university of groningen / faculty of science and engineering / artificial intelligence

# Demographic Biases in Real-Fake Image Classification Tasks

Bachelor's Project Thesis

Efe Görkem Şirin, s4808746, e.g.sirin@student.rug.nl,
Supervisors: M. Zullich

**Abstract:** The rapid advancements in artificial intelligence and machine learning have led to significant progress in image generation and classification, with techniques like Generative Adversarial Networks (GANs) and diffusion models like Stable Diffusion producing highly realistic synthetic images. However, these developments also present challenges in distinguishing real from fake images, which is crucial for various applications. One less explored, yet critically important aspect is the presence of demographic biases in image classification tasks. This thesis focuses on examining the demographic biases present in the classification of real and synthetic face images generated by Stable Diffusion and GAN models. By evaluating the performance of Convolutional Neural Networks (CNN) and with transfer learning (using Resnet-18) models across male and female images, the study aims to uncover potential disparities in accuracy and assess the implications of these biases. The evaluation involves two primary classification tasks: distinguishing face images generated by Stable Diffusion from real face images, and differentiating between face images generated by GANs and real face images. The results of this study provide valuable insights into the demographic biases present in the classification of real and synthetic images, which can inform the development of more robust and fair image classification systems. All of the code used in this paper is available at https://efesirin.com/realfakeclassification.html.

## 1 Introduction

In recent years, the rapid advancement of artificial intelligence (AI) and machine learning (ML) has led to significant progress in the field of image generation and classification. Techniques such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and diffusion models like Stable Diffusion (SD) (Rombach et al., 2022) have demonstrated the ability to create highly realistic synthetic images that are often indistinguishable from real photographs. These developments, while groundbreaking, also present substantial challenges, particularly in the context of distinguishing between real and synthetic images—a task crucial for various applications ranging from digital forensics to social media moderation.

One of the less explored, yet critically important aspects of this field is the presence of demographic biases in image classification tasks. Demographic bias occurs when the performance of an AI model varies significantly across different demographic groups, such as gender, age, or ethnicity (Rawal et al., 2024). In the context of real versus

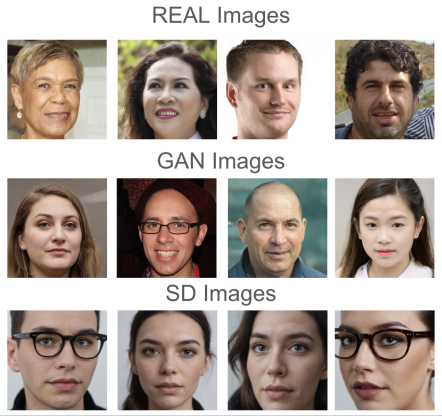



**Figure 1.1: Examples of training images: The 1st row shows real images, the 2nd row displays images generated by Generative Adversarial Networks (GANs), and the 3rd row showcases images generated by Stable Diffusion models. I trained models that classify between Real-GAN and Real-SD and I performed bias testing on the trained models.**

fake image classification, such biases can lead to unequal treatment of images based on the depicted demographic characteristics, potentially resulting in systemic inaccuracies and unfair outcomes.

In this thesis, I focus on examining the demographic biases present in the classification of real and synthetic face images generated by Stable Diffusion and GAN models. By evaluating the performance of Convolutional Neural Networks (CNN) and ResNet-18 models across male and female images, I aim to uncover potential disparities in accuracy and assess the implications of these biases. My evaluation involves two primary classification tasks: distinguishing face images generated by Stable Diffusion from real face images, and differentiating GAN-generated face images from real ones. To statistically analyze the disparities in accuracy between male and female images, I will perform a two-proportion Z-test and also will create saliency maps to see which parts of the images are considered by the model in classification process.

The motivation for this research stems from the increasing integration of AI-generated images in various sectors and the need to ensure fairness and reliability in image classification systems. Understanding and mitigating demographic biases is essential for developing equitable AI technologies that serve diverse populations without perpetuating existing inequalities. This study contributes to the broader effort of identifying and addressing bias in AI systems, promoting the creation of more inclusive and fair technological solutions.

My personal observations suggest that there are more female images than male images in AI-generated datasets, as female images are generally more abundant and easier to obtain in public datasets. This leads to the question of whether classification models exhibit demographic biases as a result. The central hypothesis of this thesis is that such biases will be present, with differences in accuracy between male and female images. Specifically, I hypothesize that classification models will achieve higher accuracy for female images compared to male images. This anticipated disparity is hypothesized to stem from the greater number of female images used in training these models. To support this hypothesis, I will conduct an analysis to determine if the observed trends are indeed factual.

The evolution of generative AI models has led to diverse architectures, each with its unique strengths. Figures Figure 1.2 and Figure 1.3 illustrate two prominent approaches: Generative Adversarial Networks (GANs) and Stable Diffusion models. While GANs rely on a competitive dynamic between generator and discriminator networks, Stable Diffusion employs a more intricate process involving latent space transformations and conditioning. These diagrams highlight the fundamental differences in their architectures, from GANs' adversarial training to Stable Diffusion's use of diffusion processes and denoising techniques. Understanding these distinct approaches provides insight into the rapid advancements in AI-driven image generation and manipulation techniques.

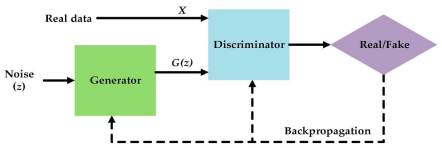



**Figure 1.2: Schematic diagram of a Generative Adversarial Network (GAN). The Generator creates synthetic data from random noise, while the Discriminator attempts to distinguish between real and generated data. Through iterative training and backpropagation, the Generator improves at producing realistic fake samples that can fool the Discriminator (Kyang, 2024).**

## 1.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), are a class of machine learning frameworks designed for generating synthetic data that closely resembles real data. A GAN consists of two neural networks: the generator and the discriminator. The generator creates fake data, such as images, while the discriminator evaluates the authenticity of these images, distinguishing between real and generated data. This adversarial process, where the generator aims to fool the discriminator, and the discriminator strives to improve its detection accuracy, results in the generation of increasingly realistic images over time. GANs have been widely used in various applications, including image synthesis, style transfer Isola et al. (2017), and super-resolution

Ledig et al. (2017), due to their ability to produce high-quality, detailed images that are often indistinguishable from real photographs.

## 1.2 Stable Diffusion

Stable Diffusion is a more recent advancement in the field of image synthesis, falling under the category of diffusion models. Diffusion models generate images by iteratively refining a noisy initial image through a series of steps, gradually enhancing the quality and reducing the noise to produce a coherent, realistic final image. Stable Diffusion specifically leverages a probabilistic framework to ensure the stability and consistency of the generated images throughout the refinement process.

Stable Diffusion incorporates textual prompts into its image generation process, a feature absent in GANs. Consequently, in your workflow, GANs generate images unconditionally, whereas Stable Diffusion produces outputs that are conditioned on textual inputs. This distinction underscores Stable Diffusion's ability to align generated images closely with specific textual descriptions, offering greater control and precision in the synthesis of images from textual prompts.

In the context of text-to-image generation, Stable Diffusion leverages textual prompts to guide the image creation process. Users input a text description, and the model generates an image that visually represents the described scene or object. This capability is particularly valuable in applications requiring precise and contextually accurate image synthesis, such as digital art, media content creation, and various forms of visual storytelling. The model achieves this by encoding the textual information into a latent space, which then conditions the image generation process to ensure the output aligns with the given description. This approach allows for fine control over the generated image content and style, making it a powerful tool for creating detailed and realistic images from textual prompts (Zhang et al., 2023).

Prompts can be classified as either positive or negative. Positive prompts specify elements that should be included in the generated image, while negative prompts indicate aspects to avoid. For example, a positive prompt might be "Close up face", instructing the model to generate an image focusing on a close-up view of a face. Conversely, a negative prompt like "sketch, cartoon, drawing, anime" guides the model to avoid these styles.

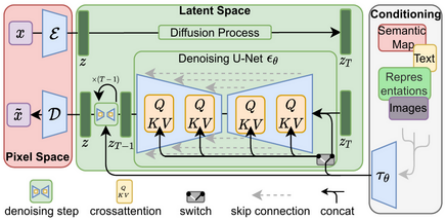



**Figure 1.3: Architecture diagram of a Stable Diffusion model. The process flows from Pixel Space through a Latent Space with a Diffusion Process and Denoising U-Net, to Conditioning incorporating semantic information. The model uses techniques like cross-attention, skip connections, and concatenation to generate high-quality images from text prompts or other inputs (AIModels, 2024).**

Additionally, Classifier-Free Guidance (CFG) is a parameter that balances the fidelity of the image to the prompt and the overall quality of the generated image. A higher CFG value makes the model adhere more strictly to the prompt, potentially at the cost of image quality, while a lower CFG value allows for more flexibility, potentially improving image aesthetics but with less adherence to the prompt details. For instance, a CFG value of 7 might ensure strict adherence to the prompt, while a value of 5 offers a balance between adherence and quality Zhang et al. (2023).

## 1.3 Challenges in Fake Image Detection

The field of image classification has seen significant advancements with the development of technologies such as Generative Adversarial Networks (GANs) and diffusion models (Papa et al., 2023). Researchers explore the use of Stable Diffusion for generating realistic face images and the subsequent challenges in detecting these synthetic images. Their study highlights the sophistication of diffusion models in creating lifelike images and the critical need for effective detection mechanisms to distinguish these images from real ones. They emphasize the role of prompt engineering and human-in-the-loop strategies in enhancing detection accuracy, which are vital in the fight against deepfakes.

The detection of fake images remains a challenging task, particularly when models are expected to generalize across different types of generative models. The study by (Ojha et al., 2023) on developing universal fake image detectors that generalize across generative models revealed limitations in accuracy, especially for detecting fake images. Their results indicate that while the model could identify real images with reasonable accuracy, it struggled significantly with distinguishing between different types of synthetic images, leading to lower overall performance.This highlights the need for more robust detection methods.

In contrast, the method proposed by (Guarnera et al., 2023) shows better overall accuracy in fake image detection. Their approach involves a two-part system: first, distinguishing between AI-generated and real images, and then, if the image is AI-generated, further classifying it as being produced by either GAN or Stable Diffusion. This hierarchical approach allows for more precise identification and has demonstrated higher accuracy rates compared to other methods, particularly in the challenging task of discriminating between different types of synthetic images. (Wang & Deng, 2021) provide a comprehensive survey on deepfake recognition, detailing various methods and their performances. They discuss the influence of different training data and model architectures on the effectiveness of face recognition systems. This survey underscores the importance of diverse and representative training datasets, noting that, due to the inherent biases in the training data composition, the models tend to exhibit bias. The survey provides examples of female, black, and younger cohorts being detected with less accuracy. In this thesis I will be focusing on the biases arising from the CNN classification tasks and I will be solely focusing on SD and GAN generated image detection. I will also be using a separate dataset from FairFace (Karkkainen & Joo, 2021). Different from Wang & Deng (2021), the focus will be on SD and GAN deepfake detection.

Despite the advances in detection methods, there are still significant gaps. Current methods often fail to generalize well across different generative models, and there is a lack of research focusing specifically on demographic biases in fake image detection. This thesis aims to fill these gaps by providing a detailed analysis of gender bias in the classification of real and synthetic images generated by GAN and Stable Diffusion models. My approach differs from existing studies by not only assessing the performance of CNN and ResNet-18 models but also by conducting a thorough statistical analysis of the observed biases.

The detection of deepfakes and other forms of synthetic media is crucial for several reasons. Firstly, deepfakes can be used to spread misinformation and disinformation, significantly impacting public opinion and trust. For instance, convincingly altered videos and images can be used in political campaigns to damage reputations or manipulate election outcomes. Secondly, deepfakes pose a serious threat to personal privacy and security. Individuals can be impersonated in compromising situations, leading to potential blackmail or psychological harm. Additionally, deepfakes undermine the authenticity of digital media, making it increasingly difficult to discern real content from fake. This erosion of trust in visual and audio media can have far-reaching consequences for journalism, legal proceedings, and social interactions. Therefore, developing robust methods for detecting fake images is essential to maintain the integrity and reliability of information in our increasingly digital world.

By addressing these gaps and focusing on demographic biases, this thesis aims to contribute to the development of fairer and more reliable AI systems. The unique contribution of this research lies in its dual focus on the performance disparities between male and female images and the application of statistical methods to analyze these biases. This comprehensive approach not only enhances our understanding of demographic biases in image classification but also provides practical insights for improving the fairness and accuracy of AI-generated image detection systems.

# 2 Datasets Used

## 2.1 GAN and Real Faces Dataset

I used the publicly available Flickr dataset (XHLULU, 2024) for real people faces, which contains a total of 77 856 images of both male and female faces. For the GAN face images, I used a publicly available dataset on Kaggle (TUNGUZ, 2024) that contains 1 million images. However, I used only the


**Figure 1.4: A diverse selection of nine facial portraits randomly sampled from the FairFace dataset, showcasing a range of ages, ethnicities, and genders to represent the dataset's broad demographic coverage.(Karkkainen & Joo, 2021)**

first 77 856 images to ensure equal sample sizes for both GAN and real faces. I reserved 1000 images from both GAN and real face datasets exclusively for bias measurements, and these images are not used in any part of model training or validation and also not used in accuracy testing. I sized all GAN and real images to $1024 \times 1024$.

## 2.2 Stable Diffusion Faces Dataset

I generated the Stable Diffusion dataset for this paper using the open-source repository "Automatic 1111" (36DB et al., 2024). I performed the computations on the University of Groningen's computational cluster Hábrók. Information about the Hábrók can be found here. I generated 9319 images for training and an additional 4000 images specifically for bias measurements. I generated all images at a resolution of $1024 \times 1024$. When I inspected the real image dataset I planned to use (Flikr dataset, XHLULU (2024)), I realized that approximately one-sixth of the images had people with glasses on their faces. To prevent the models from learning that people with glasses are always real, I ensured that one-sixth of the Stable Diffusion images were generated with the phrase "with glasses" added to the positive prompt.

I used two models with equal amounts in the stable diffusion faces dataset generation. Both models were taken from a popular model sharing website (Civitai, Year of access: 2024). The first model I used is named Realistic Vision (RV) v60B1 (SG_161222, Year of access: 2024), and the second is Realistic Stock Photo (RSP) v20 (PromptSharingSamaritan, Year of access: 2024). I chose these two models because they were the most popular models downloaded in the realism context. Because the models had different weights, the prompts I used for generation differed slightly. The prompts used for image generation for both models can be seen in TableTable 2.1. The number of images with the prompt "with glasses" is the same for all parts of the sets.

To ensure unique images, I generated seeds with a script that ensures each seed is used only once. I visualized example images from the training dataset in Figure 1.1.

## 2.3 FairFace Dataset

In this paper I have also used an external dataset for the purpose of bias measurement. I used the FairFace Dataset (Karkkainen & Joo, 2021) because it is a labeled dataset with gender, race and age. Example images from this dataset can be found at Figure 1.4. This dataset had 2 versions in which the image sizes were different. I used the version with higher resolution. The set had already train and validation split. For bias testing only the train set was used and it had 86 745 images.

## 2.4 Data Splits and Normalization

I meticulously divided the dataset into training, validation, test sets, and a bias measurement subset to ensure robust model evaluation and mitigate potential biases. As shown in Table 2.3 and Table 2.2, I included 130 000 images in the training set, dedicating 83% to training the model on a balanced distribution of real and GAN-generated images. Similarly, I maintained this balance in the validation and test sets, which comprised 8000 and 15 712 images respectively, to validate and assess the generalization of the model effectively. Additionally, I specifically generated a bias measurement subset of 2000 images to evaluate the model's performance on synthetic data, ensuring comprehensive assessment across different data distributions.

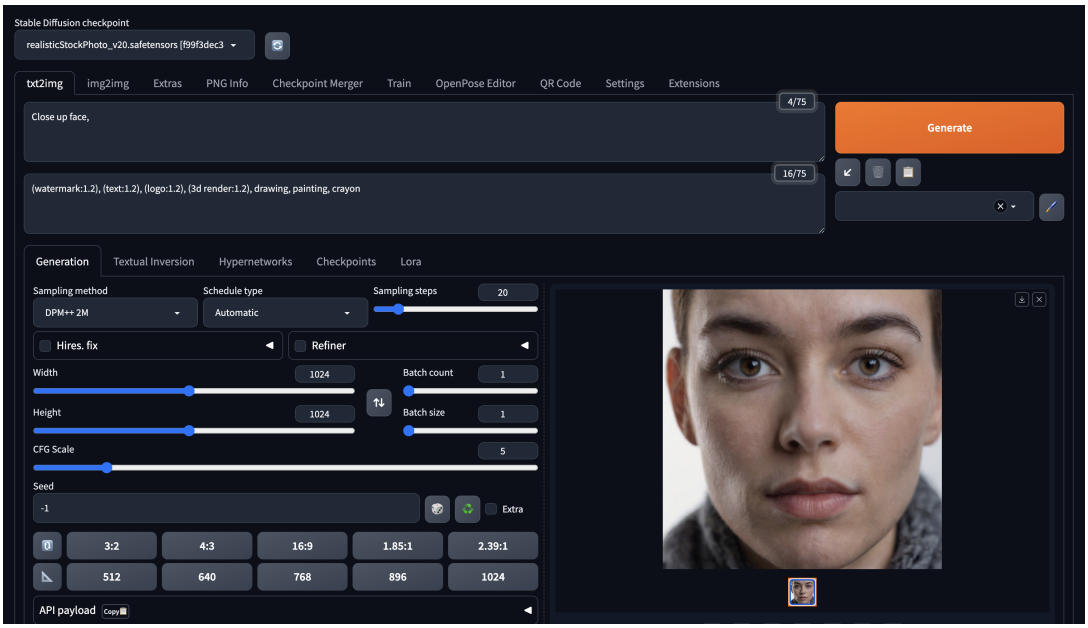



**Figure 2.1: Screenshot of the Automatic1111 web interface for Stable Diffusion, showing the generation of a close-up face image. The interface displays various controls for text prompts, negative prompts, sampling methods, and image parameters, alongside the generated output.**

To standardize the dataset and mitigate potential biases arising from varying resolutions, I uniformly downscaled all images to a resolution of $256 \times 256$ pixels. I have used LANCZOS (Burger & Burge, 2009) method in every downscaling application in order to make sure the models do not solely learn the artifacts created by the downscaling algorithm. This downscaling process ensured consistent image quality across the dataset. Notably, this standardization was achieved using the LANCZOS method, an interpolation technique known for preserving image details and producing high-quality downscaled images. The LANCZOS method achieves its high-quality downsampling by employing a mathematical algorithm that considers a weighted average of nearby pixels when generating new pixel values. This approach ensures that each pixel in the downscaled image is derived from a combination of multiple pixels in the original image, preserving important details and minimizing the loss of information. Additionally, the method uses a windowed sinc function to interpolate pixel values, resulting in smooth transitions between neighboring pixels and reducing artifacts such as jagged edges or blurring. Overall, these characteristics of the LANCZOS method contribute to its effectiveness in maintaining image clarity and preserving fine details even after downsampling.

# 3 Methods and Model Architecture

## 3.1 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) I used in this study consists of several key layers that work together to classify images as either real or fake. The input to the network is an image with three color channels (RGB). The first layer is a convolutional layer with 32 filters of size $3 \times 3$ and a padding of 1, which ensures that the spatial dimensions of the input are preserved. This layer is followed by a ReLU (Rectified Linear Unit) activation function. ReLU is a non-linear activation function that introduces non-linearity to the model, allowing it to learn complex patterns in the data more effectively.

The output of the first convolutional layer is passed to a max-pooling layer with a 2x2 filter and

| Parameter | Realistic Vision (RV) v60B1 |
|---|---|
| Positive Prompt | Close up face |
| Negative Prompt | sketch, cartoon, drawing, anime |
| Sampling Steps | 20 |
| CFG | 7 |

| Parameter | Realistic Stock Photo (RSP) v20 |
|---|---|
| Positive Prompt | Close up face |
| Negative Prompt | (watermark:1.2), (text:1.2), (logo:1.2), (3d render:1.2), drawing, painting, crayon |
| Sampling Steps | 20 |
| CFG | 5 |

**Table 2.1: Parameter settings for RV and RSP**

where, parentheses mean a weight for the given keyword is given (Higher is more important), Sampling Steps indicate the number of iterations for refining the image, and CFG (Classifier-Free Guidance) controls the trade-off between adherence to the prompt and image quality (Higher values make the model adhere more closely to the prompt). I added "with glasses" to the positive prompt for the images with glasses.

| Set | Images | Percentage of Total |
|---|---|---|
| Training Set | 14 910 | 80% (50% real, 25% RV(SD), 25% RSP(SD)) |
| Validation Set | 1864 | 10% (50% real, 25% RV(SD), 25% RSP(SD)) |
| Test Set | 1864 | 10% (50% real, 25% RV(SD), 25% RSP(SD)) |
| Bias Measurement | 4000 | Additional SD generated (50% RV(SD), 50% RSP(SD)) |

**Table 2.2: Data Split For SD Vs Real, CNN and TL18**

| Set | Images | Percentage of Total |
|---|---|---|
| Training Set | 130 000 | 83% (50% real, 50% GAN) |
| Validation Set | 8000 | 5% (50% real, 50% GAN) |
| Test Set | 15 712 | 10% (50% real, 50% GAN) |
| Bias Measurement | 2000 | 1% (50% real, 50% GAN) |

**Table 2.3: Data Split For GAN Vs Real, CNN and TL18**

a stride of 2. Max-pooling is a pooling operation that reduces the spatial dimensions of the feature map by taking the maximum value from each patch of the feature map.

The next layer is another convolutional layer with 64 filters of size $3 \times 3$ and padding of 1, again followed by a ReLU activation function and a subsequent max-pooling layer with the same parameters as before. These convolutional layers extract higher-level features from the input image through their learned filters, while the max-pooling layers downsample the feature maps, reducing computational complexity and focusing on the most important features.

The feature maps produced by the second convolutional layer are then flattened into a single vector, which serves as the input to a fully connected layer with 128 units and a ReLU activation function. Fully connected layers integrate the features extracted by the convolutional layers and make the

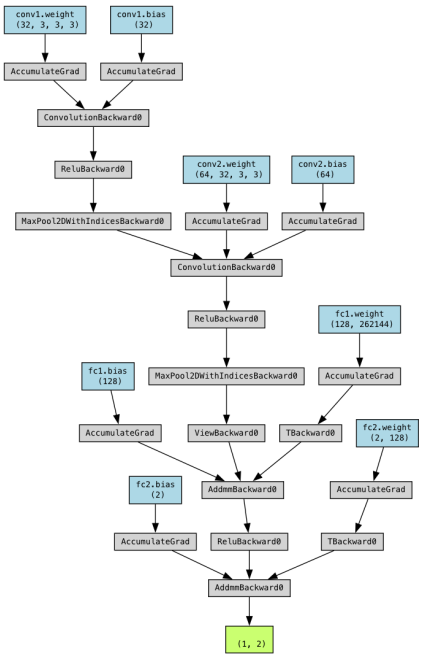



**Figure 3.1: Neural network backpropagation graph using torchviz (Zagoruyko, 2020), illustrating gradient flow through convolutional and fully connected layers.**

final decision based on these features.

Another fully connected layer with 2 output units follows, corresponding to the two classes: real and fake. The final output layer produces a score for each class, determining the classification of the input image.

This architecture, implemented in PyTorch (Paszke et al., 2017), is designed to effectively learn spatial hierarchies in the image data through its convolutional and pooling layers, and to make classification decisions based on the extracted features through its fully connected layers.

The visual for the CNN architecture used can be found in Figure 3.1.

## 3.2 Transfer Learning with ResNet-18 (TL18)

Additionally, I utilized transfer learning by adapting a pre-trained ResNet-18 model (He et al., 2015). ResNet-18, a deep neural network architecture pre-trained on a large dataset for image recognition tasks, was repurposed for the classification purposes. In this adaptation, the pre-trained parameters were frozen, and a new classification layer was added. By freezing the pre-trained parameters, the model retained the knowledge acquired from the original dataset while fine-tuning its ability to classify real and synthetic images accurately. While ResNet-18 comprised numerous layers, the focus of modification was primarily on incorporating the new classification layer, ensuring that the depth and complexity of the original architecture were maintained. This approach allowed for effective leveraging of pre-existing knowledge while tailoring the model to the specific classification task at hand.

For GAN images and SD images both CNN and TL18 models are trained. Each with equal amount of deepfake and real images.

## 3.3 Early Stopping

To prevent overfitting and ensure optimal model performance, I employed early stopping during the training of both the Convolutional Neural Network (CNN) and the transfer learning model based on ResNet-18 (TL18) (Prechelt, 1998). Early stopping is a technique where training is halted if the model's performance on a validation set does not improve after a specified number of epochs, referred to as "patience".

For this study, a patience value of 3 was used. This means that if the model's validation accuracy did not improve for 3 consecutive epochs, the training process was stopped. By doing so, the models were able to generalize better on unseen data, avoiding the common pitfall of overfitting to the training set. Early stopping also contributed to more efficient training by reducing unnecessary computation after the model had essentially converged.

## 3.4 Two Proportion Z-Test

To statistically evaluate the differences in classification accuracy between male and female images, I conducted a Two Proportion Z-Test. This test assesses whether there is a significant difference between the proportions of two groups—in this case, the classification accuracies for male and female images.

The Two Proportion Z-Test is formulated as follows:

- **Null Hypothesis ($H_0$)**: There is no difference in the classification accuracies between male and female images ($p_1 = p_2$).
- **Alternative Hypothesis ($H_a$)**: There is a difference in the classification accuracies between male and female images ($p_1 \neq p_2$).

Where $p_1$ is the proportion of correctly classified male images, and $p_2$ is the proportion of correctly classified female images.

The test statistic $Z$ for the Two Proportion Z-Test is calculated using the formula:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Here,

- $p_1$ and $p_2$ are the sample proportions of correctly classified male and female images, respectively.
- $n_1$ and $n_2$ are the sample sizes of male and female images, respectively.
- $\hat{p}$ is the pooled sample proportion, calculated as:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

where $x_1$ and $x_2$ are the number of correctly classified male and female images, respectively.

## 3.5 Saliency Maps

In this study, I implemented Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to interpret the decisions of the convolutional neural network (CNN) model used for image classification. The CNN model, pretrained on a dataset of real and fake images, distinguishes between these two classes. I began by preprocessing each image, converting it into a tensor and normalizing its pixel values. Next, I applied Grad-CAM to generate heatmaps that visually highlight the image regions crucial for the model's classification decisions. This method provided valuable insights into the CNN's behavior, revealing the specific areas of each image that significantly influenced its predictions.

# 4 Results

## 4.1 Model Training

According to the plan of this paper 4 models are trained. The number of epochs they are trained for are shown on table Table 4.1. As mentioned in the methods section, a patience value of 3 was used for the training of the models for early stopping. Training and Validation Loss Charts of the models can be seen on Figure 4.1.

| CNN/TL | GAN - REAL | SD - REAL |
|---|---|---|
| CNN | 7 | 4 |
| TL18 | 10 | 20 |

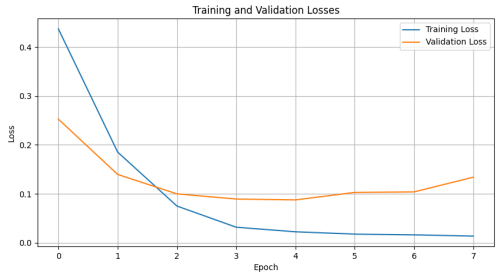
**Table 4.1: Table for number of epochs the models trained.**

### 4.1.1 CNN Models

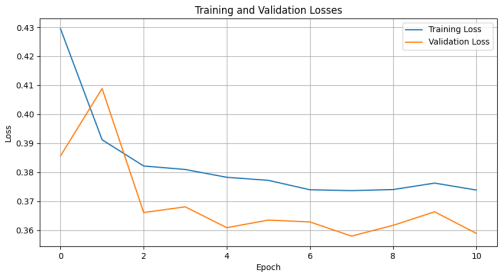
As can also be seen from the Table 4.1, models have been trained with 7 and 4 epochs, respectively for GAN-REAL and SD-REAL training. For the GAN-REAL training the reason for the early stop is the early stopping rule with a patience value of 3. Because the loss value of the model kept increasing for 3 times consecutively, the training stopped at $8^{\text{th}}$ epoch and took the 7 epoch trained model. As for the SD-REAL training this was not the case. The training stopped because the loss of the model had reached 0 and therefore the training could not proceed further, stopping at $4^{\text{th}}$ epoch.

### 4.1.2 Transfer Learning Models

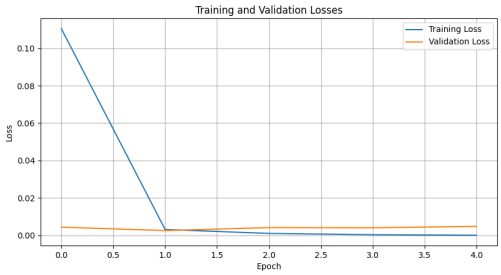
In the transfer learning models implemented with the Resnet-18 model, the number of epochs they were trained on were 20 and 10 for SD-REAL and GAN-REAL, respectively. In SD-REAL no stopping was applied and the target 20 epochs was achieved. However in GAN-REAL, the training stopped at 10 epochs without an error and not with early stopping rule. The reason for the stop is still unknown, despite re-running the experiment several times and conducting thorough debugging.



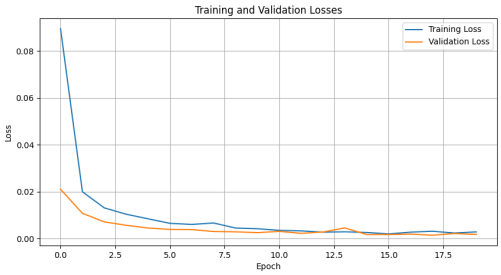

**(a) Training and Validation Loss Chart for GAN-Real training on basic CNN.**




**(b) Training and Validation Loss Chart for GAN-Real training on Resnet-18.**




**(c) Training and Validation Loss Chart for SD-Real training on basic CNN.**




**(d) Training and Validation Loss Chart for SD-Real training on Resnet-18.**

**Figure 4.1: Training and validation loss charts showing loss values on the y-axis and epochs on the x-axis. Yellow lines represent losses in validation and blue lines represent training losses.**

## 4.2 Statistical Tests

### 4.2.1 Supplementary Set

After training the models, they are evaluated using a supplementary dataset reserved specifically for this purpose. This dataset, which is manually labeled, allows for a two-proportion Z-Test to be conducted on the results. For each gender, the total number of images from both the fake and real sets is combined and used in the analysis. Likewise, the total number of incorrect predictions for both sets, separated by gender, is summed and utilized in the Z-Test. These values can be observed in the Table 4.2.

### 4.2.2 FairFace Dataset

Similarly the two-proportion Z-Test is also performed for the FairFace Dataset. Models predictions are taken for each image and because the FairFace set consists of real people whenever a "fake" output is received it is considered a wrong answer. The proportions of the results can be observed on the Figure 4.2. The table that has the p-values for the Z-Test is in Table 4.3 and the table with numeric raw data is Table 4.9.

## 4.3 Statistical testing

### 4.3.1 Bias Testing Test

I ran the trained models on the supplementary set, which I did not use in the training or testing of the model. I then obtained the number of correct and incorrect classifications and used those numbers to calculate the p-value for the two proportion z-test. The results are on Table 4.2. It can be seen that all of the p-values are below 0.05 meaning that we reject the null hypothesis. Except for the GAN-REAL Transfer Learning model, all of the models showcased higher accuracy towards female images. For the GAN-REAL Transfer learning model this was the opposite with a higher accuracy towards male images.

| Model | Proportion Males (Correct) | Proportion Females (Correct) | P-value |
|---|---|---|---|
| GAN-REAL CNN | 0.9684 | 0.9830 | **0.032 300** |
| SD-REAL CNN | 0.9894 | 0.9969 | **0.001 300** |
| GAN-REAL TL | 0.9153 | 0.8448 | **0.000 002** |
| SD-REAL TL | 0.9379 | 0.9976 | **0.000 000** |

**Table 4.2: Bias assessment for the models using the supplementary set that is labeled by hand.**

#### 4.3.2 FairFace Dataset

I ran the trained models on the FairFace dataset, which is an external dataset with gender labels. After getting the number of correct and incorrect classifications with the models, I have performed a two proportion z-test and its results are on Table 4.3. Except for the GAN-REAL CNN model every model had a significant p-value with 0.05. SD-REAL Transfer learning had a higher accuracy for males and SD-REAL CNN and GAN-REAL TL had a higher accuracy towards females.

### 4.4 Saliency Maps

Saliency maps for the CNN models are on Figure 4.3. As it can be seen from the figure SD had more points on the face compared to the GAN classifier. Furthermore, the GAN model mostly focuses on the specific points of the face where in the SD model it is a bit different. In the SD model the model also often checks the background of the image. This might be because the stable diffusion models tend to keep the background clean as nothing is specified in the prompt. Although not always, this is mostly the case.

## 5 Discussion

### 5.1 Biases

Overall, the analysis reveals significant biases in classification accuracy between male and female images across different models and datasets. The GAN-REAL CNN model shows a significant bias towards females in the supplementary dataset but not in the FairFace dataset. Conversely, the GAN-REAL Transfer Learning model demonstrates a significant bias towards males in both datasets. Similarly, the SD-REAL CNN and SD-REAL Transfer Learning models exhibit significant biases, favoring females and males, respectively, in both datasets.

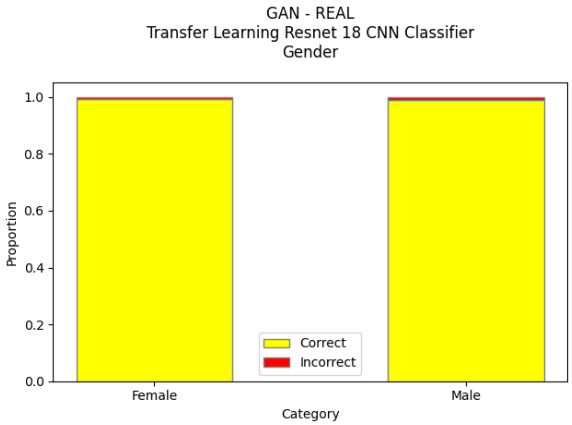
Considering the perspective of fake image generation techniques, my hypothesis holds true to some extent. It appears that female images exhibit higher accuracies in classifying real and fake images for the Stable Diffusion models. However, this pattern does not consistently apply to GAN-generated images, where biases towards male or female images vary between models. Therefore, it is challenging to assert a consistent higher accuracy for female images across all scenarios.

### 5.2 Saliency Maps

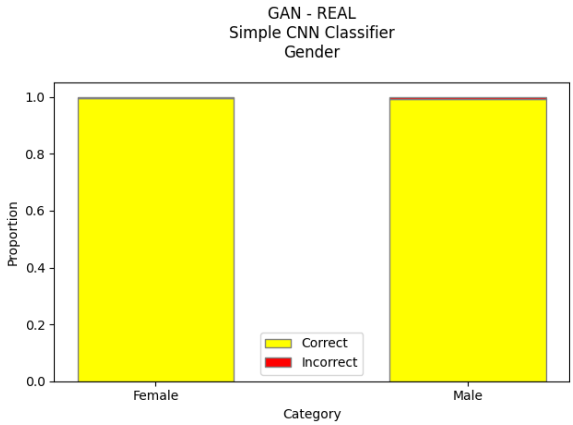
As mentioned in the results section Section 4.4 SD also considers the background of the image GAN seems to be focusing on the deformities on the face of the person in the image. To avoid this before the training one of the ideas was to implement a rotating set of background prompts though that method then ruled out because the model could also learn specific backgrounds as real or fake directly. For further research a solution to this problem might be looked for. Perhaps instead of 10-20 different environmental prompts a huge number of it could be utilized.

### 5.3 Limitations

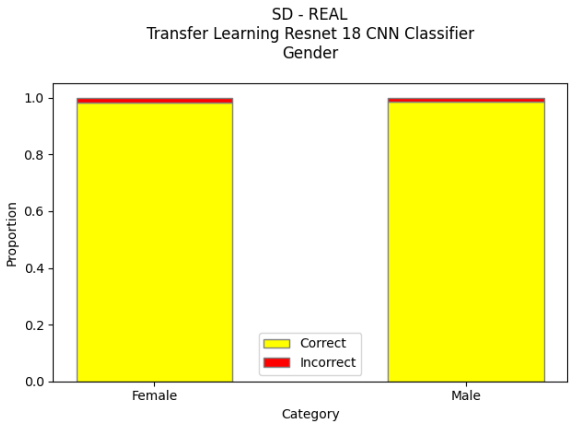
I have observed that the dataset I have created had more female images for the stable diffusion dataset. This might be because of the median human perspective of the stable diffusion models. Different models have different median points and the realistic models that I used generated mostly female faces when prompted with “Close up face”. Though this situation was different for the “with glasses” prompt. The amount of male and female was more equalized when a glasses prompt is added.



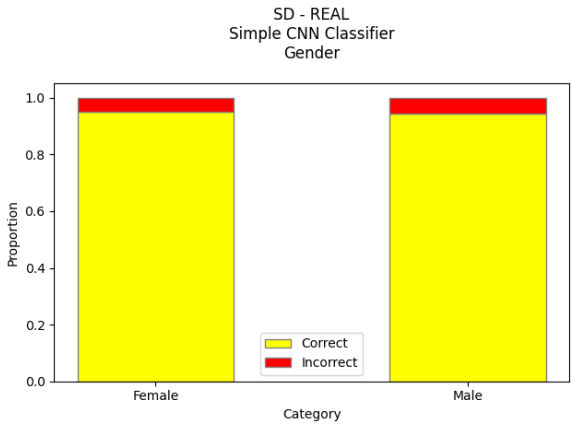

**(a) Gender Accuracy proportions for GAN-Real Resnet-18 classifier.**




**(b) Gender Accuracy proportions for GAN-Real basic CNN classifier.**




**(c) Gender Accuracy proportions for SD-Real Resnet-18 classifier.**




**(d) Gender Accuracy proportions for SD-Real basic CNN classifier.**

**Figure 4.2: Gender Accuracy proportions for the trained models. Tested on FairFace Dataset**

So I think that a more accurate result regarding the gender biases in this field could be done with an ensured equal male female amounts in the training of the models. This way I think a more reliable result can be obtained even though I suspect biases will also be present in that case.

## 5.4 Future Work

As discussed in Section 5.2, stable diffusion models often produce images with plain, typically white backgrounds. This characteristic causes the models to focus more on the background in the SD versus real image classification task, rather than on facial deformities introduced by the SD algorithms. To address this, future research could explore solutions such as incorporating a list of diverse environments into the positive prompts during SD generation. By mitigating this issue, more effective models could be developed.

# References

36DB, AUTOMATIC1111, C43H66N12O12S2, CodeHatchling, EllangoK, KohakuBlueleaf, ... yfszzx (2024). *Automatic1111/stable-diffusion-webui.* https://github.com/AUTOMATIC1111/

| Model | Proportion Males (Correct) | Proportion Females (Correct) | P-value |
|---|---|---|---|
| GAN-REAL CNN | 0.9940 | 0.9945 | 0.342 900 |
| SD-REAL CNN | 0.9430 | 0.9497 | **0.000 013** |
| GAN-REAL TL | 0.9905 | 0.9920 | **0.022 100** |
| SD-REAL TL | 0.9863 | 0.9836 | **0.001 200** |

**Table 4.3: Bias assessment in model predictions based on gender with the FairFace Dataset. Proportions and p-values.**


stable-diffusion-webui. (Accessed: 2024-04-13)

AIModels. (2024). *Brain mri axial slices generative diffusion.* https://aimodels.org/ai-models/healthcare-medical-imaging/brain-mri-axial-slices-generative-diffusion/. (Accessed: 2024-07-15)

Burger, W., & Burge, M. J. (2009). *Principles of digital image processing: Core algorithms.* Springer.

Civitai. (Year of access: 2024). *Stable diffusion models.* https://civitai.com/. (Accessed: 2024-04-13)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *arXiv preprint arXiv:1406.2661*.

Guarnera, L., Giudice, O., & Battiato, S. (2023). Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *arXiv preprint arXiv:2303.00608*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.* Retrieved from https://arxiv.org/abs/1512.03385

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134).

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1548–1558).

Kyang, Y. (2024). *Deep learning gan (generative adversarial networks).* https://medium.com/@kyang3200/deep-learning-gan-generative-adversarial-networks-67ba72173d0c. (Accessed: 2024-07-15)

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4681–4690).

Ojha, U., Li, Y., & Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.*

Papa, L., Faiella, L., Corvitto, L., Maiano, L., & Amerini, I. (2023). On the use of stable diffusion for creating realistic faces: from generation to detection. In *2023 11th international workshop on biometrics and forensics (iwbf)* (pp. 1–6). Barcelona, Spain. doi: 10.1109/IWBF57495.2023.10156981

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in pytorch. In *Nips-w.*

Prechelt, L. (1998). Early stopping - but when? *Neural Networks*, *11*(4), 761–767.

PromptSharingSamaritan. (Year of access: 2024). *Realistic stock photo.* https://civitai.com/models/139565/realistic-stock-photo. (Accessed: 2024-04-13)

| Result | GAN Male | GAN Female | Real Male | Real Female |
|---|---|---|---|---|
| Total Samples | 436 | 564 | 449 | 551 |
| Correct | 410 | 545 | 447 | 551 |
| Incorrect | 26 | 19 | 2 | 0 |
| Percentage | 94.04% | 96.63% | 99.55% | 100% |

**(a) GAN - Simple CNN Classifier**

| Result | GAN Male | GAN Female | Real Male | Real Female |
|---|---|---|---|---|
| Total Samples | 436 | 564 | 449 | 551 |
| Correct | 361 | 480 | 449 | 462 |
| Incorrect | 75 | 84 | 0 | 89 |
| Percentage | 82.80% | 85.11% | 100% | 83.85% |

**(b) GAN - Transfer Learning Resnet 18 CNN Classifier**

| Result | SD Male | SD Female | Real Male | Real Female |
|---|---|---|---|---|
| Total Samples | 2000 | 2000 | 449 | 551 |
| Correct | 1974 | 1992 | 449 | 551 |
| Incorrect | 26 | 8 | 0 | 0 |
| Percentage | 98.70% | 99.60% | 100% | 100% |

**(c) SD - Simple CNN Classifier**

| Result | SD Male | SD Female | Real Male | Real Female |
|---|---|---|---|---|
| Total Samples | 2000 | 2000 | 449 | 551 |
| Correct | 1848 | 1994 | 449 | 551 |
| Incorrect | 152 | 6 | 0 | 0 |
| Percentage | 92.40% | 99.70% | 100% | 100% |

**(d) SD - Transfer Learning Resnet 18 CNN Classifier**

**Table 4.4: Numeric Results for Supplementary Data.**

| Result | Real Male | Real Female |
|---|---|---|
| Total Samples | 45 986 | 40 758 |
| Correct | 45 712 | 40 535 |
| Incorrect | 274 | 223 |
| Percentage | 94.40% | 99.45% |

**Table 4.5: GAN - Simple CNN Classifier**

| Result | Real Male | Real Female |
|---|---|---|
| Total Samples | 45 986 | 40 758 |
| Correct | 45 549 | 40 430 |
| Incorrect | 437 | 328 |
| Percentage | 99.05% | 99.20% |

**Table 4.6: GAN - Transfer Learning Resnet 18 CNN Classifier**

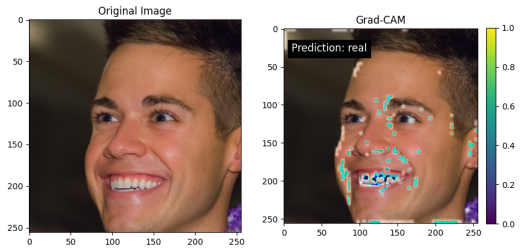
| Result | Real Male | Real Female |
|---|---|---|
| Total Samples | 45 986 | 40 758 |
| Correct | 43 364 | 38 707 |
| Incorrect | 2622 | 2051 |
| Percentage | 94.30% | 94.96% |

**Table 4.7: SD - Simple CNN Classifier**

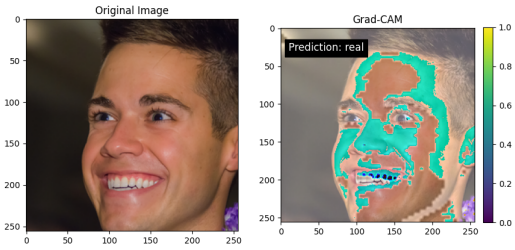
| Result | Real Male | Real Female |
|---|---|---|
| Total Samples | 45 986 | 40 758 |
| Correct | 45 356 | 40 091 |
| Incorrect | 630 | 667 |
| Percentage | 98.63% | 98.36% |

**Table 4.8: SD - Transfer Learning Resnet 18 CNN Classifier**

**Table 4.9: Numeric Results for FairFace Dataset**




**(a) Saliency map of GAN-REAL CNN via Gradcam Implementation.**




**(b) Saliency map of SD-REAL CNN via Gradcam Implementation.**

**Figure 4.3: Comparison of saliency maps for GAN-REAL CNN and SD-REAL CNN via Gradcam Implementation. Heatmaps highlight the regions of the image that were most influential in the model's decision-making process. Warmer colors (closer to yellow) indicate higher importance, while cooler colors (closer to blue) indicate lower importance. The prediction label for the image is displayed in the top left corner of the heatmap.**

**Figure 5.1: Example image generated with the "Close up face" positive prompt white background mentioned in the Section 5.4.**


Rawal, A., Dietrich, S. L., & McCoy, J. (2024). *Explainable artificial intelligence for bias identification and mitigation in demographic models* (Working Paper No. SEHSD-WP2024-02). U.S. Census Bureau. Retrieved from https://www.census.gov/library/working-papers/2024/demo/SEHSD-WP2024-02.html

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.*

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

SG_161222. (Year of access: 2024). *Realistic vision v6.0 b1.* https://civitai.com/models/4201/realistic-vision-v60-b1. (Accessed: 2024-04-13)

TUNGUZ, B. (2024). *1 million fake faces on kaggle.* https://www.kaggle.com/c/deepfake-detection-challenge/discussion/121173. (Accessed: 2024-04-13)

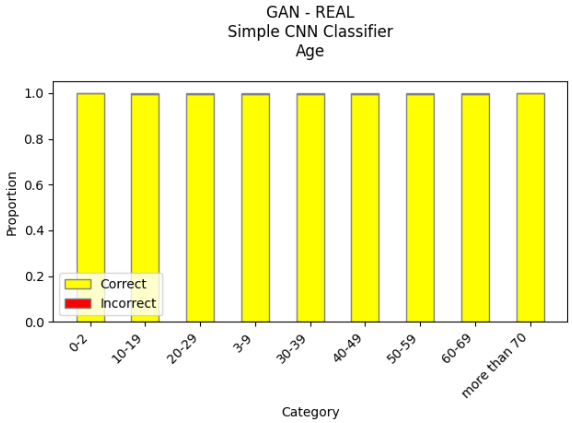
Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244. doi: 10.1016/j.neucom.2020.10.081

XHLULU. (2024). *70k real faces.* https://www.kaggle.com/c/deepfake-detection-challenge/discussion/122786. (Accessed: 2024-04-13)

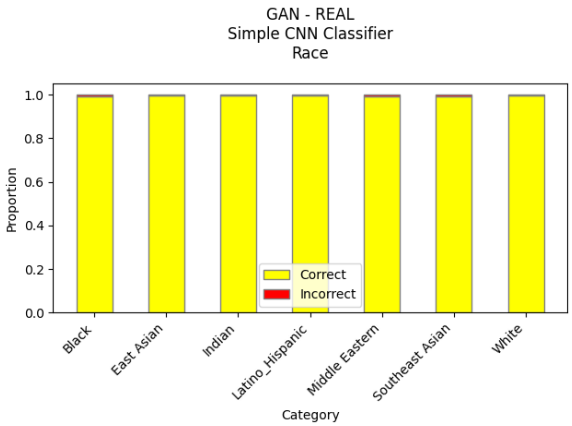
Zagoruyko, S. (2020). *pytorchviz.* Retrieved from https://github.com/szagoruyko/pytorchviz (GitHub repository)

Zhang, B., Kim, T., Kim, D., Chai, Y., Fu, Y., & Li, C. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 18120-18129).

# A Appendix

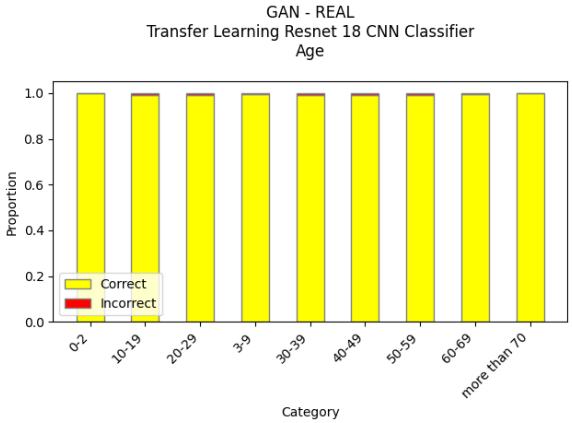



**(a) Age**

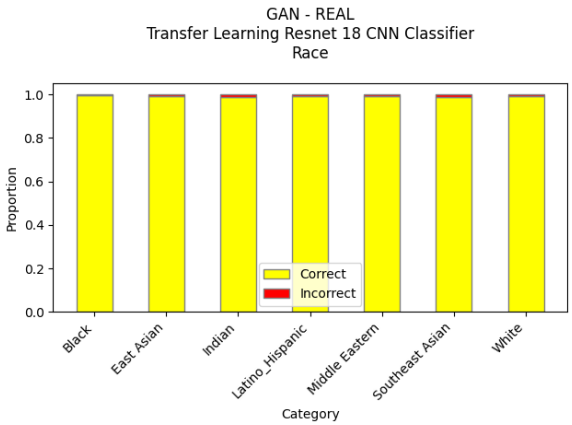



**(b) Race**

**Figure A.1: Additional tests with FairFace dataset for GAN-REAL CNN model.**

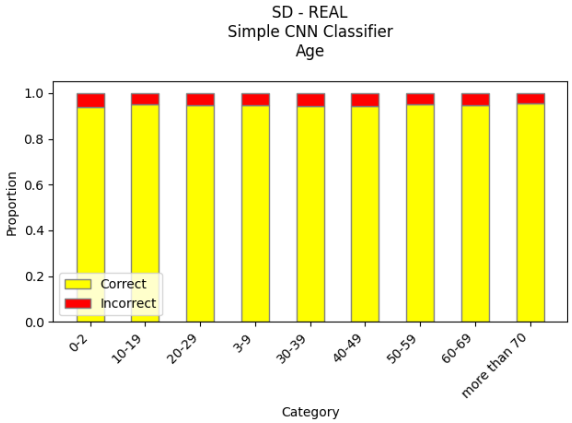



**(a) Age**

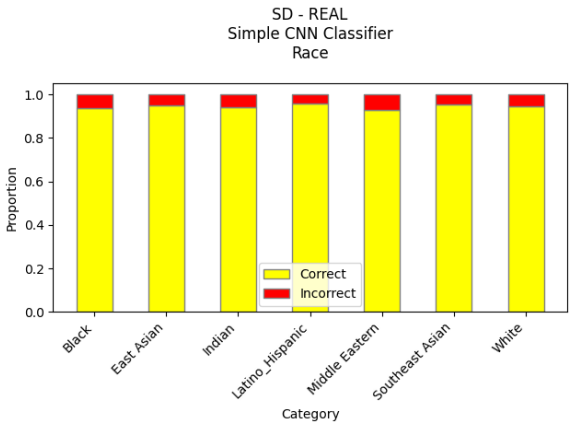



**(b) Race**

**Figure A.2: Additional tests with FairFace dataset for GAN-REAL TL model.**

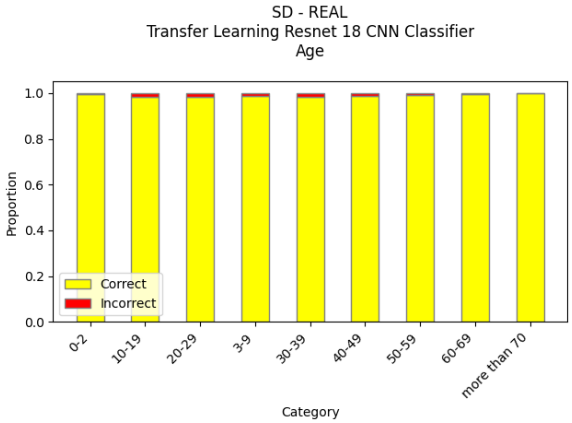



**(a) Age**

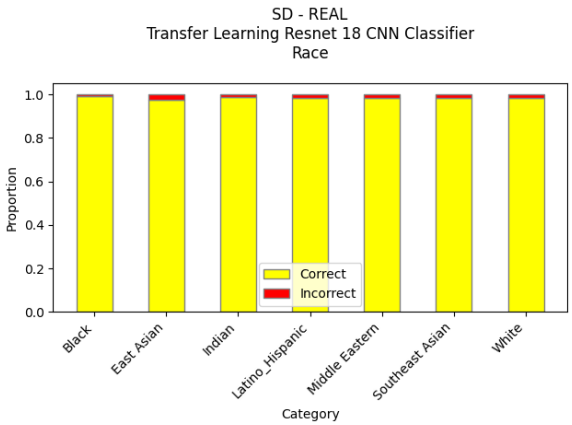



**(b) Race**

**Figure A.3: Additional tests with FairFace dataset for SD-REAL CNN model.**




**(a) Age**




**(b) Race**

**Figure A.4: Additional tests with FairFace dataset for SD-REAL TL model.**