

Investigation of Visual Bias in Generative AI

Jerome Agius

Supervisor: Prof. Dylan Seychell

Co-Supervisor: Dr John Abela

June 2024

*Submitted in partial fulfilment of the requirements
for the degree of Bachelor of Science in Information Tech (Hons) - Artificial
Intelligence.*



L-Università ta' Malta

Faculty of Information &
Communication Technology

Abstract

In the realm of Artificial Intelligence (AI), the emergence of text-to-image generators, such as Stable Diffusion, Dall-E-3 and Midjourney has brought about new avenues for creativity. However, as with any innovation concerns have been raised in regards to the presence of bias within images generated by such means, particularly those depicting individuals.

This, thesis explored and analysed the biases within such models by conducting a comparative analysis between the aforementioned models alongside the publicly available LAION-400M training dataset in relation to real-world bias.

The research approach revolved around the retrieval or generation of images coinciding with the biased terms doctor and nurse. These terms were used to leverage real-world biases throughout the bias identification process thereby exposing how each generative model deals with this innate bias and by extension discover any bias mitigation techniques along with their effectiveness in comparison to the other models.

This was achieved by annotating the images using feature extraction models in particular DeepFace and FairFace, whose accuracy was evaluated on a human annotated subset of LAION-400M images. Furthermore, the bias present within the images was concluded due to a series of metrics particularly gender, race and age distribution, person prominence along with Shannon and Simpson diversity/evenness measures. This research highlighted the bias present within the LAION-400M dataset along with the Stable Diffusion and Midjourney models whilst outlining the inverse bias within the Dall-E model and the effectiveness of its bias mitigation process.

The findings of this research shed light on the pervasiveness of bias in generative AI, highlighting the urgent need for proactive mitigation strategies whilst contributing to the understanding of bias and the development of fairer models and datasets.

Acknowledgements

I would like to thank my supervisor Dr Dylan Seychell for guiding me throughout the process of this final year project and aiding me throughout various challenges encountered. I would also like to thank my parents, Reno and Graziella, and my brother Julian for their continuous support.

Contents

Abstract	i
Acknowledgements	ii
Contents	v
List of Figures	vi
List of Tables	vii
List of Abbreviations	viii
Glossary of Symbols	1
1 Introduction	1
1.1 Problem Definition	1
1.2 Motivation	2
1.3 Aims and Objectives	2
1.4 Document Structure	3
2 Background	4
2.1 Prompting	4
2.2 Diffusion Models	4
2.3 Facial Analysis	5
2.4 Image Bias	6
2.5 Chapter Summary	7
3 Literature Review	8
3.1 Prompting	8
3.1.1 Prompt target	8
3.1.2 Prompt structure	8
3.2 Facial Analysis	10
3.3 Image Annotations	11
3.3.1 Human based image annotation	11

3.3.2	Computer assisted image annotation	12
3.4	Measuring Bias (Metrics/Techniques)	13
3.4.1	Reduction to tabular data	13
3.4.2	REVISE	15
3.5	Chapter Summary	16
4	Specification and Design	17
4.1	Dataset and Generative Models	17
4.2	Image Retrieval	18
4.3	Image Annotation	18
4.4	Metric Extraction	20
4.5	Chapter Summary	20
5	Implementation	21
5.1	Human Image Annotation	21
5.2	Computer Assisted Image Annotation	21
5.3	Implemented Measures	25
5.3.1	Dulhanty and Wong measurement	25
5.3.2	Zhao et al. measurements	25
5.3.3	Merler et al. measurements	25
5.3.4	REVISE measurements	26
5.4	Chapter Summary	26
6	Evaluation	27
6.1	Real World Bias	27
6.2	Human and AI annotation comparison	28
6.3	Bias in LAION-400M Dataset	29
6.4	Stable Diffusion result analysis	30
6.5	Dall-E result analysis	30
6.6	Midjourney result analysis	32
6.7	Discussion	32
6.7.1	LAION-400M Dataset	32
6.7.2	Stable Diffusion	33
6.7.3	Dall-E	33
6.7.4	Midjourney	33
6.7.5	Discussion	34
6.8	Chapter Summary	34
7	Conclusion	35
7.1	Future Work	35

7.1.1	Revisiting the Aims and Objectives	35
7.1.2	Critique and Limitations	35
7.1.3	Future Work	36
7.2	Final Remarks	36
A	Human Annotation Data Analysis	44
B	Results	46
B.1	LAION-400M	46
B.2	Stable Diffusion	49
B.3	Dall-E	52
B.4	Midjourney	55

List of Figures

Figure 2.1	Forward diffusion process [20].	5
Figure 2.2	Revere diffusion process [20].	5
Figure 3.1	Midjourney prompt structure [39].	9
Figure 3.2	DeepFace age model architecture [45].	11
Figure 3.3	The distribution of Monk Skin Tone Scale annotations for this image from a sample of 5 photographers in the U.S. and 5 photographers in India in [47].	12
Figure 4.1	FYP Process	17
Figure 4.2	UTK-Face Model Comparison	19
Figure 5.1	Metric Dictionary Structure	22
Figure 6.1	Google Form Human Annotated Images Demographic Graphs	29
Figure A.1	Google Form Respondent Demographic Graphs	44
Figure A.2	Google Form	45
Figure B.1	LAION-400M Demographic Graphs	46
Figure B.2	LAION-400M Prominence Graphs	47
Figure B.3	StableDiffusion Demographic Graphs	49
Figure B.4	Stable Diffusion Prominence Graphs	49
Figure B.5	Dall-E Demographic Graphs	52
Figure B.6	Dall-E Prominence Graphs	52
Figure B.7	Midjourney Demographic Graphs	55
Figure B.8	Midjourney Prominence Graphs	55

List of Tables

Table 5.1 Average Fliess Kappa Values For Doctor/Nurse Responses 21

Table B.1 LAION-400M Shannon & Simpson Measurements 47

Table B.2 LAION-400M Positive Correlation Measurements 48

Table B.3 Stable Diffusion Shannon & Simpson Measurements 50

Table B.4 Stable Diffusion Positive Correlation Measurements 51

Table B.5 Dall-E Shannon & Simpson Measurements 53

Table B.6 Dall-E Positive Correlation Measurements 54

Table B.7 Midjourney Shannon & Simpson Measurements 56

Table B.8 Midjourney Positive Correlation Measurements 57

List of Abbreviations

AAMC Association of American Medical Colleges.

AI Artificial intelligence.

API Application programming interface.

CLIP Contrastive language-image pre-training.

CNN Convolution Neural Network.

FYP Final year project.

MAE Mean absolute error.

OECD Organisation for Economic Cooperation and Development.

RBF Radial Basis Functions.

ResNet Residual neural network.

SVM Support Vector Machine.

UNET U-shaped encoder-decoder network architecture.

URL Uniform resource locator.

WHO World Health Organization.

1 Introduction

1.1 Problem Definition

In recent years, the field of Generative AI has experienced remarkable advancements in visual content generation, with a primary focus on images. Notably, generative models such as Midjourney, DALL-E and Stable Diffusion have been at the forefront of this progress [1–3], by providing users with the capability to generate numerous images through the use of simple text prompts. However, the generation of visual content brings to the forefront a variety of critical issues such as lack of control over output, over fitting as well as privacy and ethical concerns [4, 5].

This study focuses on a particular issue, that of bias. Bias in relation to visual AI systems tends to refer to cases in which systems showcase prejudice in relation to particular demographic features, gender and race being the primary focus of this paper [6]. Several instances exist in which this bias driven prejudice led to negative consequences in relation to recidivism scoring [7], online advertisement [8], facial recognition [9], and credit scoring [10].

Bias serves to affect a large majority of computer vision systems such as classification algorithms, face recognition systems, object detectors and many more [11]. To address this problem tools can be created which aid in the identification of bias, these are crucial as bias is not attributed to a singular cause rather a variety of factors varying from the composition of the dataset and the framing of images to the characteristics of the latent space employed during the generative process [11].

Tools such as this already exist, a prime example is the REVISE implementation which given an annotated dataset can provide object-based, person-based and geography-based insights on the presence of bias [12]. However, such systems tend to be cumbersome to set-up and utilise. The initial aim of this study was to detect if bias is present in traditionally gender biased prompts such as doctor and nurse. This was initially going to be carried out by looking at relevant images from the Stable Diffusion model and the LAION-5B training dataset, however due to recent proceedings [13] with the LAION-5B dataset, its access has been revoked and thus the study will instead attempt to outline the presence of bias within the LAION-400M dataset¹ whilst also considering multiple generative models in particular Stable Diffusion, DALL-E and Midjourney. This study also aims to develop a simple to use python notebook which will facilitate image feature extraction and metric visualisation to allow individuals to easily detect bias and replicate the results shown.

¹ Accessible via a Kaggle repository [14]

1.2 Motivation

The motivation behind this research stems from the growing importance of addressing bias in artificial intelligence (AI) systems, particularly within the realm of generative models and visual datasets. As AI technologies continue to play an increasingly integral role in shaping various aspects of our lives, understanding and mitigating biases becomes imperative. The LAION-400M dataset along with the Midjourney, DALL-E and Stable Diffusion models serve as focal points for this study, representing key components in the landscape of generative AI. By investigating and uncovering biases present in these specific entities, this research aims to contribute valuable insights to the broader discourse on ethical AI development. The implications of biased AI systems are far-reaching, with potential consequences in areas such as image generation, facial recognition, and algorithmic decision-making. Through a meticulous examination of biases, this study strives to enhance our understanding of the challenges inherent in generative models but also to pave the way for more ethical and unbiased AI systems in the future.

1.3 Aims and Objectives

The aim of this study as outlined above is to determine the presence of bias within the generative AIs mentioned prior as well as the LAION-400M dataset. In line with this no final deliverable or program will result from this research paper excluding the program containing the feature extraction and analysis pipeline leading to the insights and conclusion presented in this thesis. This aim will be achieved via the following set of objectives:

1. Investigate how each generative model processes their prompts and determine an optimal prompt structure. Determine the requirements needed to carry out valid human annotation.
2. Generate images of doctors and nurses using the Stable Diffusion, Dall-E and Midjourney models and retrieve the associated images from the LAION-400M dataset. Annotate the retrieved images using feature extraction models in relation to gender, race and age, similarly human annotate a subset of the LAION-400M dataset.
3. Determine the annotation bias within the DeepFace and FairFace models via comparison with the human annotated LAION-400M subset. Furthermore, extract the generated image metrics consisting of gender, race and age

distributions, person prominence and Shannon/Simpson diversity and evenness measures.

4. Through qualitative analysis regarding the resultant metrics, uncover relationships within the data to identify the innate bias within the LAION-400M training dataset and the aforementioned models, whilst concluding on the common ways by which bias presents itself, the least biased model and the effectiveness of any implemented bias mitigation techniques.

1.4 Document Structure

This theses is divided into seven sections, with this section providing an introduction to the field in which the research resides whilst outlining the primary aims and objectives. The background section provides additional information, technical or otherwise which is required to fully understand the means by which the research was carried out. The literature review section delves into complementary research covering prompting and its structure, facial analysis models, bias types and measurement techniques alongside the two image annotation techniques used throughout this paper. The specification and design section provides a detailed explanation and justification on the decisions made throughout the research pipeline. The implementation section delves into the program used to facilitate image annotation and metric extraction in addition to outlining the human annotation results and the agreeableness therein. The evaluation section presents an in depth look at the metric results obtained comparing them to real world data and arriving to a conclusion on the dataset and model bias. The conclusion section revisits the aims and objectives highlighting how they were achieved, whilst critiquing and outlining the limitations of this study in addition to presenting areas in which the research conducted could be further expanded upon. Furthermore, it summarises the above sections, whilst reiterating the findings and conclusion of this theses.

2 Background

This chapter provides a foundation of knowledge required for understanding the techniques employed within the bias detection pipeline. The chapter is divided into four subsections covering prompting, diffusion models, facial analysis, and image bias, going over a variety of relevant research and challenges associated with each section. Furthermore, the chapter outlines how each section fits into this research paper.

2.1 Prompting

Prompting consists of guiding generative models towards generating appropriate text, code, images and other outputs. The guiding instructions can involve various mediums primarily text, code and images. Given that this research paper concerns itself with text-to-image generation and the bias therein only text inputs and image outputs were considered.

Along the same vein, prompting introduces a variety of challenges revolving around the generation of relevant images. These challenges are closely related to identifying a suitable prompt to achieve the required output. This is a non-trivial issues as slight alterations to the prompt can have a major impact on model performance and output, as such finding the appropriate prompt is a time consuming endeavor [15]. Prompt engineering addresses this by altering prompt length and wording to effectively depict the required output, rather than just specifying the desired image [16]. Automated prompt engineering can further enhance this process, although it was beyond the scope of this paper.

2.2 Diffusion Models

Generative models encompass a variety of different approaches, including GANs, VAEs, and diffusion models. The latter offers several advantages over its counterparts. Unlike GANs, diffusion models excel in both training stability and diverse image generation, avoiding the pitfalls that often plague GANs. Additionally, they bypass the surrogate loss issue inherent in VAEs. This allows diffusion models to achieve superior performance and efficiency. The models considered in this paper all fall under the diffusion category [17–19].

Diffusion models are traditionally composed of two steps, these being the forward and reverse diffusion processes. Forward diffusion adds Gaussian noise to an image until the resultant image no longer resembles the input as can be seen in Figure 2.1.

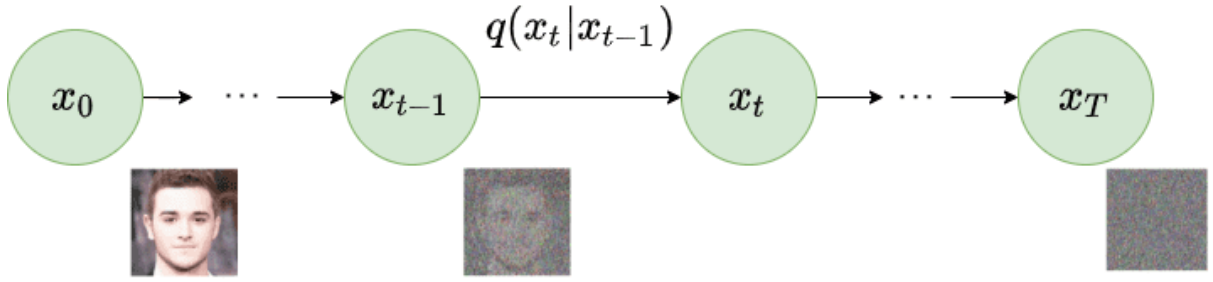


Figure 2.1 Forward diffusion process [20].

The reverse diffusion process resembles the inverse of the forward diffusion process as depicted in Figure 2.2, which employs a noise prediction model to iteratively denoise the input image. The noise predictor iteratively estimates and subtracts noise from the image's latent space thereby enhancing image details. Contrary to pixel space, latent space serves as a compressed representation of the image. Its use throughout the diffusion process offers significant computational advantages including vastly reduced processing demands, enhanced performance, and improved overall efficiency [18, 21]. Furthermore, conditioning prompts serve to guide the diffusion process towards specific image themes or styles [21]. The UNET architecture originally developed for image segmentation in bio-medicine serves as the core component within this process, with the majority of generative models adopting the ResNet variant developed for computer vision, throughout their image generation pipeline.

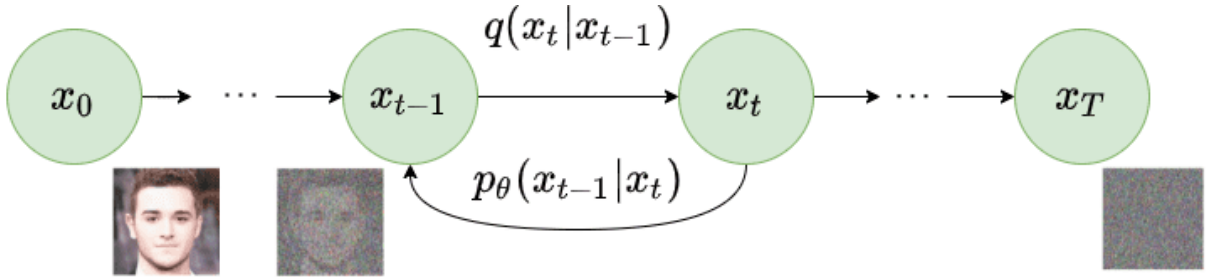


Figure 2.2 Reverse diffusion process [20].

2.3 Facial Analysis

Facial analysis involves a three step process consisting of face detection, feature extraction and facial analysis. Face detection involves the extraction of face regions which are used for face tracking and pose estimation. These serve as input to the feature extraction process which retrieves several facial features varying based on the extraction model used. These generally consist of colour-based, spatial, textural, geometric, and deep learning features with the type extracted varying based on the

use case. These are then fed to the facial analysis component which uses said data to extract faces, age, gender, race, head pose and so on [22]. The outputs present unique challenges due to their use of different facial features and extraction processes.

However, seeing as they are interconnected, advancements in one area can benefit another. This feature diversity leads to various use cases for facial analysis, including but not limited to using facial analysis to direct the attention of a surveillance system based on what is capturing peoples attention [23], discerning an advertisements level of engagement based on an individuals attention [24], ensuring driving safety by monitoring the emotional state of the driver [25], and estimating attributes such as expression, gender, age, and race to aid in tasks like image annotation. However, facial analysis encounters several challenges, these include; pose variation, the obstruction of facial features, varying facial expressions and image quality (lighting, image size) which can all negatively affect the resultant output [26–28].

2.4 Image Bias

Image bias in relation to visual AI systems as defined in Section 1.1 tends to primarily refer to cases in which systems showcase prejudice in relation to certain demographic features [6]. However, bias can present itself in a variety of different forms, these can be broadly categorised as; selection bias which occurs when visual data is unevenly gathered, leading to inaccurate and biased representations, framing bias which arises from how images are composed, influencing perception through angles, lighting, and expressions, potentially leading to unfair interpretations and label bias resulting from inaccurately tagged images, which distort data meaning and hinder accurate analysis [29].

These types of biases in most cases are not intentional rather they occur due to some unforeseen consequences of the data collection and annotation process. Thus, it is crucial to identify and mitigate such bias. Bias detection techniques can be categorised as either subjective or objective. The latter using statistical and algorithmic approaches whereas the former utilises human judgment to come to a conclusion based on the resultant data. These approaches usually go hand in hand as can be seen in [12] wherein the tool itself utilises various algorithmic techniques to extract various metrics, in turn allowing an individual to carry out the final judgement on bias. This joint approach is useful as the individual can contextualise the presented metrics and thus, come to a sound conclusion.

Bias mitigation techniques vary in their implementation however there are certain aspects one must keep in mind in order to mitigate bias, these include but are not limited to [11]:

- Selection bias
 - Data representativeness - do we need balanced or statistically representative data?
 - Negative set coverage - are the negative sets adequately represented?
 - Excluded groups - are there any essential categories which are missing?
- Framing bias
 - Image interpretation - can the interpretation of an image change based on the viewer?
 - Subject depiction - are certain subjects depicted in a particular manner more than others?
 - Stereotype adherence - does data perpetuate harmful biases?
- Label bias
 - Automated labelling biases - has the innate machine bias been taken into consideration or mitigated?
 - Annotator bias control - is there a diverse team of annotators such that human bias is mitigated?
 - Label clarity - are fuzzy labels (gender/race) being used ?

2.5 Chapter Summary

This chapter introduces the key concepts and techniques required to understand the content of this paper as well as its importance. It covers prompting, diffusion models, facial analysis and image bias, explaining their purpose and relevance to the research.

3 Literature Review

This chapter offers a comprehensive review of relevant studies, beginning with an overview of prompting and its structure whilst outlining the reason behind the prompt targets chosen. It delves into facial analysis models, addressing associated issues, and discusses the two types of image annotation methods along with their respective advantages and disadvantages. The chapter concludes by examining bias measurements employed in analogous scenarios, laying the groundwork for the research proposed in this paper.

3.1 Prompting

In accordance with section 2.1, prompting is the process by which a person guides an artificial intelligence model to generate a specific output. This section explores the main challenges associated with using said models in particular the process of selecting the prompt target and the formulation of the prompt structure.

3.1.1 Prompt target

In accordance with the nature of this research, the prompt subjects will consist of traditionally gender biased professions. Specifically, doctor and nurse as they are male and female dominated respectively as showcased in [30, 31]. Furthermore, said bias perpetuates itself online as seen in [32] thereby increasing its relevance to this research as the majority of these generative AIs leverage training data retrieved from the Internet. For instance, Stable Diffusion was trained on the LAION-5B dataset, a successor to LAION-400M, with images sourced from the common crawl [33, 34].

3.1.2 Prompt structure

Stable Diffusion

The Stable Diffusion WebUI repository [35] outlines the tools functionality covering upscaling, img2img, negative prompting, face restoration, model merging and so on. It further specifies that Stable Diffusion accepts prompts of up to 75 tokens, with additional 75 token chunks allotted in instances of longer prompts. However, there is an overall limit that, if surpassed, triggers a warning and prompt truncation. Despite this, the FYP results are unaffected as the prompts used were relatively short as outlined below. Additionally, the prompt should specify the subject, image medium (digital art, sketch, painting), image style (hyper realistic, fantasy) whilst utilising

negative prompts in accordance with the stable diffusion prompt guide [36] to achieve the best results.

DALL-E 3

The official DALL-E 3 documentation [37] outlines how the DALL-E 3 input prompt is automatically rewritten for safety reasons and enhanced detail. Furthermore, this functionality currently cannot be removed as such it is recommended to precede the prompt with *I NEED to test how the tool works with extremely simple prompts. DO NOT add any detail, just use it AS-IS:* to produce images closer to the initial prompt. Lastly, the OpenAI Developer Forum [38] offers general prompting tips such as be specific and detailed, use descriptive adjectives, avoid prompt overloading, specify desired styles or themes.

Midjourney

The official Midjourney documentation [39] outlines how the Midjourney bot breaks down the prompt into smaller chunks called tokens, which are then compared with its training data to generate prompt relevant images. It suggests using simple, short sentences as opposed to a long list of requests and instructions to obtain the best results. Furthermore, it outlines the creation of advanced prompts composed of image prompts, text prompts and parameters. Image prompts consist of an image URL which influences the style and content of the generated image. Text prompts consist of a text description of what image you want to generate. Parameters alter the resultant image by changing its aspect ratios, upscaling and so on. This prompt structure can be seen in Figure 3.1.

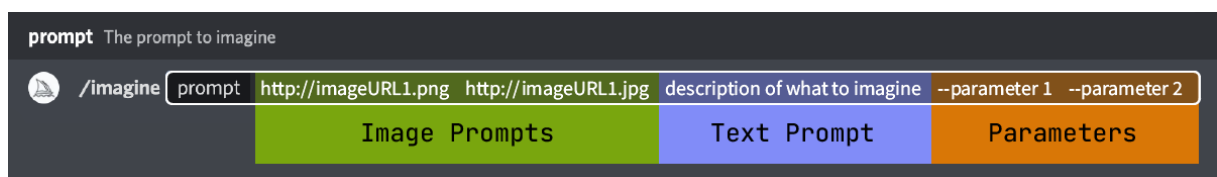


Figure 3.1 Midjourney prompt structure [39].

The guide further emphasises the significance of word choice, advocating for the use of synonyms, numbers, and collective nouns. It underscores the importance of directing prompts towards desired image elements rather than exclusions, suggesting the use of the `-no` parameter for the latter. Additionally, it highlights the impact of prompt length, noting that shorter prompts encourage model creativity, while longer, detailed prompts provide greater control.

According to the literature reviewed, it is evident that while the models vary, their prompt features generally remain consistent. Thus, a general prompt structure for image generation can be proposed: *A picture of a [subject] facing forward*, with *Disfigured* and *Art* as negative prompts for clear, realistic depictions. This streamlined prompt integrates various suggestions, it specifies *picture* to emphasize the medium and realistic nature of the required image, *subject* is used to focus on the desired aspects, *facing forward* is used to ensure facial clarity, whilst the short length of the prompt allows the model creativity limiting human influence and enhancing clarity with Negative Prompts.

3.2 Facial Analysis

Facial analysis, as outlined in section 2.3 refers to the extraction of varied facial features, each with their own challenges and issues. Focusing on the features relevant to this paper several techniques exist by which they can be retrieved ranging from Support Vector Machines (SVM), Radial Basis Functions (RBF) and Deep Learning based methods, with the latter being the most commonly used.

Deep Learning based methods involve the training of a Convolution Neural Network (CNN) using a vast and expansive labelled dataset, thereby allowing for gender, age, race and emotion estimation and classification. Instances of these models include Googles Google Vision API and Amazons Rekognition API, the latter implementing only gender and emotion classification whilst the former only implementing emotion classification. These APIs implement other functionalities such as object detection, text detection and so on, however they minimise their classification functionalities to just gender and emotion as the implementation of such models requires access to large unbiased datasets to produce accurate results. Additionally, given the size and influence of these companies they have to take into consideration the possible affects that releasing such models can have on society which can be quite problematic as can be seen with Meta's discontinuation of its face recognition system in the wake of sustained privacy and ethical concerns such as the abuse of marginalised groups and further racial bias [40].

Contrarily, the open-source DeepFace API implements age, gender, race and emotion estimation and classification with varying degrees of success. The age and gender models were implemented using the VGG-Face model in which the initial layers were frozen whilst the remainder were trained on a subset of the IMDB+Wikipedia dataset, the race model underwent similar training on the FairFace dataset. The implementation of the emotion model required a custom architecture depicted in Figure 3.2. and was trained on the FER-2013 dataset. These models achieved varying

degrees of accuracy on their respective test sets, with the gender classification model having an accuracy of 97.44%, the race classification model had an accuracy of 68% with the emotion model having a 57.42% accuracy. Finally, the age model achieved an MAE of 4.65 meaning that the age can be predicted with plus and minus 4.65 years [41–43]. Similar to DeepFace the FairFace model implemented the same functionalities save for emotion classification, with a gender classification accuracy of 94.89% and 92.95% across images of people varied in race and age respectively. Age and race classification was also noted to be on par or surpassing other commercial APIs [44].

	1 conv	2 mpool	3 conv	4 conv	5 apool	6 conv
Filters	64	-	64	64	-	128
Kernel	5	-	3	3	-	3
Pool	-	5	-	-	3	-
Strides	-	2	-	-	2	-
Units	-	-	-	-	-	-
	7 conv	8 apool	9 fc	10 fc	11 fc	12 softmax
Filters	128	-	-	-	-	1
Kernel	3	-	-	-	-	-
Pool	-	3	-	-	-	-
Strides	-	2	-	-	-	1
Units	-	-	1024	1024	7	0

Figure 3.2 DeepFace age model architecture [45].

3.3 Image Annotations

Image annotation is the process by which labels are assigned to an image or image set. This is a crucial component of any study particularly those revolving around bias as the metrics used to deduce a conclusion need a basis on which to be made. Image annotation is commonly carried out in either of two ways, these being computer assisted and human based image annotation.

3.3.1 Human based image annotation

Human based image annotation makes use of human annotators to correctly identify and label images. Although computer assisted image annotation has become more prevalent, there is still a place for human based image annotation in various applications such as computer vision and machine learning.

However, this process has both strengths and challenges. Human annotators excel at understanding complex visual information and incorporating context into annotations, as evidenced by research showing their ability to recognize positive and negative expressions with minimal facial expression information [46]. They can adapt to varying conditions, as seen in instances where annotators adjusted to differences in image hue, saturation, and brightness for skin tone annotation [47]. Moreover, humans demonstrate high accuracy in specific annotation tasks, such as face and gender

annotation, achieving a 96% accuracy excluding contextual cues like hairstyle and makeup [48].

On the other hand, challenges include annotator bias, where cultural and experiential differences influence image annotations as is depicted in Figure 3.3 [47], consistency issues due to subjective interpretations of ambiguous visual cues, impacting data reliability, and scalability and cost concerns, as manual annotation is time-consuming and expensive, particularly for large datasets, leading to the adoption of computer-assisted image annotation [49].

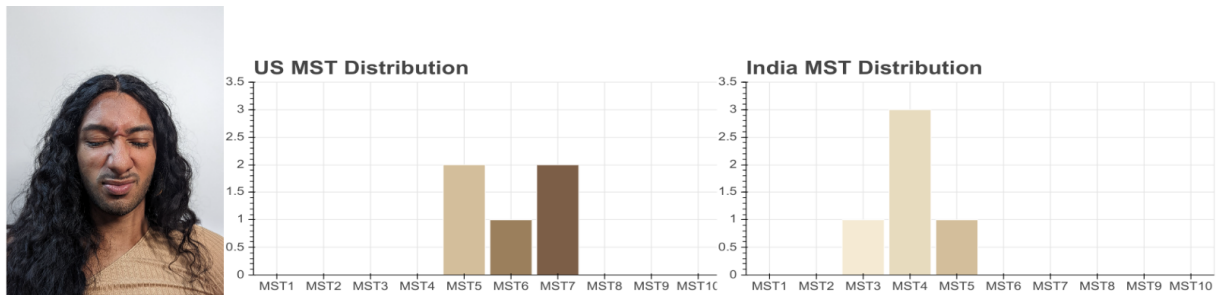


Figure 3.3 The distribution of Monk Skin Tone Scale annotations for this image from a sample of 5 photographers in the U.S. and 5 photographers in India in [47].

Furthermore, considerations for effective human-based image annotation include selecting annotators from geographically diverse backgrounds to ensure accurate annotations, as advocated in [47], implementing a standard set of labels and measures throughout the annotation process to maintain a cohesive standard, and integrating annotation tools like Roboflow or similar to reduce required annotation time, as recommended by [49].

3.3.2 Computer assisted image annotation

Computer assisted image annotation makes use of AI models such as those discussed in Section 3.2 to remove the human component from the annotation process. Similar to human based image annotation this comes with a varying degree of strengths and challenges as outlined below.

The primary strengths of computer-assisted image annotation include scalability and cost-effectiveness, as it can handle large volumes of images without additional costs, ensuring consistency across annotations when using consistent model architecture and pipelines, and efficiency in quickly annotating large datasets, saving time and resources [50].

Contrarily, challenges and limitations include varied model performance depending on the training dataset used, the replication of innate training dataset bias, and limitations in understanding nuances and context specific to task at hand, leading

to misinterpretations and errors, especially in complex or ambiguous scenarios [51].

Furthermore, considerations for effective computer-assisted image annotation involve assessing the quality of the training data to determine a model's applicability to a particular task [52].

3.4 Measuring Bias (Metrics/Techniques)

Bias can present itself in a myriad of ways as outlined in Section 2.4 such as selection, framing and label bias, in line with the concerns of this research paper selection and framing bias are the main types considered. The study of these biases can prove useful in exposing the presence of baser biases in particular gender, race and age bias.

Considering the research carried out by Fabbri et al. in [11] a total of twenty four papers and the strategies they used to discover bias were reviewed, these strategies were then grouped into four categories, those that measure bias using already present or extracted attributes and labels as if it were a tabular dataset. These encompass the majority of techniques used throughout this paper then there are those techniques which discover bias through observing lower-dimensional representations of the data and those which uncover bias via cross-dataset comparisons. The remaining techniques discussed which did not fall under these categories were categorised as other.

3.4.1 Reduction to tabular data

Most strategies presented in this section utilise automatic feature extraction processes which are prone to errors and bias, this can in turn result in said bias being reflected or amplified in the final output. It is also noted that the impact of this additional source of bias is typically ignored, only being mentioned as an aside when interpreting the results. This section covers those strategies which are relevant to this research.

Count / demographic parity.

Dulhanty and Wong in [53] determined the presence of gender and age bias in the ImageNet dataset by extracting the age and gender of images using relevant recognition models and thereby determine the distribution of age and gender across the dataset. This method provides insight into selection bias but also on the framing of the protected attributes, given a suitable labelling of the dataset. Finally, it was noted by the authors that such a method relies on the assumption that the recognition models involved are not biased, which is a far claim from the truth and thus, the analysis is not fully reliable.

Yang et al. [54] similarly opted to address selection and label bias in relation to the person category of the ImageNet dataset. To address label bias, they firstly had annotators remove images which could be offensive or sensitive (e.g., sexual/racial slur) and those with ambiguous labels, this was then followed by the annotators labelling the remaining images according to the categories of interest (gender, age, and skin colour). This was done so as to understand the bias present within the dataset. The annotation process was validated by measuring the agreeableness of the annotators on a small, controlled set of images.

Zhao et al. [55] measured the correlation between protected attributes and the occurrences of certain objects/actions. This was carried out via equation 3.1 wherein g_n was a protected attribute and o an occurrence of an object or action in the image. The bias score is denoted by $b(o, g)$ whereas $c(o, x)$ counts the co-occurrences of the object/action o and the protected attribute's value x . Assuming that $b(o, g_i) > \frac{1}{n}$, where n is the number of possible values that the protected attribute can be, this implies that the attribute g is positively related to object/action o .

$$b(o, g) = \frac{c(o, g)}{\sum_{x \in \{g_1 \dots g_n\}} c(o, x)} \quad (3.1)$$

Information theoretical

Merler et al. [56] presented four measurements for a balanced dataset. The Shannon entropy and Simpson Index as calculated in equations 3.2 and 3.3 measure diversity with the larger values depicting greater diversity. Additionally the Shannon and Simpson evenness measures as calculated in equations 3.4 and 3.5 denote how evenly distributed the dataset labels are across the entire dataset with the maximum value being 1.

$$H(X) = - \sum_{i=1}^n P(X = x_i) \cdot \ln(P(X = x_i)) \quad (3.2)$$

$$D(X) = \frac{1}{\sum_{i=1}^n P(X = x_i)^2} \quad (3.3)$$

$$E_{Shannon} = \frac{H(X)}{\ln(n)} \quad (3.4)$$

$$E_{Simpson} = \frac{D(X)}{n} \quad (3.5)$$

3.4.2 REVISE

The revise tool [12] adopts a multi-variant approach to detecting bias considering object, person, and geography-based insights. This section will go over person-based insights as the object and geography-based insights are irrelevant to this research paper. The relevant metrics used for detecting person-based bias are outlined below:

Person Prominence

This considers the proportion of the image that the subject takes up in addition to the distance of the subject from the centre of the image. These measures are then treated as a proxy for the subjects importance. This analysis was carried out on the COCO dataset for images separated by gender and skin tone, for which the Cohen's D measurements was used to facilitate a comparison between the different groups whilst Jonckheere's trend test was used to visualise an a priori ordering of the data.

Contextual Representation

This considers the context in which individuals are primarily featured in through the objects and scenes with which they are primarily associated with. Taking into consideration the COCO dataset it was concluded that woman tend to be greatly associated with shopping and dining whilst being depicted in images containing furniture, accessories, and appliances. Contrarily men tend to be associated with sports fields and water, ice, snow, whilst depicted in images mostly containing sport items and vehicles. This reflects traditional gender stereotypes present in society.

Instance counts and distances

This opts to look deeper than simply the number of times certain object appear with individuals rather it considers the distance said object is from the subject to determine if the subject is interacting with the object or whether it is simply in the background. This is achieved via a scaled distance metric depicted in equation 3.6 wherein p denotes the person, o the object and the $area_p/area_o$ are calculated on a normalised image of total area equal to 1. This metric in turn outlines whether certain demographics are depicted interacting with certain objects as opposed to the image simply containing the two.

$$dist = \frac{\text{distance between } p \text{ and } o \text{ centres}}{\sqrt{area_p * area_o}} \quad (3.6)$$

Appearance differences

This opts to analyse appearance differences in images of people of varying demographics in relation to particular objects. This was carried out to further disambiguate situations where occurrence counts, and distance are not depicting the entire situation. This involves extracting FC7 features from a subset of images to get scene-level features, projecting them into $\sqrt{\text{number of samples}}$ dimensions to prevent over-fitting and then fitting a Linear SVM to see if it learns the difference between images containing n images of the same object with people of different demographics. This results in insights such as men being portrayed playing outdoor sports whilst woman playing indoor sports when considering the object *sports uniform*.

3.5 Chapter Summary

This chapter delved into the research and techniques crucial for the execution of this research. Elaborating on prompting, covering its structure and target selection. It explored various facial analysis models, highlighting their inherent challenges. Subsequently, the chapter dove into the two primary image annotation methods, outlining their strengths and limitations. Finally, concluding with an exploration of various bias measurement techniques employed in similar research areas, laying the groundwork for the construction and execution of the research proposed in this paper.

4 Specification and Design

The specification and design chapter provides a detailed explanation of the steps taken to arrive at the conclusion presented in chapter 6, going over its various components whilst detailing and justifying the decisions taken throughout. It further links the research carried out in chapter 3 with the implementation whilst offering a critical review of the entire process, going over the challenges encountered and their solutions. The process is outlined in Figure 4.1.

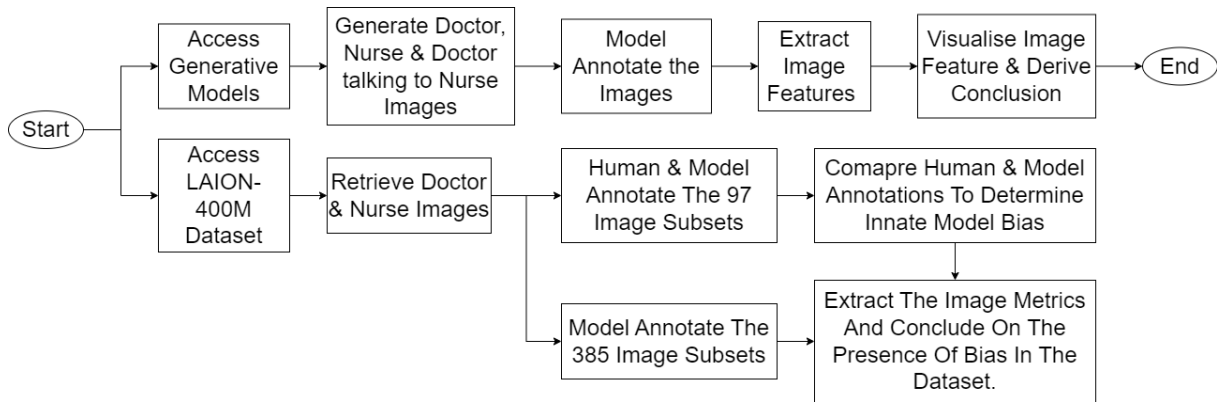


Figure 4.1 FYP Process

4.1 Dataset and Generative Models

In line with the aim of this research paper the initial decision was taken to perform analysis on the Stable Diffusion model alongside the LAION-5B training dataset, this was done as both were freely accessible and the model widely used. However, when it came to facilitate image retrieval from the LAION-5B dataset, said dataset had been taken down due to the alleged presence of illegal content [13]. Although steps were taken in order to facilitate access to the dataset the issue persisted. Fortunately, Kaggle still hosted access to some of LAIONs datasets in particular LAION-400M [14]. Given that said dataset preceded the LAION-5B and was curated in the same manner as the LAION-5B dataset, all be it on a smaller scale it served as a valid substitute. Thus, the decision was taken to utilise said dataset instead, however this resulted in further issues particularly regarding image retrieval. Due to the change in dataset it was deemed appropriate to expand the scope of the research by considering other popular generative models such as Dall-E and Midjourney. These models were chosen due to their popularity and usage, they also served as a point of comparison with the Stable Diffusion model even though said models have distinct training datasets.

4.2 Image Retrieval

The initial challenge in this section was deducing the image subjects, i.e., what the images will depict. In line with Section 3.1.1 *Doctor* and *Nurse* were chosen due to the innate bias associated with the professions. Additionally, image retrieval proved challenging both in accessing the relevant LAION-400M images as well as generating said images via the aforementioned models.

The main issue with retrieving images from the LAION-400M dataset revolved around the lack of CLIP integration which had previously been used to carry out image retrieval on the LAION-5B dataset. CLIP was useful as it streamlined the image retrieval process through its search functionality allowing for efficient image retrieval. This lack of CLIP integration for the LAION-400M dataset was resolved by parsing the text descriptor assigned to each image and removing those images which did not contain the word *Doctor* or *Nurse*. Furthermore each image had an NSFW label which allowed for the removal of unsavoury images, however filtering by hand was still carried out to clear out any images which had been incorrectly labelled as well as those depicting children and multiple individuals. The latter was carried out using the YOLOv8 model to flag images depicting multiple people, with the purpose being to remove ambiguity in relation to the subject of the image.

Image generation was carried out using the prompt outlined in Section 3.1.2, where subject was substituted for *doctor*, *nurse* and *doctor and a nurse*. The latter label was added simply to deduce if the presence of both labels in a singular prompt would affect the image depiction. The Stable Diffusion model was accessed and used via [18] whilst utilising the interface provided in [35]. Furthermore, DALL-E was accessed through OpenAIs API whereas MidJourney provided no official API integration and as such the images were generated via discord as is outlined on their official page.

Once the means by which image retrieval was established another issue presented itself, this being the number of images to consider. Given the time and monetary constraints it was determined that 385 was a sufficient number of images as it guaranteed a confidence score of 95% with a 5% margin of error. In the case of the LAION-400M dataset these images were randomly selected from the doctor and nurse subsets resulting in 2 subsets respectively. Furthermore, subsets were selected from the generative images resulting in 9 subsets total 3 per model.

4.3 Image Annotation

Following image retrieval the next step was image annotation. This was divided into two parts human and computer assisted image annotation. The former was carried out

on a subset of the selected LAION-400M images, said subset consisted of 97 images per label resulting in a confidence score of 95% with a 10% margin of error. The increase in margin of error was deemed acceptable given the reduction in image count from 385 to 97 whilst allowing for human annotation within a reasonable time frame. The human annotation process was carried out via the use of several google forms as depicted in Appendix A wherein users had to label images in terms of their gender, race and age. Finally, only the image subsets from the LAION-400M dataset were human annotated as carrying out human annotation on all generated images would conflict with our time constraint and this subset was sufficient in gauging any pre-existing bias within the DeepFace and FairFace models.

The computer assisted annotation process was carried out on the LAION-400M images as well as all the generated images as it was the only feasible way to annotate such a large set. This process initially made use of the DeepFace model as outlined in Section 3.2 as it provided adequately accurate gender, age and race classification. The emotion classification was initially considered however it was deemed redundant as no possible relevant insights were identified. Furthermore the FairFace model was also implemented as it provided the same classification capabilities as DeepFace whilst performing better when tested on subsets of the UTK-Face dataset as can be seen in Figure 4.2. The UTK-Face dataset [57] was used as it possessed varied images in terms of the age, gender and race whilst facilitating easy comparison between the models. Given these results the initial idea was to combine these two methods through ensemble techniques in the hope of creating a model with reduced bias and better performance however the ensemble model performed worse than the FairFace model and only marginally better than the DeepFace model as such the decision was taken to use the two models individually and interpret their results accordingly.

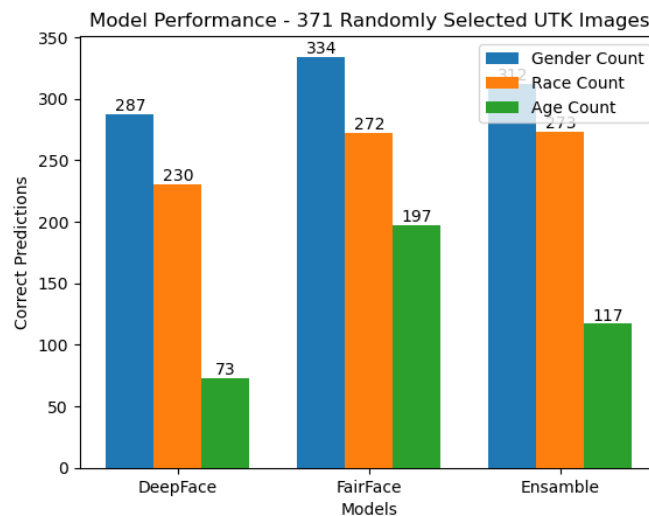


Figure 4.2 UTK-Face Model Comparison

4.4 Metric Extraction

The final component of the pipeline involves the extraction of the image metrics such that a conclusion on the bias present can be made. Several metrics previously outlined in Section 3.4 were used, these include count, person prominence, Shannon entropy, Simpson index, Shannon evenness, Simpson evenness and correlation. These metrics were chosen to systematically determine the form of bias present within the image as well as outline its severity. This is accomplished as follows:

1. The count reveals potential bias, such as an overabundance of male-labeled images, indicating possible gender bias in the generative AI.
2. Correlation confirms and quantifies bias severity. For example, a high correlation between "male" and "doctor" suggests bias, with values above 0.5 indicating less severe bias (e.g., 0.56) and values like 0.89 suggesting more pronounced bias.
3. Shannon entropy/Simpson index measure diversity, while evenness measures indicate how evenly distributed the labels are in the dataset.
4. Person prominence assesses whether identified biases also manifest in image framing.

4.5 Chapter Summary

This chapter delved into the design decisions taken throughout the research paper, going over how the dataset and models were chosen, the process behind image retrieval and annotation and concluding with the metrics used to arrive to the conclusion. The upcoming section will delve deeper and explain how these were fully implemented and carried out from a technical point of view.

5 Implementation

This chapter delves into the implemented program used for image annotation and metric extraction, going over the primary steps in detail whilst outlining any issues presented throughout the implementation process.

5.1 Human Image Annotation

The LAION-400M images distributed via several google forms where annotated by a diverse group of individuals as outlined in appendix A. The responses were primarily gathered via distribution of the forms via social media and then processed simply by carrying out a majority vote for each response received thus, arriving at a singular label per attribute, this majority vote approach was adopted as alternate approaches were beyond the scope of this research paper. Furthermore a weighted approach was initially going to be used however seeing as there is no clear hierarchy to the response given (i.e., no response is worth more than another seeing as none of the respondents are experts) all responses would have equal weight. In addition to processing these labels the Fleiss Kappa [58] scale was used to identify the level of agreement between the respondents, this outlined how there was almost perfect agreement amongst the gender annotation with regards to both doctor and nurse images, there was moderate agreement with regards to racial labels but only slight to fair agreement with regards to age labels (age ranges were used as opposed to the exact age) as denoted in table 5.1. These results are crucial as they support the use of the majority vote approach seeing as agreeableness between annotators was high except for age labels, furthermore it supports the usage for these labels to serve as baseline for model comparison as carried out in Section 6.2.

Fleiss Kappa Avg	Gender	Race	Age
Doctor Values	0.941	0.505	0.217
Nurse Values	0.857	0.589	0.15

Table 5.1 Average Fleiss Kappa Values For Doctor/Nurse Responses

5.2 Computer Assisted Image Annotation

The annotation process starts by loading the images from disk through the LoadImagesFromFolder function. Once the images are loaded they are fed to the YOLODetectPerson function alongside a confidence threshold which utilises the

YOLOv8 model to detect individuals within the images. Predictions below the specified threshold are filtered out. Furthermore the remaining predictions are processed such that the detected individuals are cropped out and their percentage area and distance from image center is calculated. These metrics are crucial for determining person prominence. Finally the cropped images alongside the aforementioned metrics are returned. The pseudo-code for this function can be seen in Algorithm 1.

Once the set of cropped images are extracted they are then fed to either of the two models via the FairFaceProcess or DeepFaceProcess functions. These functions are identical to each other except for the manner in which the model predictions are processed. These models, given a list of images extract the faces of the individuals depicted via the mtcnn model available through the DeepFace library. This model was primarily chosen as it produced the most accurate results when tested with the DeepFace model and as such was applied to the FairFace pipeline for the sake of consistent results. Furthermore the face predictions are passed through the FindObjWithLargestArea function which returns the index of the largest object detected, with the purpose being to determine the primary face (taking up the most area) in an image in cases where there are individuals in the background. Following this the face is then cropped from the image and passed to the FairFace or DeepFace models respectively wherein they output the gender, race and age predictions. These are then stored in a dictionary wherein each image entry consists of the cropped face image alongside the predictions made. The pseudo-code for the FairFaceProcess function can be seen in Algorithm 2, with the DeepFaceProcess function following a similar structure.

Finally the resultant image features from the models alongside the results from the YOLOv8 process are combined into a singular dictionary for ease of use and then saved to disk. The structure of the dictionary is outlined in Figure 5.1. These results can then be loaded for later processing via the LoadMetricCSVFromFolder function.



Figure 5.1 Metric Dictionary Structure

Algorithm 1 YOLODetectPerson Pseudocode

```

1: function YOLODetectPerson(images, confidence_threshold  $\leftarrow$  0.5)
2:   cropped_images  $\leftarrow$  []
3:   dist_between_centers  $\leftarrow$  []
4:   space_takeup  $\leftarrow$  []
5:
6:   #Loading the YOLO model.
7:   model = YOLO('yolov8n.pt')
8:
9:   for input_image in images do
10:    predictions  $\leftarrow$  model.predict(input_image, classes  $\leftarrow$  0)
11:
12:    #Filtering out predictions below the confidence threshold.
13:    predictions  $\leftarrow$  torch.where(scores > confidence_threshold)
14:
15:    #Finding the area of the image and its center.
16:    area_of_image  $\leftarrow$  input_image.width * input_image.height
17:    image_center  $\leftarrow$  work out image center
18:
19:    for person_bounding_box in predictions do
20:      #Crop the image based on the bounding box
21:      cropped_image  $\leftarrow$  input_image[image boundaries]
22:      cropped_images.append(cropped_image)
23:
24:      area_of_cropped_image = cropped_image.width * cropped_image.height
25:      space_takeup.append(area_of_cropped_image/area_of_image * 100)
26:
27:      #Determining the center of the bounding box
28:      bounding_box_center  $\leftarrow$  work out cropped image center
29:
30:      distance_between_centers  $\leftarrow$  math.dist(bounding_box_center, im-
age_center)
31:      dist_between_centers.append(distance_between_centers)
32:    end for
33:  end for
34:  return cropped_images, dist_between_centers, space_takeup
35: end function

```

Algorithm 2 FairFaceProcess Pseudocode

```

1: function FairFaceProcess(listOfImages, fairFaceModel)
2:   FairFaceDict  $\leftarrow$  {}
3:
4:   for image_index, image in enumerate(listOfImages) do
5:     gender_mapping  $\leftarrow$  {"man":0,"woman":1}
6:     race_mapping  $\leftarrow$  {"white":0,"black":1, ... ,"middle eastern":5}
7:     age_mapping  $\leftarrow$  {"0-2":0,"3-9":1,"10-19":2, ... ,"60-69":7,"70+":8}
8:
9:     #Using the mtcnn detector to detect faces in the image.
10:    face_image  $\leftarrow$  DeepFace.analyze(image,detector_backend="mtcnn")
11:
12:    outputs  $\leftarrow$  fairFaceModel(face_image)
13:
14:    #Setting the appropriate scores for each category
15:    for gender in list(gender_mapping.keys()) do
16:      gender_mapping[gender]  $\leftarrow$  gender_confidence
17:    end for
18:
19:    for race in list(race_mapping.keys()) do
20:      race_mapping[race]  $\leftarrow$  race_confidence
21:    end for
22:
23:    for age_range in list(age_mapping.keys()) do
24:      age_mapping[age_range]  $\leftarrow$  age_confidence
25:    end for
26:
27:    #Determining the label with the highest confidence value
28:    gender_label  $\leftarrow$  GetLabelHighestConfidence(gender_mapping)
29:    race_label  $\leftarrow$  GetLabelHighestConfidence(race_mapping)
30:    age_label  $\leftarrow$  GetLabelHighestConfidence(age_mapping)
31:
32:    ImageInfo  $\leftarrow$  {"age":age_label, "gender":gender_label, "race":race_label}
33:    imageConfidenceData  $\leftarrow$  {"race":race_mapping, "gender":gender_mapping,
34: "age":age_mapping}
35:    FairFaceDict[image_index]  $\leftarrow$  [face_image, ImageInfo, imageConfidenceData]
36:  end for
37:  return FairFaceDict
38: end function

```

5.3 Implemented Measures

5.3.1 Dulhanty and Wong measurement

The count was implemented as outlined in [53] to provide an easily interpretable measure of the innate dataset bias, however due to the unreliable nature of deriving a conclusion from the count metric alone it was further supported by the measures outlined below. This metric was implemented via the use of a simple tally which tracked the occurrence of each label within the image subsets.

5.3.2 Zhao et al. measurements

The correlation calculation outlined in [55] was implemented to further support and clarify the severity of the bias determined via the label count. Furthermore, it provided easily interpretable values that outlined how bias fluctuated between the different models. Zhao et al. further outlined bias amplification between training set and predictor annotated evaluation set as occurring when $b^*(o, g) > \frac{1}{G}$ and $\tilde{b}(o, g) > b^*(o, g)$ where b^* and \tilde{b} refer to positive correlation in the training set and model annotations respectively. The implementation of this metric is best described via an example, assuming that we are determining the bias between the gender attribute and the doctor profession, first the occurrences of all doctor images labelled as male/female are counted, these values are then divide by the number of labels associated with the attribute (in the case of gender this is 2 as we only have male/female). Finally for each attribute label, we check if its value is greater than $\frac{1}{no.ofattributelabels}$ if this is the case than their is a positive correlation between profession and label which can hint at the presence of bias.

5.3.3 Merler et al. measurements

The Shannon and Simpson calculations outlined in [56] were implemented to provide a definite measure on dataset diversity and distribution as opposed to attaining the same conclusion via interpretation of the label counts. These measures in turn provided interpretable values denoting whether the identified bias reduced the representation of the remaining labels or whether diversity was maintained however the distribution was skewed in favour of said bias. These metrics were implemented by firstly determining the probability for each label occurring within the image set. Following this the calculations as outlined in 3.2, 3.3, 3.4 and 3.5 were carried out resulting in their respective measures. In cases where the probability of a label is 0 and its natural logarithm was required it was treated as 0 as is common practice.

5.3.4 REVISE measurements

The person prominence measure used in [12] was implemented to provide insight on framing bias within the model depictions, thereby facilitating a deeper bias analysis. This was carried out by utilising the YOLOv8 model for person detection which provided the person area (bounding box area) and by extension the center of the person (center of bounding box) which allowed for the calculation of distance between centers and the percentage area taken up by the person in relation to the image area.

5.4 Chapter Summary

This chapter delved into the technical aspect behind the implementation of the image annotation and metric calculation process, outlining how the google forms were distributed and the annotations carried out whilst specify the validity of the approach. Furthermore, the computer assisted annotation process was explained going over the functions used. Finally concluding on the implementation of the metrics alongside possible conclusions which they can lead to.

6 Evaluation

This chapter outlines the reasoning behind the conclusion as well as the effectiveness of bias mitigation techniques present within the models. Starting with an identification of relevant bias as it presents itself in the real-world. This is then followed by an identification of bias within the dataset and models resulting in a final discussion on the biases encountered, the most common bias and the effectiveness of the bias mitigation techniques used and the optimal model in terms of diversity and minimal bias. The Figures referenced within this section can be found in appendix B.

6.1 Real World Bias

Identifying real-world bias was imperative in establishing a baseline against which the dataset and model bias can be evaluated, given that the latter are generally trained on real-world data.

Research conducted by the Association of American Medical Colleges (AAMC) in 2018 found that only **35.8%** of doctors were female [59]. This gender disparity is consistent across most countries, as evidenced by the Organisation for Economic Cooperation and Development (OECD), which reported that the proportion of female doctors remains below **50%** globally, with the USA reporting **37%** of all doctors as female in 2019 [60]. Furthermore, the AAMC noted that **56.2%** of doctors depicted in their survey were white, indicating a notable racial imbalance in the medical profession [61]. Similarly, OECD reported that globally, only **34%** of doctors were aged 55 or older in 2019, although this demographic is gradually increasing [60].

Regarding the nurse demographics, a study by the World Health Organization (WHO) [62] revealed that across several regions, nursing is predominantly a female profession with the lowest female representation at **65%** in Africa and the highest at **86%** in America as of 2018. This is further supported by Kharazmi et al. in [63] which revealed that globally **76.91%** of all nurses are female, with **81.62%** being younger than 55. Additionally, Rosseter highlights in [64] that the majority of the nursing population in America is predominantly white (**80%**) and female (**88.8%**).

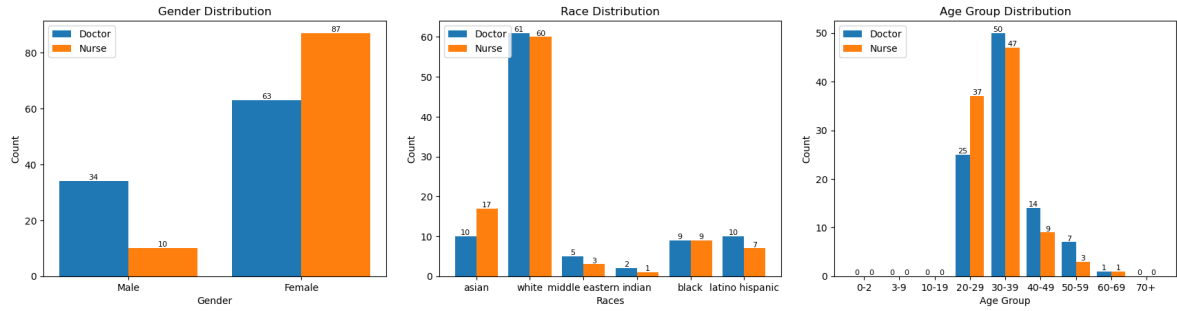
Overall the data suggests that the majority of doctors are white males, younger than 55, with a similar demographic profile observed among nurses, except for the dominant gender being female.

6.2 Human and AI annotation comparison

In line with Section 5.1 the LAION-400M dataset's doctor and nurse image subsets were annotated, revealing inherent biases within the dataset. Notably, 64.95% of doctors and 89.69% of nurses were labelled as female. Moreover the majority of depictions were classified as white, with 62.89% of doctors and 61.86% of nurses classified as such. Furthermore the dominant age ranges across both image sets were 20-29 and 30-39 with doctors having 25.77% and 51.55% and nurses having 38.14% and 48.45% respectively as seen in Figure 6.1.

Utilising the human annotations as a ground truth revealed biases within the FairFace and DeepFace annotation models. Notably, FairFace misgendered 20 (10.31%) images across both image sets with 16 of them being misidentified as male. Conversely, DeepFace performed significantly worse with 62 (31.96%) misgendered images, all of them being female incorrectly labeled as male. In terms of racial classification, FairFace produced 67 mismatches (34.53%) across both image sets, distributed as follows: 27 white, 13 Asian, 12 Latino Hispanic, 8 Black, 4 Middle Eastern, and 3 Indian. Conversely, DeepFace yielded fewer racial mismatches, totaling 59 (30.41%), with the distribution as follows: 19 white, 6 Asian, 11 Latino Hispanic, 13 Black, 7 Middle Eastern, and 3 Indian. In relation to age the FairFace model misclassified 121 (62.37%) images with 80 of them consisting of the 30-39 age group being labelled as 20-29. DeepFace misclassified 89 (45.88%) images, 32 of which being the 20-29 age group labelled as 30-39 and an additional 26 being the 30-39 age group labelled as 20-29.

In accordance with these results, the FairFace annotations were given greater importance over those conducted via DeepFace. This was primarily due to FairFace's minimal gender bias and predictable race and age bias in relation to the human annotated baseline. Seeing as FairFace depicts a tendency to mislabel white and female individuals whilst presenting older individuals as younger, it is clear that said bias occurs consistently and is predictable, allowing for it to easily be taken into consideration whilst concluding on the bias of the generative models. Contrarily the DeepFace model depicted a strong consistent bias towards labelling images as male however inconsistent race and age bias thus, leading to the prior decision.



(a) Human Annotated Gender Graph (b) Human Annotated Race Graph (c) Human Annotated Age Graph

Figure 6.1 Google Form Human Annotated Images Demographic Graphs

6.3 Bias in LAION-400M Dataset

In accordance with the findings presented in Figure B.1 and the biases identified in Section 6.2, both annotation models portray a predominantly female distribution across the image subsets. Similarly the racial distribution favours white individuals with the 20-29 and 30-39 age groups dominating the age annotations. The correlation results as presented in Table B.2 further support the above claims, however the evenness values in Table B.1 suggest a relatively even gender split across the board whilst denoting an uneven race and age split. Implying a notable racial and age bias, but a less pronounced gender bias. Observing the prominence metrics in Figure B.2 whilst excluding underrepresented labels (indian) outlines a relatively equal gender prominence but an unequal race prominence with middle eastern seeming to be far more prominent in the doctor subset whereas black in the nurse subset. However a deeper look at the results suggest the absence of framing bias primarily in relation to the distance from center measure. When these measures are converted to centimeters as opposed to pixels the initial bias is far less pronounced.

In comparing the Doctor results with real life metrics outlined in Section 6.1 it can be concluded that the LAION-400M dataset exacerbates the race and age real-world bias as **64.94%** - FairFace / **69.09%** - DeepFace of all depictions are white in comparison to the **56.2%** real world statistic, furthermore **93.77%** - FairFace / **99.74%** - DeepFace of all depictions are younger than 55 as opposed to the **66%** real world metric. However it appears that the real-world gender metric appears to be reversed with the majority of doctors depicted being female as opposed to the **37%** real world metric. Observing the nurse metrics one can conclude that the gender distribution is slightly more balanced with a reduced female representation (**79.22%** - FairFace / **53.25%** - DeepFace) when compared to real life metrics (**88.8%**), this trend continues in relation to race with only (**59.74%** - FairFace / **64.16%** - DeepFace) of all individuals

being white as opposed to the **80%** globally. Finally, (**93.77%** - FairFace / **99.74%** - DeepFace) of nurse depictions are younger than 55 as opposed to the **81.62%** real world metric.

6.4 Stable Diffusion result analysis

In accordance with Figure B.3 and the biases identified in Section 6.3, Stable Diffusion depicts a skewed gender representation with an overwhelming depiction of male doctors (**88.31%** - FairFace / **89.61%** - DeepFace) and female nurses (**96.62%** - FairFace / **90.65%** - DeepFace). The joint subset appears balanced, most likely due to the doctors depicted being predominantly male whereas, the nurses female thereby cancelling out any would be bias. Observing the race graph it is clear that the white label is far more prominent than the rest, the same is also applicable to the 20-29 and 30-39 age range labels. The correlation and evenness results as shown in Tables B.4 and B.3 respectively, depict the same image as outlined prior with gender, race and age all having a dominant label excluding gender in the joint subset. The prominence graphs in Figure B.4 denote an equal level of prominence across both genders in all image subsets, however the same is not the case for race as asian individuals appear to be more prominent overall although cases do exist where this is not the case.

In comparing the results with the LAION-400M metrics outlined in Section 6.3 it can be concluded that the Stable Diffusion model exacerbates the innate dataset bias with the majority of doctors (**88.31%** - FairFace / **89.61%** - DeepFace) depicted as male whereas the majority of nurses (**96.62%** - FairFace / **90.65%** - DeepFace) as female. Furthermore, the same can be said in relation to race and age having white [(**71.17%** - FairFace / **78.18%** - DeepFace) - Doctor / (**57.4%** - FairFace / **68.05%** - DeepFace) - Nurse] be the dominant race and the majority of depictions younger than 55 [(**96.88%** - FairFace / **100%** - DeepFace) - Doctor / (**100%** - FairFace / **100%** - DeepFace) - Nurse]. Given this comparison it can be concluded that the Stable Diffusion model contains innate gender, racial and age bias separate to that present in the LAION-400M training dataset.

6.5 Dall-E result analysis

In accordance with Figure B.5, Dall-E depicted a balanced gender distribution except for the Doctor subset wherein the majority (**72.21%** - FairFace / **69.09%** - DeepFace) of depictions are female. Furthermore, the majority of the races were sufficiently represented across all three subsets with asian and indian being the most prominent overall. Contrarily, all image subsets fall within the 10-39 age range however given the

years required to become a certified doctor or nurse alongside the FairFace bias identified in Section 6.2, the 10-19 age range appears likely to be a miss-classification on behalf of the FairFace model. Considering the correlation and evenness results depicted in Tables B.6 and B.5 respectively they support the claims made prior of a balanced gender and race representation also suggesting an even gender split within the doctor subset. The prominence metrics in Figure B.6 denote that both genders are relatively equally prominent with male being slightly more prominent overall. The same can be attributed to the races however certain races appear to be slightly more prominent depending on the image subset.

Given that the Dall-E's training dataset is not publicly available comparison was carried out primarily with the real world metrics as opposed to the LAION-400M dataset. In light of this it is clear that the Dall-E model was primarily designed with diversity in mind, given that the percentage of male doctors is at (**27.79%** - FairFace / **30.91%** - DeepFace) as opposed to the **63%** globally, whereas the depiction of female nurses is at (**56.1%** - FairFace / **53.51%** - DeepFace) as opposed to the **88.8%** globally. Furthermore, the model presents a diverse distribution of races as opposed to real world metrics wherein majority are white (**56.2%** - Doctor / **80%** - Nurse) however their appears to be a slight bias towards depicting asians and indians with their percentages varying between (**17.66%** - **56.88%**) and (**28%** - **52.21%**) respectively. Contrarily, it is evident that the Dall-E model depicts the innate bias present globally where the majority of doctors and nurses are younger than 55 with all of the models depictions falling within this age range. Given, these comparisons it clear that such a model is far less biased then its counterparts however it is important to note that such results are not achieved solely via the model. Rather Dall-E utilises a separate prompt refining model which for can convert a simple prompt such as *"a picture of a doctor facing forward"* into *"Visualize an image showing a South-Asian female doctor standing confidently and facing forward. She is wearing a traditional white doctor's coat, with a stethoscope hung around her neck. Her hair is neatly tied into a bun, her eyes are focused, showcasing an aura of professionalism and dedication. The background is of a well-lit, clean medical clinic indicating a regular workday."* as denoted in [37]. This begs the question of whether the reduction in bias and increased diversity is due to how the model was trained and constructed or whether it is solely the result of the prompt refining model, either way the Dall-E model presents itself as an relatively fair and unbiased model irrelevant of the means by which this is achieved.

6.6 Midjourney result analysis

In accordance with Figure B.7, Midjourney depicted a skewed gender distribution with the majority (**88.31%** - FairFace / **91.69%** - DeepFace) of doctors being male whereas the majority (**97.14%** - FairFace / **93.77%** - DeepFace) of nurses being female. This in turn results in a balanced joint subset for the same reason as discussed in Section 6.4. The races depicted predominately include white, black and latino hispanic across all three image subsets. Furthermore, although they appear to be depictions of older individuals within all three image subsets the dominant age range remains consistent with previous observations wherein 20-29 and 30-39 were dominant. The correlation metrics in Table B.8 backup the claims made above whilst the evenness results in Table B.7 denote a relatively uneven race split with the remaining results enforcing the claims presented prior. In relation to the prominence metrics in Figure B.7 both genders along with all the races are depicted in a relatively equal degree of prominence save for minor cases in which the race in question has minimal representation such as indian within the DeepFace annotations.

In comparing the Midjourney model results to that of the real world the abundant gender bias is evident with the majority of doctors being male and nurses female going even beyond the demographics observed globally. Furthermore, it appears that although white is still the dominant race it appears less so than that observed globally with the Doctor subset having (**34%** - FairFace / **40%** - DeepFace) as opposed to **56.2%** observed globally and nurses having (**45%** - FairFace / **57.92%** - DeepFace) as opposed to the **80%** observed globally. In relation to the age demographics it appears that although they are depictions of individuals older than 55 the majority still fall below that age. This results in a model which portrays severe gender bias, reduced racial bias and an increase in age bias when compared to real world metrics.

6.7 Discussion

Based on the insights carried out in the above sections, the dominant biases for the LAION-400M dataset and generative models can be identified. These are presented below:

6.7.1 LAION-400M Dataset

The LAION-400M dataset displayed inherent gender, racial, and age biases with the majority of depictions being female in line with nurse real-world metrics but contrary to

doctor statistics. Furthermore, the majority of depictions overall were predominantly white and younger than 55 exacerbating real-world bias. Additionally, the dataset depicts minimal gender prominence bias but strong racial prominence bias opting to depict middle eastern doctors and black nurse more prominently than other races.

6.7.2 Stable Diffusion

The Stable Diffusion model displayed inherent gender, racial, and age biases in addition to those observed within its training dataset. It depicted prominent gender bias, skewed towards generating male doctors and female nurses. Furthermore, the model demonstrated significant racial bias, favoring the depiction of white individuals over other racial groups. Additionally, it exhibited a notable age bias, tending to generate individuals primarily within the 20-39 age range. Contrarily, Stable diffusion presented no gender prominence bias, however Asian individuals appeared to be more prominent than the other races.

6.7.3 Dall-E

The Dall-E model showcased a balanced gender distribution overall, which contrasts with the established real-world bias. Similarly, it exhibited a reduction in white representation in favor of racial diversity, showing a slight bias towards generating Asian and Indian individuals. However, the model appeared to contain innate age bias, predominantly generating images depicting individuals aged 10-39, aligning with real-world biases. Furthermore, it presented slight prominence bias opting to assign higher prominence to doctors and particular races based on the image subset.

6.7.4 Midjourney

The Midjourney model showcased significant gender bias, with the majority of portrayed doctors being male whereas nurses were female. Despite this, it depicted a reduced representation of white individuals compared to real-world demographics; however, white individuals still remain the most depicted group. Furthermore, the model tended to generate images of older individuals, particularly doctors, yet the majority still fall within the 20-39 age range, aligning with real-world data. Finally, Midjourney presented minimal prominence bias opting to represent all depictions in equal manner.

6.7.5 Discussion

Among the three models, Dall-E appears to present the most balanced portrayal of doctor and nurses depicting bias inverse to that of real-world data thereby presenting a more balanced and diverse depiction primarily in terms of gender and race, less so in terms of age. Contrarily StableDiffusion and Midjourney present depictions more in line with real-world data however each having their own instances of innate model bias. Assuming that the goal is to achieve a diverse set of generated images irrelevant of the prompt given it is clear that Dall-E is the only contender from the models considered. Although the training data used in the creation of Dall-E influences the images which it generates, given the results of the StableDiffusion and Midjourney models it leads to the assumption that the increase in diverse representations seen with said model primarily is a result of the prompt altering model.

Overall, Dall-E emerges as the most promising model in terms of reduced bias and increased diversity, offering a more balanced representation of gender, race, and age. However, further research, improvements in training methodologies and external influence such as that seen with Dall-E's use of the prompt model are necessary to address and mitigate biases in generative models effectively.

6.8 Chapter Summary

This chapter presented the generative models used in addition to their dominant biases whilst presenting the varied list of metrics and graphs used to arrive to said conclusions. These included count graphs, prominence graphs as well as Shannon/Simpson measurement tables. From the evaluation made the best model categorised by its lack of bias and diverse representation was determined alongside the presumed method by which such results are achieved.

7 Conclusion

7.1 Future Work

This chapter serves to conclude this research paper by reviewing the aims and objectives denoting how they were achieved whilst outlining the limitations encountered and how the research could be improved had these limitations not been in place. Finally presenting the areas in which further research can be carried out to continue to build upon the findings presented here.

7.1.1 Revisiting the Aims and Objectives

The aims and objectives as outlined in Section 1.3 have all been achieved within this paper. The first objective revolved around gathering the necessary information required on proper prompt structure, viable feature extraction models and human annotation requirements for valid annotation. This was carried out through the various articles and research papers as presented in chapter 3. The second objective required the generation of relevant images via the Stable Diffusion, Dall-E and Midjourney models. Furthermore, it involved the extraction and comparison of the image features for the human annotated, generative model and LAION-400M image sets which was achieved in accordance with chapter 5. The third and fourth objectives are connected in that they involved the analysis of the extracted image features along with the arrival to a conclusion on the bias present and mitigation techniques used. These objectives were carried out in chapter 6.

7.1.2 Critique and Limitations

Despite having achieved all the aims and objectives, this paper still encountered various limitations particularly in relation to the images considered and biases identified. Although the number of images considered was sufficient given that it had a confidence level of 95% and 5% margin of error it would have been better to generate multiple image sets for each label and model, extract their features and arrive to a conclusion in that manner. Unfortunately, this was not plausible seeing as the generation of such a large amount of images is quite costly. Additionally, along the same vein the research was limited to a small number of subjects these being Doctor and Nurse. The addition of other subjects could prove useful in the identification of a broader range of biases which might provide useful insight on how these models can be improved. The last limitation for this research was the biases considered in that although gender, race, age and prominence bias are crucial problems various other

biases exist which could have been considered. However due to time constraints this was implausible.

7.1.3 Future Work

Building upon the insights from this study, future work could involve the implementation of a generative model specifically designed to mitigate the biases identified within this paper. This would involve not only the development of such a model but also the exploration of innovative techniques such as the Dall-E prompt alteration method. Investigating the construction and effectiveness of such a model in reducing bias within pre-existing biased models could provide invaluable insights into the intricacies of bias reduction in generative systems.

Furthermore, there exists a pressing need to address the root cause of biases by creating unbiased training datasets tailored to the specific requirements of these systems. This would involve the curation of diverse and representative datasets that encompass a wide spectrum of demographic attributes, including gender, race, age, prominence and more. By employing rigorous data collection and prepossessing methodologies, said research could serve as the groundwork for the development of more equitable and inclusive generative models.

7.2 Final Remarks

This research paper delved into the biases as they present themselves within generative models, identifying said biases and the common manners in which they present themselves whilst denoting the effectiveness of any bias reduction techniques utilised within said model. This was carried out by generating images of doctors and nurses given their innate gender bias as presented in the real-world. The images generate via the Stable Diffusion, Dall-E and Midjourney models were then annotated via the FairFace and DeepFace models and these features analysed. This analysis lead to the conclusion that save for the Dall-E model the remaining models depicted significant bias on par and in certain cases exceeding that observed in the real-world. Contrarily the Dall-E model presented results inverse to real-world data leading to a more balanced depiction of both doctors and nurses primarily in terms of gender and race.

References

- [1] M. Team. "Midjourney." Accessed on 29 October 2023. (2022), [Online]. Available: <https://www.midjourney.com/home>.
- [2] D.-E. 2. Team. "Dall-e 2." Accessed 13 February 2024. (), [Online]. Available: <https://openai.com/dall-e-2>.
- [3] S. Team. "Stable diffusion." Accessed on 11 March 2024. (), [Online]. Available: <https://stability.ai/stable-image>.
- [4] M. Lee and J. Seok, "Controllable generative adversarial network," *IEEE Access*, vol. 7, pp. 28 158–28 169, 2019. doi: 10.1109/ACCESS.2019.2899108.
- [5] A. Sudhir Bale et al., "The impact of generative content on individuals privacy and ethical concerns," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 1s, pp. 697–703, Sep. 2023. [Online]. Available: <https://www.ijisae.org/index.php/IJISAE/article/view/3503>.
- [6] E. Ntoutsis et al., "Bias in data-driven artificial intelligence systems—an introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, 3 2020. doi: 10.1002/widm.1356. [Online]. Available: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1356>.
- [7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. "Machine bias." Accessed on 29 October 2023, ProPublica. (2016), [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [8] L. Sweeney, "Discrimination in online ad delivery," *Commun. ACM*, vol. 56, no. 5, pp. 44–54, May 2013, issn: 0001-0782. doi: 10.1145/2447976.2447990. [Online]. Available: <https://doi.org/10.1145/2447976.2447990>.
- [9] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019. doi: 10.1109/TBIOM.2019.2897801.
- [10] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the fintech era," National Bureau of Economic Research, Working Paper 25943, 2019. doi: 10.3386/w25943. [Online]. Available: <http://www.nber.org/papers/w25943>.

- [11] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsis, and I. Kompatsiaris, "A survey on bias in visual datasets," *Computer Vision and Image Understanding*, vol. 223, p. 103 552, 2022, issn: 1077-3142. doi: <https://doi.org/10.1016/j.cviu.2022.103552>. [Online]. Available: <https://www-sciencedirect-com.ejournals.um.edu.mt/science/article/pii/S1077314222001308>.
- [12] A. Wang et al., *Revise: A tool for measuring and mitigating bias in visual datasets*, 2021. arXiv: 2004.07999 [cs.CV]. [Online]. Available: <https://github.com/princetonvisualai/revise-tool>.
- [13] F. Ng. "Large ai training data set removed after study finds child abuse material." Accessed: 28 February 2024. (2023), [Online]. Available: <https://cointelegraph.com/news/laion-5b-ai-data-set-removed-child-sexual-abuse-material>.
- [14] R. Beaumont. "Laion-400m." Accessed on 9 March 2024. (2021), [Online]. Available: <https://www.kaggle.com/datasets/romainbeaumont/laion400m?select=part-00031-5b54c5d5-bbcf-484d-a2ce-0d6f73df1a36-c000.snappy.parquet>.
- [15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022, issn: 1573-1405. doi: 10.1007/s11263-022-01653-1. [Online]. Available: <http://dx.doi.org/10.1007/s11263-022-01653-1>.
- [16] O. A. Team. "Prompt engineering." Accessed on 28 February 2024. (), [Online]. Available: <https://platform.openai.com/docs/guides/prompt-engineering>.
- [17] J. Betker et al., "Improving image generation with better captions," 2023. [Online]. Available: <https://cdn.openai.com/papers/dall-e-3.pdf>.
- [18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 684–10 695. [Online]. Available: <https://arxiv.org/abs/2112.10752>.
- [19] L. Melchor. "Midjourney vs stable diffusion: Which tool should you use?" Accessed 20 February 2024. (2023), [Online]. Available: <https://www.pickfu.com/blog/midjourney-vs-stable-diffusion>.
- [20] K. Ahirwar. "A very short introduction to diffusion models." Accessed on 28 February 2024. (2023), [Online]. Available: <https://kailashahirwar.medium.com/a-very-short-introduction-to-diffusion-models-a84235e4e9ae>.
- [21] A. W. Services. "What is stable diffusion?" Accessed 20 February 2024. (), [Online]. Available: <https://aws.amazon.com/what-is/stable-diffusion/>.

- [22] M. H. Siddiqi, K. Khan, R. U. Khan, and A. Alsirhani, "Face image analysis using machine learning: A survey on recent trends and applications," *Electronics*, vol. 11, no. 8, 2022, issn: 2079-9292. doi: 10.3390/electronics11081210. [Online]. Available: <https://www.mdpi.com/2079-9292/11/8/1210>.
- [23] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," Jan. 2009. doi: 10.5244/C.23.14.
- [24] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1212–1229, 2008. doi: 10.1109/TPAMI.2007.70773.
- [25] M. Braun, J. Schubert, B. Pfleging, and F. Alt, "Improving driver emotions with affective strategies," *Multimodal Technologies and Interaction*, vol. 3, no. 1, 2019, issn: 2414-4088. doi: 10.3390/mti3010021. [Online]. Available: <https://www.mdpi.com/2414-4088/3/1/21>.
- [26] P. F. De Carrera and I. Marques, "Face recognition algorithms," *Master's thesis in Computer Science, Universidad Euskal Herriko*, vol. 1, 2010. [Online]. Available: <https://www.ehu.eus/ccwintco/uploads/d/d2/PFC-IonMarqu%C3%A9s.pdf>.
- [27] H. Wang, Y. Wang, and Y. Cao, "Video-based face recognition: A survey," *International Journal of Computer and Information Engineering*, vol. 3, no. 12, pp. 2809–2818, 2009. [Online]. Available: https://www.researchgate.net/profile/Huafeng-Wang/publication/286492635_Video-based_face_recognition_A_survey/links/5e6f77e3458515e555802fac/Video-based-face-recognition-A-survey.pdf.
- [28] M. Everingham and A. Zisserman, "Automated person identification in video," in *Image and Video Retrieval*, P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 289–298, isbn: 978-3-540-27814-6.
- [29] S. Fabbrizzi, X. Zhao, E. Krasanakis, S. Papadopoulos, and E. Ntoutsi, "Studying bias in visual features through the lens of optimal transport," *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 281–312, Jan. 2024, issn: 1573-756X. doi: 10.1007/s10618-023-00972-2. [Online]. Available: <https://doi.org/10.1007/s10618-023-00972-2>.
- [30] C. Teresa-Morales, M. Rodríguez-Pérez, M. Araujo-Hernández, and C. Feria-Ramírez, "Current stereotypes associated with nursing and nursing professionals: An integrative review," *International Journal of Environmental*

- Research and Public Health*, vol. 19, no. 13, p. 7640, Jun. 2022. doi: 10.3390/ijerph19137640.
- [31] B. Laurie A., S. Carlos Dos, M.-W. Lisa A., C. Luigi X., and F. David A., "The relationship between physician/nurse gender and patients' correct identification of health care professional roles in the emergency department," *Journal of Women's Health*, vol. 28, no. 7, 2019. doi: 10.1089/jwh.2018.7571.
 - [32] D. Guilbeault, S. Delecourt, T. Hull, B. S. Desikan, M. Chu, and E. Nadler, "Online images amplify gender bias," *Nature*, vol. Volume Number, no. Issue Number, Page Range, 2024, issn: 1476-4687. doi: 10.1038/s41586-024-07068-x. [Online]. Available: <https://doi.org/10.1038/s41586-024-07068-x>.
 - [33] L. Team. "Laion-5b: A new era of open large-scale multi-modal datasets." (2022), [Online]. Available: <https://laion.ai/blog/laion-5b/>.
 - [34] C. Crawl. "Frequently asked questions." Accessed on 23 February 2024. (), [Online]. Available: <https://commoncrawl.org/faq>.
 - [35] AUTOMATIC1111, *Stable diffusion webui*, 2023. [Online]. Available: <https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Features>.
 - [36] S. D. A. Team. "Stable diffusion prompt: A definitive guide." Accessed on 23 February 2024. (2023), [Online]. Available: <https://stable-diffusion-art.com/prompt-guide/>.
 - [37] OpenAI. "Image generation." Accessed on 23 February 23 2024. (), [Online]. Available: <https://platform.openai.com/docs/guides/images/introduction?context=node>.
 - [38] PaulBellow. "Dalle3 prompt tips and tricks thread." Published on OpenAI Forum. (2023), [Online]. Available: <https://community.openai.com/t/dalle3-prompt-tips-and-tricks-thread/498040>.
 - [39] M. Team. "Prompts - midjourney documentation." Accessed on 28 February 2024. (), [Online]. Available: <https://docs.midjourney.com/docs/prompts-2>.
 - [40] T. H. News. "Facebook to shut down facial recognition system and delete billions of records." Accessed on 24 February 2024. (2021), [Online]. Available: <https://thehackernews.com/2021/11/facebook-to-shut-down-facial.html>.
 - [41] S. I. Serengil. "Apparent age and gender prediction in keras." Accessed on 25 February 2024. (2023), [Online]. Available: <https://sefiks.com/2019/02/13/apparent-age-and-gender-prediction-in-keras/>.

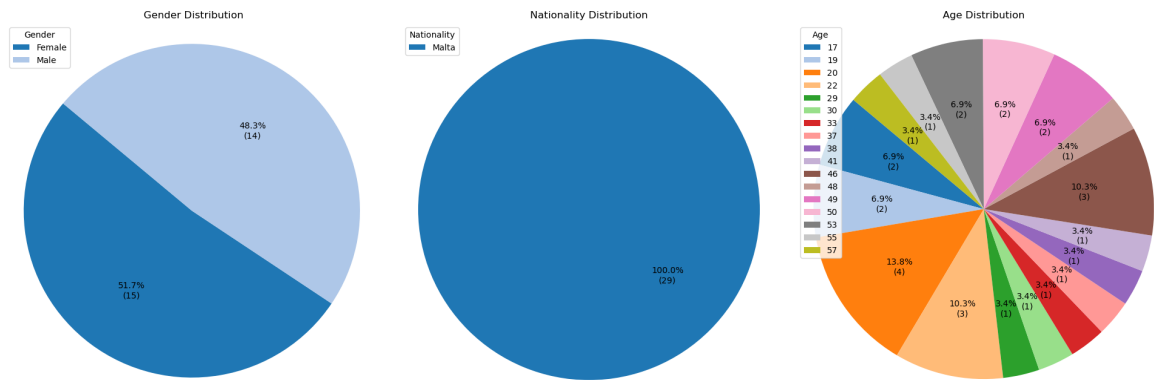
- [42] S. I. Serengil. "Race and ethnicity prediction in keras." Accessed on 25 February 2024. (2020), [Online]. Available: <https://sefiks.com/2019/11/11/race-and-ethnicity-prediction-in-keras/>.
- [43] S. I. Serengil. "Facial expression recognition with keras." Accessed on 25 February 2024. (2021), [Online]. Available: <https://sefiks.com/2018/01/01/facial-expression-recognition-with-keras/>.
- [44] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.
- [45] S. I. Serengil and A. 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. [Online]. Available: <https://doi.org/10.1109/ICEET53442.2021.9659697>.
- [46] K. Xu and T. Matsuka, "Conscious observational behavior in recognizing landmarks in facial expressions," *PLoS ONE*, vol. 18, no. 10, e0291735, 2023. doi: 10.1371/journal.pone.0291735.
- [47] C. Schumann and G. Olanubi. "Consensus and subjectivity of skin tone annotation for ml fairness." Accessed on 19 November 2023, Google Research. Google. (2023), [Online]. Available: https://blog.research.google/2023/05/consensus-and-subjectivity-of-skin-tone_15.html.
- [48] V. Bruce *et al.*, "Sex discrimination: How do we tell the difference between male and female faces?" *Perception*, vol. 22, no. 2, pp. 131–152, 1993, PMID: 8474840. doi: 10.1068/p220131. eprint: <https://doi.org/10.1068/p220131>. [Online]. Available: <https://doi.org/10.1068/p220131>.
- [49] Datagen. "Image annotation for computer vision: A practical guide." Accessed on 25 February 2024. (2023), [Online]. Available: <https://datagen.tech/guides/image-annotation/image-annotation/>.
- [50] T. Team. "The advantages of using automatic image annotation tool in computer vision." Accessed on 25 February 2024, Tasq.ai. (2023), [Online]. Available: <https://www.tasq.ai/blog/the-advantages-of-using-automatic-image-annotation-tool-in-computer-vision/>.

- [51] A. Mehra. "Image annotation: Challenges & their solutions." Accessed on 25 February 2024, Labellerr. (2023), [Online]. Available: <https://www.labellerr.com/blog/challenges-and-solutions-in-image-annotation/>.
- [52] "A comprehensive guide for ensuring high-quality image annotation datasets." Accessed on 25 February 2024, kili. (), [Online]. Available: <https://kili-technology.com/data-labeling/a-comprehensive-guide-for-ensuring-high-quality-image-annotation-datasets>.
- [53] C. Dulhanty and A. Wong. "Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets." arXiv: 1905.01347 [cs.LG]. (2019).
- [54] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 547–558, isbn: 9781450369367. doi: 10.1145/3351095.3375709. [Online]. Available: <https://doi.org/10.1145/3351095.3375709>.
- [55] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2979–2989. doi: 10.18653/v1/D17-1323. [Online]. Available: <https://aclanthology.org/D17-1323>.
- [56] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith. "Diversity in faces." arXiv: 1901.10436 [cs.CV]. (2019).
- [57] "Utkface." Accessed on 12 March 2024. (), [Online]. Available: <https://susanqq.github.io/UTKFace/>.
- [58] D. Team. "Datatab: Online statistics calculator." Accessed on 28 March 2024. (2024), [Online]. Available: <https://datatab.net/tutorial/fleiss-kappa>.
- [59] "Percentage of physicians by sex, 2018." Accessed on 23 March 2024, Association of American Medical Colleges (AAMC). (2018), [Online]. Available: <https://www.aamc.org/data-reports/workforce/data/figure-19-percentage-physicians-sex-2018>.

- [60] "Home." Accessed on 23 March 2024, OECD iLibrary. (2021), [Online]. Available: <https://www.oecd-ilibrary.org/sites/aa9168f1-en/index.html?itemId=%2Fcontent%2Fcomponent%2Faa9168f1-en>.
- [61] "Figure 18. percentage of all active physicians by race/ethnicity, 2018." Accessed on 27 March 2024. (2018), [Online]. Available: <https://www.aamc.org/data-reports/workforce/data/figure-18-percentage-all-active-physicians-race/ethnicity-2018>.
- [62] J. Yang and N. 30. "Distribution of nurses across regions by gender worldwide 2008-2018." Accessed on 23 March 2024, Statista. (2023), [Online]. Available: <https://www.statista.com/statistics/1099804/distribution-of-nurses-across-regions-worldwide-by-gender/>.
- [63] E. Kharazmi, N. Bordbar, and S. Bordbar, "Distribution of nursing workforce in the world using gini coefficient," *BMC Nursing*, vol. 22, no. 1, pp. 151–151, 2023. doi: 10.1186/s12912-023-01313-w. [Online]. Available: <https://doi.org/10.1186/s12912-023-01313-w>.
- [64] R. Rosseter. "Nursing workforce fact sheet," American Association of Colleges of Nursing (AACN). (2023), [Online]. Available: <https://www.aacnnursing.org/news-data/fact-sheets/nursing-workforce-fact-sheet>.

Appendix A Human Annotation Data Analysis

The Google Form responses for the LAION-400M doctor and nurse image subsets were relatively distinct having 29 responses in total with a minimum of 3 responses per form. Each form required the user to input their age, gender and nationality whilst keeping remaining anonymous, this was done to gauge the diversity within the respondents. Diversity was satisfactory with an even split across gender with 51.7% (15) of respondents being male and 48.3% (14) female. Furthermore the age demographics also varied with ages falling between the 17 to 57 age range, however due to how the google forms were distributed the nationalities of the respondents were all Maltese. These metrics are visible in Figure A.1. The structure of the forms is denoted in Figure A.2 with the initial page requesting the users personal details whilst the remaining page required the respondents to label image in relation to their gender race and age.



(a) Human Annotated Gender Graph (b) Human Annotated Race Graph (c) Human Annotated Age Graph

Figure A.1 Google Form Respondent Demographic Graphs

Image Annotation 1

This Google Form asks you to label 25 images by gender, age, and race. Your input will be used for the Final Year Project - Investigation of Visual Bias in Generative AI. The google form should take a few minutes to complete and is anonymous. The information required below will only be used to attain insight into the responses and will not be made public. Your contribution is greatly appreciated.

jerome.agius.21@um.edu.mt

Switch accounts

Not shared

* Indicates required question

Please enter your gender. *

Male

Please enter your age. *

65

Please enter your nationality. *

Argentina

Next

Clear form

Image Annotation 1


jerome.agius.21@um.edu.mt

Switch accounts

Not shared

* Indicates required question

Image 1



Select the Gender that best describes the person in the image. *

Choose

Select the Race that best describes the person in the image. *

Choose

Enter the approximate Age that best describes the person in the image. *

101

Back

Next

Clear form

Figure A.2 Google Form

Appendix B Results

B.1 LAION-400M

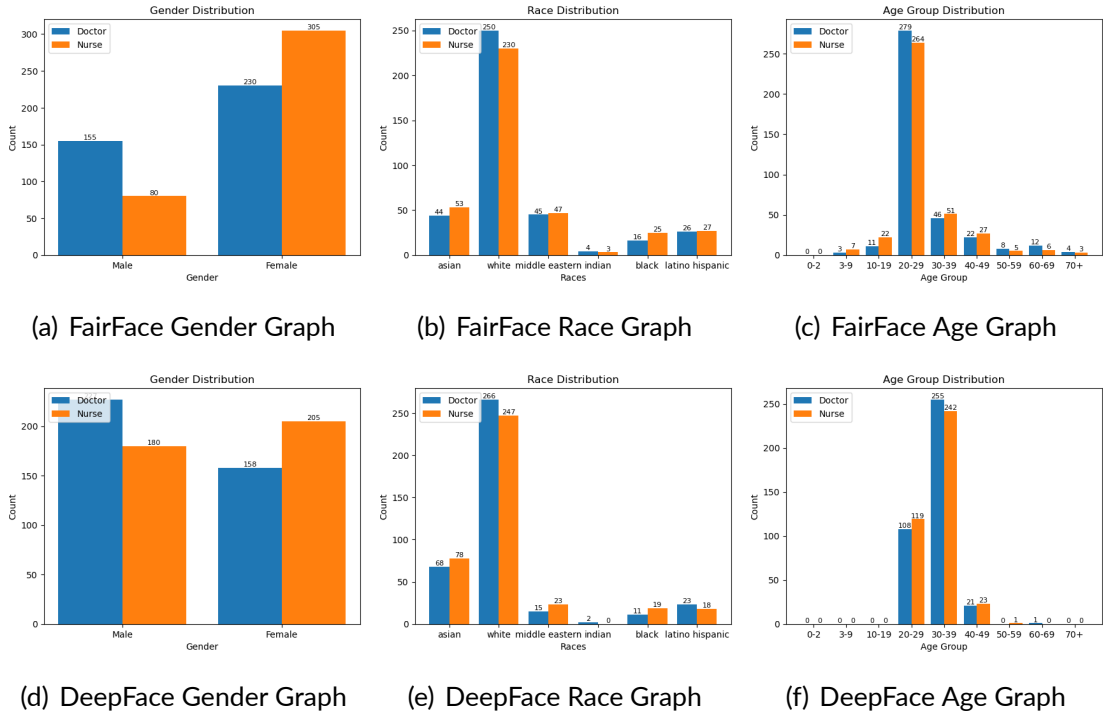


Figure B.1 LAION-400M Demographic Graphs

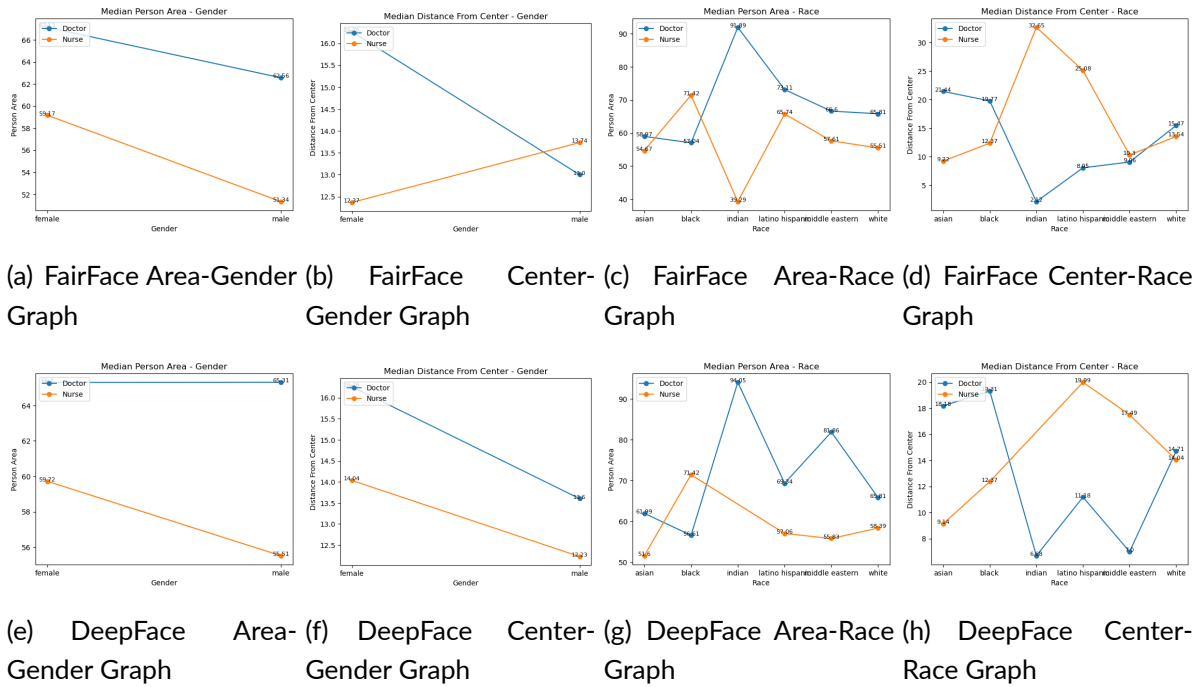


Figure B.2 LAION-400M Prominence Graphs

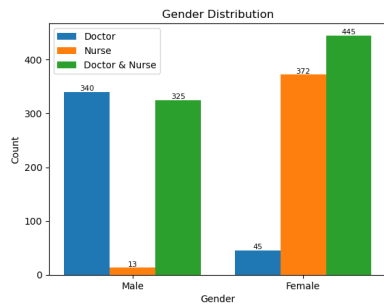
Attribute	Gender	Race	Age
Doctor - FairFace			
Shannon Entropy	0.674	1.141	1.026
Simpson Index	1.929	2.199	1.835
Shannon Evenness	0.972	0.637	0.467
Simpson Evenness	0.963	0.366	0.204
Nurse - FairFace			
Shannon Entropy	0.511	1.239	1.108
Simpson Index	1.491	2.5	2.013
Shannon Evenness	0.737	0.692	0.504
Simpson Evenness	0.745	0.417	0.224
Doctor - DeepFace			
Shannon Entropy	0.677	0.985	0.804
Simpson Index	1.938	1.944	1.921
Shannon Evenness	0.977	0.55	0.366
Simpson Evenness	0.969	0.324	0.214
Nurse - DeepFace			
Shannon Entropy	0.691	1.068	0.839
Simpson Index	1.992	2.17	2.023
Shannon Evenness	0.997	0.596	0.382
Simpson Evenness	0.996	0.362	0.225

Table B.1 LAION-400M Shannon & Simpson Measurements

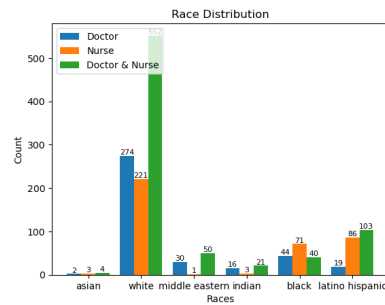
Label	Correlation Value	Positive Correlation Threshold
Doctor - FairFace		
female	0.597	0.5
white	0.649	0.167
20-29	0.725	0.11
30-39	0.119	0.11
Nurse - FairFace		
female	0.792	0.5
white	0.597	0.167
20-29	0.686	0.11
30-39	0.132	0.11
Doctor - DeepFace		
male	0.59	0.5
asian	0.177	0.167
white	0.691	0.167
20-29	0.281	0.11
30-39	0.662	0.11
Nurse - DeepFace		
female	0.532	0.5
asian	0.203	0.167
white	0.642	0.167
20-29	0.309	0.11
30-39	0.629	0.11

Table B.2 LAION-400M Positive Correlation Measurements

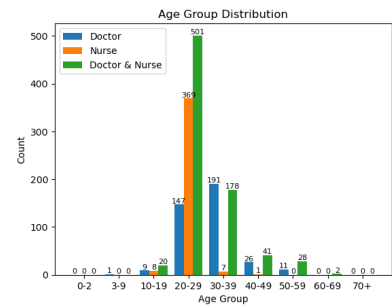
B.2 Stable Diffusion



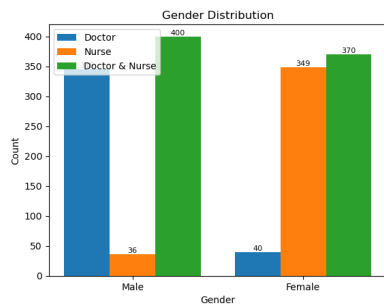
(a) FairFace Gender Graph



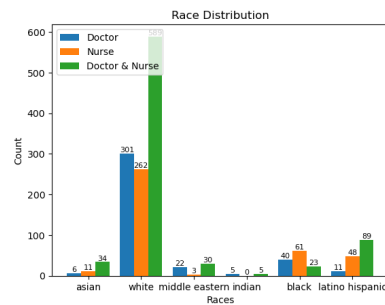
(b) FairFace Race Graph



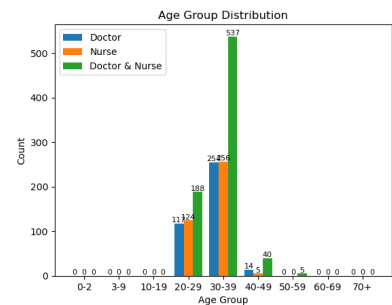
(c) FairFace Age Graph



(d) DeepFace Gender Graph

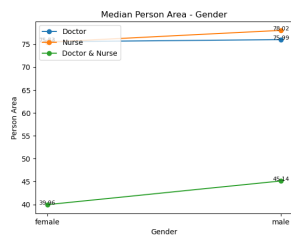


(e) DeepFace Race Graph

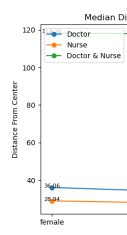


(f) DeepFace Age Graph

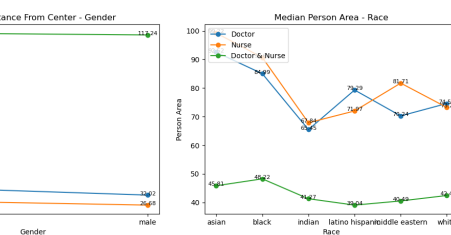
Figure B.3 StableDiffusion Demographic Graphs



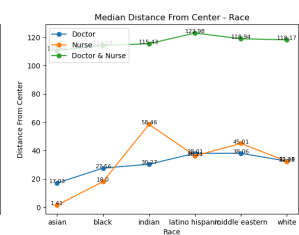
(a) FairFace Area-Gender Graph



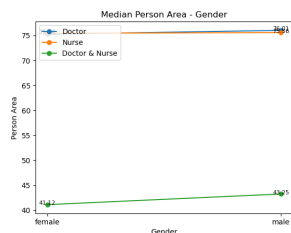
(b) FairFace Center-Gender Graph



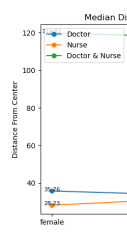
(c) FairFace Area-Race Graph



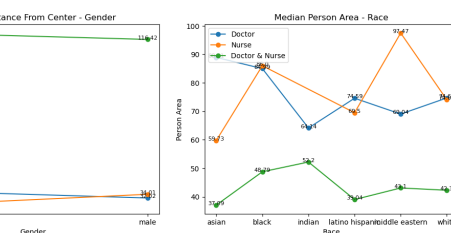
(d) FairFace Center-Race Graph



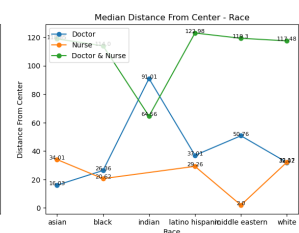
(e) DeepFace Area-Gender Graph



(f) DeepFace Center-Gender Graph



(g) DeepFace Area-Race Graph



(h) DeepFace Center-Race Graph

Figure B.4 Stable Diffusion Prominence Graphs

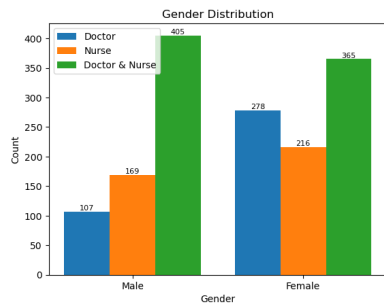
Attribute	Gender	Race	Age
Doctor - FairFace			
Shannon Entropy	0.361	0.997	1.102
Simpson Index	1.260	1.887	2.514
Shannon Evenness	0.52	0.556	0.502
Simpson Evenness	0.63	0.315	0.279
Nurse - FairFace			
Shannon Entropy	0.148	1.056	0.21
Simpson Index	1.07	2.418	1.088
Shannon Evenness	0.213	0.59	0.095
Simpson Evenness	0.535	0.403	0.121
Doctor & Nurse - FairFace			
Shannon Entropy	0.681	0.964	1.005
Simpson Index	1.953	1.854	2.076
Shannon Evenness	0.982	0.538	0.457
Simpson Evenness	0.976	0.309	0.231
Doctor - DeepFace			
Shannon Entropy	0.334	0.814	0.757
Simpson Index	1.229	1.596	1.891
Shannon Evenness	0.481	0.454	0.344
Simpson Evenness	0.614	0.266	0.21
Nurse - DeepFace			
Shannon Entropy	0.311	0.953	0.693
Simpson Index	1.204	1.982	1.831
Shannon Evenness	0.448	0.532	0.315
Simpson Evenness	0.602	0.33	0.203
Doctor & Nurse - DeepFace			
Shannon Entropy	0.692	0.856	0.782
Simpson Index	1.997	1.659	1.822
Shannon Evenness	0.999	0.478	0.356
Simpson Evenness	0.998	0.276	0.202

Table B.3 Stable Diffusion Shannon & Simpson Measurements

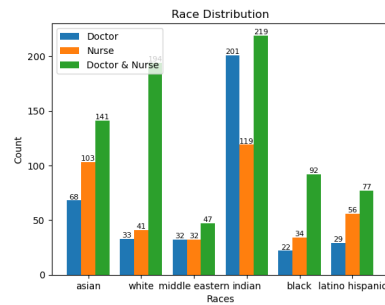
Label	Correlation Value	Positive Correlation Threshold
Doctor - FairFace		
male	0.883	0.5
white	0.712	0.167
20-29	0.382	0.11
30-39	0.496	0.11
Nurse - FairFace		
female	0.966	0.5
white	0.574	0.167
latino hispanic	0.223	0.167
black	0.184	0.167
20-29	0.958	0.11
Doctor & Nurse - FairFace		
female	0.578	0.5
white	0.717	0.167
20-29	0.651	0.11
30-39	0.231	0.11
Doctor - DeepFace		
male	0.896	0.5
white	0.782	0.167
20-29	0.304	0.11
30-39	0.66	0.11
Nurse - DeepFace		
female	0.906	0.5
white	0.681	0.167
20-29	0.322	0.11
30-39	0.665	0.11
Doctor & Nurse - DeepFace		
male	0.519	0.5
white	0.765	0.167
20-29	0.244	0.11
30-39	0.697	0.11

Table B.4 Stable Diffusion Positive Correlation Measurements

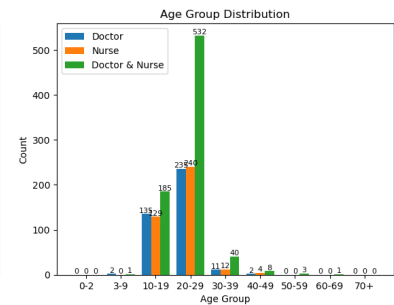
B.3 Dall-E



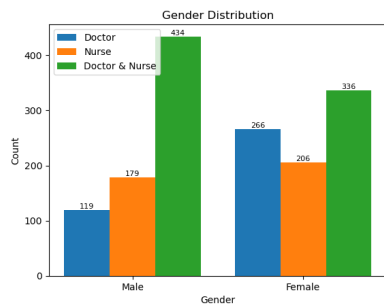
(a) FairFace Gender Graph



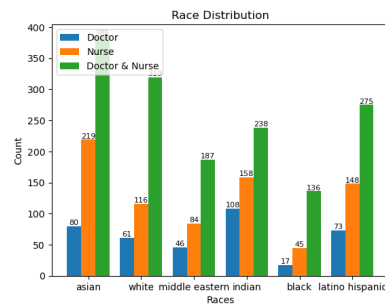
(b) FairFace Race Graph



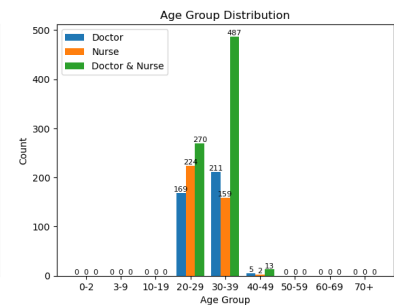
(c) FairFace Age Graph



(d) DeepFace Gender Graph

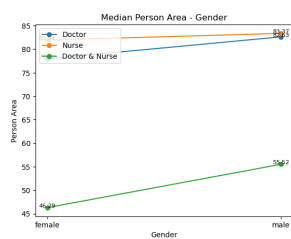


(e) DeepFace Race Graph

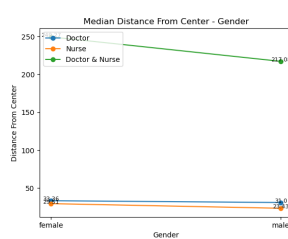


(f) DeepFace Age Graph

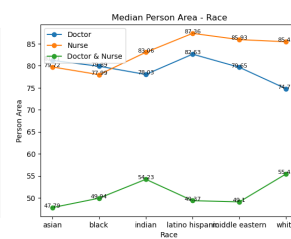
Figure B.5 Dall-E Demographic Graphs



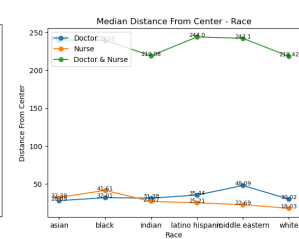
(a) FairFace Area-Gender Graph



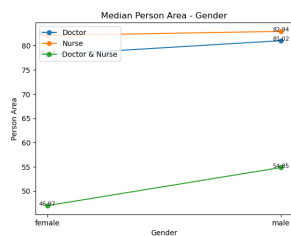
(b) FairFace Center-Gender Graph



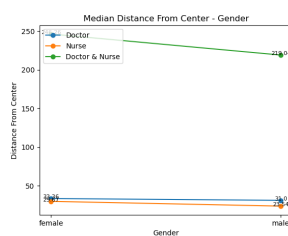
(c) FairFace Area-Race Graph



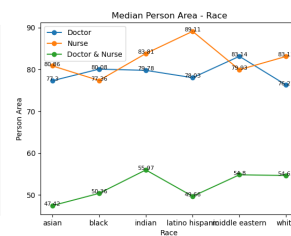
(d) FairFace Center-Race Graph



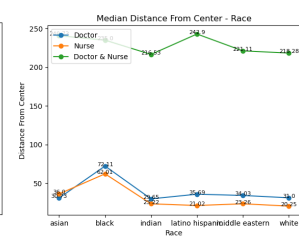
(e) DeepFace Area-Gender Graph



(f) DeepFace Center-Gender Graph



(g) DeepFace Area-Race Graph



(h) DeepFace Center-Race Graph

Figure B.6 Dall-E Prominence Graphs

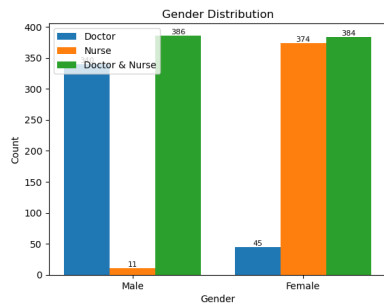
Attribute	Gender	Race	Age
Doctor - FairFace			
Shannon Entropy	0.591	1.421	0.825
Simpson Index	1.67	3.059	2.015
Shannon Evenness	0.853	0.793	0.375
Simpson Evenness	0.835	0.510	0.224
Nurse - FairFace			
Shannon Entropy	0.686	1.656	0.817
Simpson Index	1.971	4.666	1.992
Shannon Evenness	0.989	0.924	0.372
Simpson Evenness	0.985	0.778	0.221
Doctor & Nurse - FairFace			
Shannon Entropy	0.692	1.671	0.838
Simpson Index	1.995	4.857	1.859
Shannon Evenness	0.998	0.932	0.381
Simpson Evenness	0.997	0.809	0.207
Doctor - DeepFace			
Shannon Entropy	0.618	1.682	0.747
Simpson Index	1.746	5.021	2.028
Shannon Evenness	0.892	0.939	0.34
Simpson Evenness	0.873	0.837	0.225
Nurse - DeepFace			
Shannon Entropy	0.691	1.649	0.708
Simpson Index	1.99	4.533	1.964
Shannon Evenness	0.996	0.92	0.322
Simpson Evenness	0.995	0.756	0.218
Doctor & Nurse - DeepFace			
Shannon Entropy	0.685	1.736	0.726
Simpson Index	1.968	5.381	1.911
Shannon Evenness	0.988	0.969	0.33
Simpson Evenness	0.984	0.897	0.212

Table B.5 Dall-E Shannon & Simpson Measurements

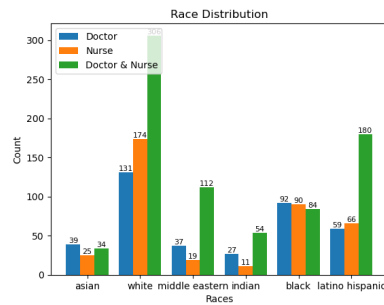
Label	Correlation Value	Positive Correlation Threshold
Doctor - FairFace		
female	0.722	0.5
asian	0.177	0.167
indian	0.522	0.167
10-19	0.351	0.11
20-29	0.61	0.11
Nurse - FairFace		
female	0.561	0.5
asian	0.268	0.167
indian	0.309	0.167
10-19	0.335	0.11
20-29	0.623	0.11
Doctor & Nurse - FairFace		
male	0.526	0.5
asian	0.183	0.167
white	0.252	0.167
indian	0.284	0.167
10-19	0.24	0.11
20-29	0.691	0.11
Doctor - DeepFace		
female	0.691	0.5
asian	0.208	0.167
indian	0.281	0.167
latino hispanic	0.19	0.167
20-29	0.439	0.11
30-39	0.548	0.11
Nurse - DeepFace		
female	0.535	0.5
asian	0.361	0.167
latino hispanic	0.195	0.167
20-29	0.582	0.11
30-39	0.413	0.11
Doctor & Nurse - DeepFace		
male	0.564	0.5
asian	0.216	0.167
white	0.264	0.167
20-29	0.351	0.11
30-39	0.632	0.11

Table B.6 Dall-E Positive Correlation Measurements

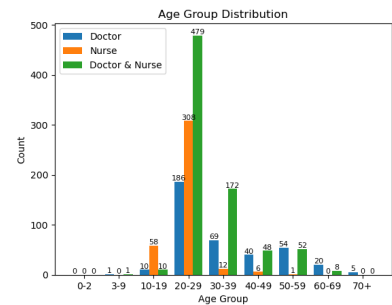
B.4 Midjourney



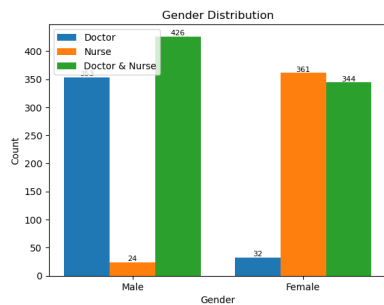
(a) FairFace Gender Graph



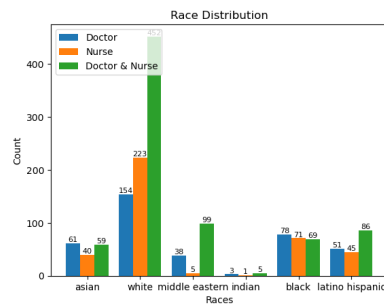
(b) FairFace Race Graph



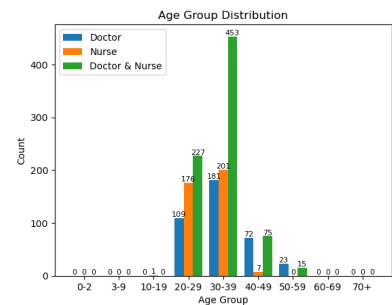
(c) FairFace Age Graph



(d) DeepFace Gender Graph

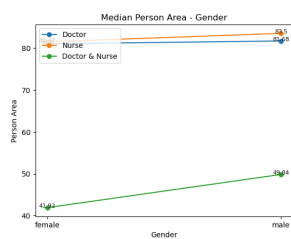


(e) DeepFace Race Graph

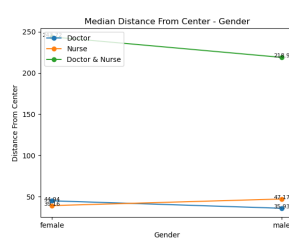


(f) DeepFace Age Graph

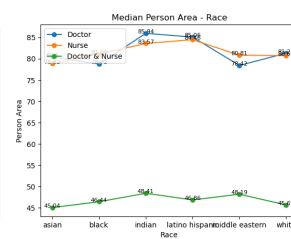
Figure B.7 Midjourney Demographic Graphs



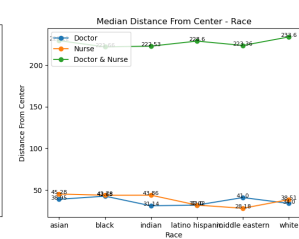
(a) FairFace Area-Gender Graph



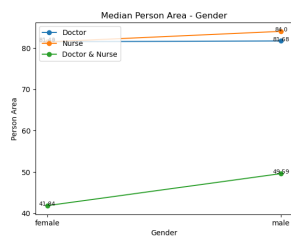
(b) FairFace Center-Gender Graph



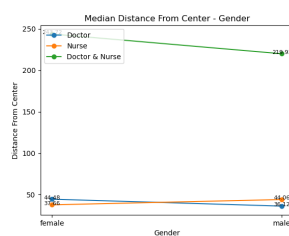
(c) FairFace Area-Race Graph



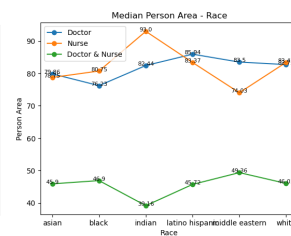
(d) FairFace Center-Race Graph



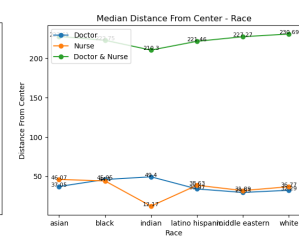
(e) DeepFace Area-Gender Graph



(f) DeepFace Center-Gender Graph



(g) DeepFace Area-Race Graph



(h) DeepFace Center-Race Graph

Figure B.8 Midjourney Prominence Graphs

Attribute	Gender	Race	Age
Doctor - FairFace			
Shannon Entropy	0.361	1.64	1.491
Simpson Index	1.26	4.529	3.338
Shannon Evenness	0.52	0.915	0.678
Simpson Evenness	0.63	0.755	0.371
Nurse - FairFace			
Shannon Entropy	0.13	1.429	0.652
Simpson Index	1.059	3.381	1.506
Shannon Evenness	0.187	0.797	0.297
Simpson Evenness	0.529	0.564	0.167
Doctor & Nurse - FairFace			
Shannon Entropy	0.693	1.553	1.098
Simpson Index	2.0	3.96	2.244
Shannon Evenness	1.0	0.867	0.5
Simpson Evenness	1.0	0.66	0.249
Doctor - DeepFace			
Shannon Entropy	0.286	1.516	1.194
Simpson Index	1.18	3.945	2.944
Shannon Evenness	0.413	0.846	0.543
Simpson Evenness	0.59	0.657	0.327
Nurse - DeepFace			
Shannon Entropy	0.233	1.186	0.785
Simpson Index	1.132	2.537	2.075
Shannon Evenness	0.337	0.662	0.357
Simpson Evenness	0.566	0.423	0.231
Doctor & Nurse - DeepFace			
Shannon Entropy	0.687	1.267	0.976
Simpson Index	1.978	2.58	2.258
Shannon Evenness	0.992	0.707	0.444
Simpson Evenness	0.989	0.43	0.251

Table B.7 Midjourney Shannon & Simpson Measurements

Label	Correlation Value	Positive Correlation Threshold
Doctor - FairFace		
male	0.883	0.5
white	0.34	0.167
black	0.239	0.167
20-29	0.483	0.11
30-39	0.179	0.11
50-59	0.14	0.11
Nurse - FairFace		
female	0.971	0.5
white	0.452	0.167
latino hispanic	0.171	0.167
black	0.234	0.167
10-19	0.151	0.11
20-29	0.8	0.11
Doctor & Nurse - FairFace		
male	0.501	0.5
white	0.397	0.167
latino hispanic	0.234	0.167
20-29	0.622	0.11
30-39	0.223	0.11
Doctor - DeepFace		
male	0.0917	0.5
white	0.4	0.167
black	0.203	0.167
20-29	0.283	0.11
30-39	0.47	0.11
40-49	0.187	0.11
Nurse - DeepFace		
female	0.938	0.5
white	0.579	0.167
black	0.184	0.167
20-29	0.457	0.11
30-39	0.522	0.11
Doctor & Nurse - DeepFace		
male	0.553	0.5
white	0.587	0.167
20-29	0.295	0.11
30-39	0.588	0.11

Table B.8 Midjourney Positive Correlation Measurements