

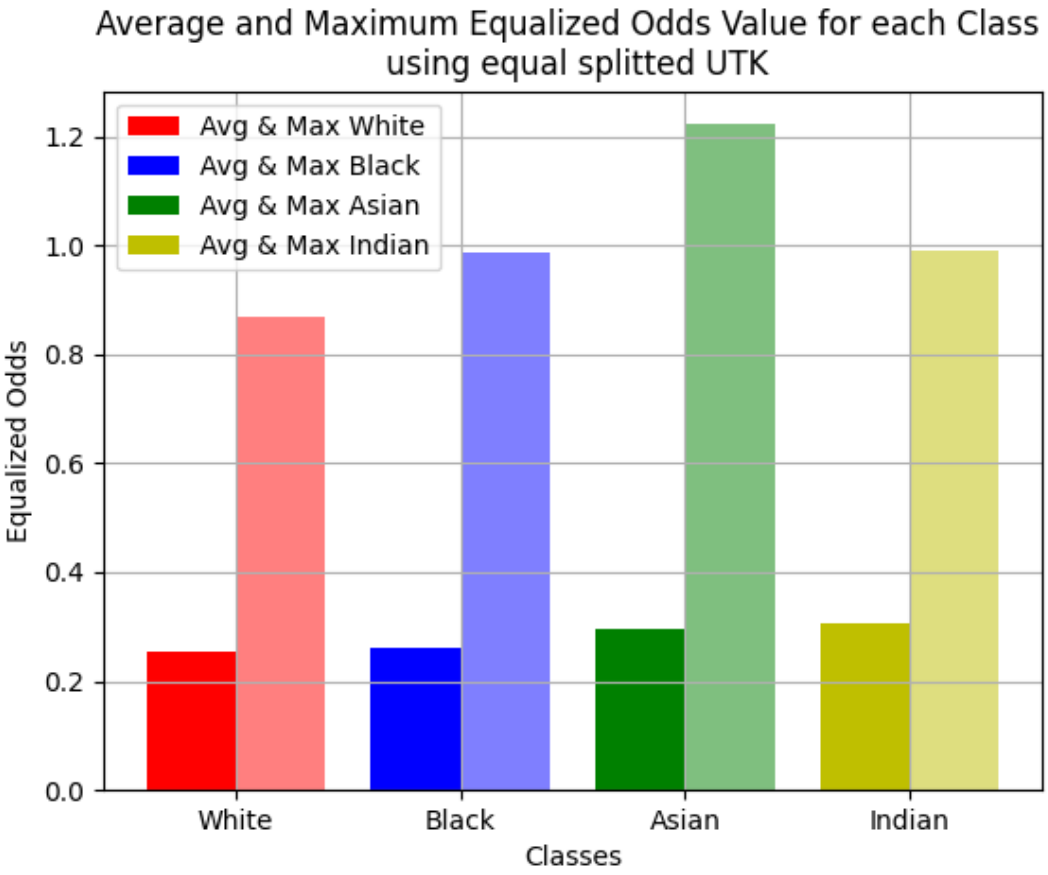

Figure 5.14: Equalized Odds values of balanced UTKFace dataset

Method (UTK)	**Accuracy**	**Equalized Odds**	**Macro F1**	**Disparate Impact**
Baseline	0.86	0.30	0.86	0.46
Undersample	0.87	0.23	0.87	0.50
Generative images	0.89	0.18	0.89	0.48
Oversample	0.87	0.27	0.87	0.47
Data Augmentation	0.87	0.27	0.87	0.52

Table 5.7: UTKFace test set

Method (UTK)	**Accuracy**	**Equalized Odds**	**Macro F1**	**Disparate Impact**
Baseline	0.77	0.40	0.76	0.36
Undersample	0.75	0.46	0.74	0.44
Generative images	0.78	0.33	0.77	0.40
Oversample	0.78	0.29	0.77	0.50
Data Augmentation	0.77	0.38	0.76	0.40

Table 5.8: FairFace test set

Chapter 6

Discussion

6.1 StyleGAN training data

From the experiments in Section 3.6, we observe that the generated images from the GAN are less effective when used for additional training of the model. If we look at the generated images manually, it becomes apparent that the generated images from the "Indian" label have a lower quality than the images from the "White" label, for example. This research shows that training data should have a high fairness level to make the models able to classify across different ethnicities, which might be lacking from the perspective of the training data used by the StyleGAN model. We can see the training data used by the StyleGAN model in Table 3.4. The dataset with the highest amount of images is the VGGFace2 dataset, with around 33.1 million images. From Figure 3.1 we can see that the VGG2 dataset is massively biased, with an expected amount of around 75%-80% "White" images. This skewed ethnicity split also translates to the performance of the StyleGAN model, which performs the best in creating images from the "White" label. This is a big problem for creating balanced and fair generated images. To be able to create fair and balanced generated data, the GAN models should be trained on balanced datasets, which translates to the output of the model. In future work, the state-of-the-art image datasets must become more racially balanced, such that all models trained on the dataset can improve on its fairness and bias towards different ethnicity groups. Another dataset with a high amount of images used by the StyleGAN model is the FFHQ dataset. This dataset does not have ethnicity labels. From the research of Maluleke et al., [24], a random subset is selected from the FFHQ dataset and manually labeled, to get an understanding of the ethnicity split in the dataset. The research of Maluleke finds that FFHQ is composed of around 69% white, 4% black, and 27% non-black and non-white facial images. It becomes transparent that also the FFHQ dataset contains a massive bias in ethnicity, which directly translates to the results made by the StyleGAN model.

6.2 Additional training methods

The experiments demonstrate that additional training can be accomplished through various methods of acquiring new image data. We have experimented with data from different datasets, augmented data, and generated data using the StyleGAN implementation. If we look at the effect of these additional data methods, we can see that these methods help in getting similar Equalized Odds values for all classes after some retraining steps. In Figure 5.7 we can see that

in the 6th and 9th epochs, the Equalized Odds values for all classes get very close to each other, where the results show strong fairness when comparing the classes to each other. From Figure **??** we can also see that after three epochs the Equalized Odds values remain very similar to each other, where no class is significantly worse than the other. Another interesting aspect of the experiments with additional training is that the generated data using StyleGAN performs very well in comparison with using images from another dataset or using data augmentation. When there are limited images available, generating new images might be a good way to experiment with improving the fairness of the model.

From Figure 5.6 we can see that after performing additional training to the Indian class in the first retraining step, all other classes get a worse Equalized Odds value. Then after performing additional training on the White class, all classes seem to drop quite heavily. The average trend of the classes seems to be going down, but most of the time the classes have a similar Equalized Odds score.

When training on Augmented Data, we can see that Figure 5.7 shows promising results for using augmented data instead of real data. In contrast to Figure 5.6, the Equalized Odds values seem to differ a lot more initially. The only two classes that received additional training are the Indian and Black classes, which are more than often the two classes with worse results in comparison to White and Asian. At the 9th retraining step, the four classes seem to be very close to each other regarding their Equalized Odds value, with the worst value for Black being 0.41 and the best value for Indian being 0.36. The experiment of helping the outlier classes to improve their Equalized Odds value seems to work quite well, mainly by adjusting the outlier values of the Indian and Black classes.

6.3 Equalizing image count in datasets to improve fairness

From the experiment using equalized image count in the UTKFace and FairFace dataset in Figure 5.14 we can see that the Equalized Odds values for all classes are quite similar to each other. We can see that the model still underperforms for the "Indian" class, in comparison with the other classes. When we compare this to the other results, we see a clear pattern that the model keeps its bias towards images with the "Indian" label, but in this experiment, the differences are not substantial. Another interesting observation is that the Equalized Odds values for the "White" class are worse than expected, and also worse than in other experiments where the image count of the "White" class is in favor. Removing images from the "White" class looks to affect the model in labeling images with low confidence. It might be the case that because of the high amount of images in the "White" class in the original UTKFace dataset, the uncertain images are more often than not classified with the "White" label, causing the accuracy to also increase. When using an equal amount of images, we remove this characteristic of the AI model to guess more often than not labels with a higher amount of images, which causes the model to make decisions more based on the actual attributes of the images.

When improving fairness using additional images created on various techniques from Section 5.4, we can see that the three different lines seem to decrease in the first few epochs, but after the fourth epoch, the three lines all increase their Equalized Odds value. From this point, the model does not seem to optimize any further. The goal of this experiment is to balance out the number of images across the different classes, where classes that have lower equalized odds values to increase its image count, such that the model has more images to learn from. In this

experiment almost all images were added to the Indian class, the model never seemed to get the equalized odds values for the Indian class to increase above any of the other classes. After the fourth epoch, the Indian class has the most images of all classes, which causes the model to guess more based on the fact that there are more Indian images than other classes, instead of classifying based on characteristics. This causes the model to have a high amount of false positives on the Indian class, which negatively impacts the Averaged Equalized Odds values for all classes. Another interesting result is that it does not look like the differences between the FairFace, Augmented, and GAN data affect the results much, with all types of image data performing quite similarly. Initially, we thought that the GAN image data would perform poorly, but in the experiment, it is only slightly worse than the other two methods. To get a better understanding of the effect of the GAN data on the training process of the model, in the next experiment, we will add the different additional data before training, such that we get equally balanced images over all classes.

From the results in Figure 5.13 we can see that the UTK+FairFace dataset achieves the lowest equalized odds value when averaged over 10 runs. We can also see that the training data with the worst results is the dataset where generated images are used, created using StyleGAN. This is likely due to the fact that the StyleGAN model is trained on unbalanced ethnicity training data, resulting in its lacking performance for the "Indian" class for example. We can also see that using data augmentation is a viable option to add data, which can be used to balance the number of images each class contains, which can also remove potential bias. The dataset with data-augmented images looks to perform similarly to the UTKFace dataset. There are still a lot of popular image datasets that have a greater imbalance in images per class, where using data augmentation techniques can really improve the stability of the training set and to remove potential bias. However, experimentation using different fairness metrics is needed to get an understanding about the effect of adding images to the original dataset. In this case, we have two datasets with the same characteristics and labels (UTKFace and FairFace), where merging the datasets might be a good idea to improve fairness. However, in a lot of cases, there is no other dataset available that can be used to improve fairness. In those cases it might benefit to use other techniques such as generating new images using various GAN implementations, or by using data augmentation to get more training data and balance the classes of the training data.

Chapter 7

Conclusion

This paper explores various techniques to enhance the fairness of facial image datasets and models. Our literature review reveals that many models continue to utilize heavily biased image datasets, which adversely affects their practical usability. When models lack fairness, not only can they discriminate between ethnic groups, but also make the models less useful in countries across the globe. To improve the fairness of these models, we have looked at adding data during training time based on the fairness metrics on the test set, where the class with the worst fairness receives more training attention. We can see that both the additional training and the balancing of the training data have promising results in terms of improving fairness. In Figure 5.13 we can see that adding generated images using the StyleGAN implementation has the worst results in terms of fairness, which can be explained by the bias the StyleGAN model has based on its training data, see 3.4. To answer the research questions, we find that several existing face datasets still contain bias due to unequal ethnicity splits, leading to unequal learning of the model. We see this happening in the implementation of StyleGAN, where images from certain ethnicity groups have far better generation quality than others. We can improve fairness by performing modifications to the training data, leading to a more balanced ethnicity split. We can also evaluate the model based on fairness metrics to showcase possible bias, which can help in the development process of making fair facial detection models. In this research we have found that both using data augmentation and borrowing images from similar datasets are the most useful method of mitigating unwanted bias and improving fairness, where data augmentation can benefit datasets without similar datasets available, and data augmentation can be used for small datasets to increase its image count and improve the balance between different ethnicity groups. Using generated images may be helpful in future work when the models used to create generated images become more balanced themselves.

In future work, it would be interesting to see if new state-of-the-art GAN implementations will improve on its fairness and balance, where the models are more likely to be used for improving fairness. Furthermore, when improving on fairness, models are more likely to be used in real applications, where models should not base its decisions based on ethnicity or other aspects. There is still a lot of work to do in terms of fairness, which will likely get more attention when AI will more and more be used in real applications.

Bibliography

[1] Alejandro Acien et al. "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition". In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales. Cham: Springer International Publishing, 2019, pp. 584–593. ISBN: 978-3-030-13469-3.

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings". In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, 2019, pp. 556–572. ISBN: 978-3-030-11009-3.

[3] John Angileri et al. "Ethical considerations of facial classification: Reducing racial bias in AI". In: *Retrieved February* 21 (2019), p. 2020.

[4] Julia Angwin et al. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks". In: *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[5] Cagdas Bak et al. *Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction*. 2017. arXiv: 1607.04730 [cs.CV].

[6] Guha Balakrishnan et al. "Towards Causal Benchmarking of Bias in Face Analysis Algorithms". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 547–563. ISBN: 978-3-030-58523-5.

[7] Sebastian Benthall and Bruce D. Haynes. "Racial categories in machine learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 289–298. ISBN: 9781450361255. DOI: 10.1145/3287560.3287575. URL: https://doi.org/10.1145/3287560.3287575.

[8] Miranda Bogen and Aaron Rieke. "Help wanted: an examination of hiring algorithms, equity, and bias". In: 2018. URL: https://api.semanticscholar.org/CorpusID:158203520.

[9] F. Zuiderveen Borgesius. "Discrimination, artificial intelligence, and algorithmic decision-making". In: *Strasbourg: Council of Europe, Directorate General of Democracy* (2018), p. 49. DOI: https://hdl.handle.net/11245.1/7bdabff5-c1d9-484f-81f2-e469e03e2360.

[10] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/p81-buolamwini18a.html.

[11] Roxana Daneshjou et al. "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review". In: *JAMA Dermatology* 157.11 (Nov. 2021), pp. 1362–1369. ISSN: 2168-6068. DOI: 10.1001/jamadermatol.2021.3129. URL: https://doi.org/10.1001/jamadermatol.2021.3129.

[12] David Danks and Alex London. "Algorithmic Bias in Autonomous Systems". In: Aug. 2017, pp. 4691–4697. DOI: 10.24963/ijcai.2017/654.

[13] Judy Wawira Gichoya et al. "AI pitfalls and what not to do: mitigating bias in AI". In: *British Journal of Radiology* 96.1150 (Sept. 2023), p. 20230023. ISSN: 0007-1285. DOI: 10.1259/bjr.20230023. eprint: https://academic.oup.com/bjr/article-pdf/96/1150/20230023/54904835/bjr.20230023.pdf. URL: https://doi.org/10.1259/bjr.20230023.

[14] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[15] Lena Hafner, Theodor Peter Peifer, and Franziska Sofia Hafner. *Equal accuracy for Andrew and Abubakar-detecting and mitigating bias in name-ethnicity classification algorithms - ai amp; society*. Feb. 2023. URL: https://link.springer.com/article/10.1007/s00146-022-01619-4.

[16] Ayanna Howard and Jason Borenstein. "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity". In: *Science and Engineering Ethics* 24 (Oct. 2018). DOI: 10.1007/s11948-017-9975-2.

[17] Brian Hu et al. "Xaitk-Saliency: An Open Source Explainable AI Toolkit for Saliency". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.13 (June 2023), pp. 15760–15766. DOI: 10.1609/aaai.v37i13.26871. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26871.

[18] Ahsan Ul Islam. "Gender and Ethnicity Bias in Deep Learning". PhD thesis. 2023.

[19] Kimmo Kärkkäinen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age". In: *ArXiv* abs/1908.04913 (2019). URL: https://api.semanticscholar.org/CorpusID:260536756.

[20] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: 1812.04948 [cs.NE].

[21] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2018. arXiv: 1710.10196 [cs.NE].

[22] Os Keyes. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: 10.1145/3274357. URL: https://doi.org/10.1145/3274357.

[23] Z. Khan and Y. Fu. "One label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision". In: *Virtual Event, Canada ACM, New York* (2021).

[24] Vongani H. Maluleke et al. *Studying Bias in GANs through the Lens of Race*. 2022. arXiv: 2209.02836 [cs.CV].

[25] Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. 2022. arXiv: 1908.09635 [cs.LG].

[26] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/, 2019.

[27] Yotam Nitzan et al. "Face identity disentanglement via latent space mapping". In: *ACM Transactions on Graphics (TOG)* 39 (2020), pp. 1–14.

[28] Inioluwa Deborah Raji et al. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. 2020. arXiv: 2001.00973 [cs.CY].

[29] Christian Schorr et al. "Neuroscope: An Explainable AI Toolbox for Semantic Segmentation and Image Classification of Convolutional Neural Nets". In: *Applied Sciences* 11.5 (2021). ISSN: 2076-3417. DOI: 10.3390/app11052199. URL: https://www.mdpi.com/2076-3417/11/5/2199.

[30] Sefik Ilkin Serengil and Alper Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE. 2021, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: https://doi.org/10.1109/ICEET53442.2021.9659697.

[31] Yujun Shen et al. "InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs". In: *TPAMI* (2020).

[32] Yujun Shen et al. "Interpreting the Latent Space of GANs for Semantic Face Editing". In: *CVPR*. 2020.

[33] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (July 2019), p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0. URL: https://doi.org/10.1186/s40537-019-0197-0.

[34] Ramya Srinivasan and Ajay Chander. "Biases in AI Systems". In: *Commun. ACM* 64.8 (July 2021), pp. 44–49. ISSN: 0001-0782. DOI: 10.1145/3464903. URL: https://doi.org/10.1145/3464903.

[35] Anthony W Flores, Kristin Bechtel, and Christopher Lowenkamp. "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."". In: *Federal probation* 80 (Sept. 2016).

[36] Tianlu Wang et al. *Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations*. 2019. arXiv: 1811.08489 [cs.CV].

[37] Zhifei Zhang, Yang Song, and Hairong Qi. "Age Progression Regression by Conditional Adversarial Autoencoder". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017.