Investigation of Visual Bias in Generative AI

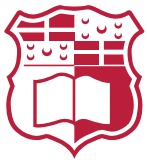
Jerome Agius

Supervisor: Dr Dylan Seychell

Co-Supervisor: Prof. 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 field 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 regarding the presence of bias within AI generated images, 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 the DeepFace and FairFace feature extraction models, whose accuracy was evaluated on a human annotated subset of LAION-400M images. Furthermore, the bias present within the images was deduced via a series of metrics these being; label count, correlation, person prominence, Shannon entropy, Simpson index and the latter twos evenness counterparts. 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 wish to express my sincere appreciation to my supervisor Dr. Dylan Seychell and co-supervisor Prof. John Abela, for their guidance and support throughout the completion of this final year project. Their expertise and mentorship played a crucial role in navigating the challenges encountered during this endeavor. Furthermore, I am deeply grateful to my parents, Reno and Graziella, and my brother Julian, for their encouragement and support throughout this academic pursuit.

Contents

List of Figures

List of Tables

List of Abbreviations

AAMC Association of American Medical Colleges.

AI Artificial intelligence.

API Application programming interface.

CLIP Contrastive language-image pre-training.

GAN Generative adversarial network.

OECD Organisation for Economic Cooperation and Development.

SVM Support vector machine.

VAE Variational autoencoder.

1 Introduction

1.1 Problem Definition

In recent years, the field of Generative artificial intelligence (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: 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, race, age, and prominence 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] systems. Furthermore, techniques such as Word2Vec have also fallen under scrutiny due to their innate gender and racial biases as outlined in [11], particularly when applied within systems such as Web search and parsing Curricula Vitae. This paper aims to build upon this research however focusing on generative AI.

Bias serves to affect a large majority of computer vision systems such as classification algorithms, face recognition systems, object detectors and many more [12]. 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 [12].

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 [13]. 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 the LAION-5B and Stable Diffusion models when considering 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 [14] with the LAION-5B dataset, its access has been revoked. In light of this the aim of the study remained the same however the focus shifted to the LAION-400M dataset [1] whilst also considering

[1] Accessible via a Kaggle repository [15]

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, allowing individuals to easily detect bias and replicate the results shown.

1.2 Motivation

The motivation behind this research stems from the growing importance of addressing bias in 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 conclusions presented throughout this thesis. This aim was 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, correlation, 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 thesis is divided into six sections, with this section providing an introduction to the field in which the research resides whilst outlining the primary aims and objectives. The background and literature review section provides additional information, technical or otherwise which is required to fully understand the means by which the research was carried out, whilst delving 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 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.

2 Background and Literature Review

This chapter provides the knowledge base required for understanding the techniques employed within the bias detection pipeline. Reviewing studies related to prompting, diffusion models, facial analysis, image bias, image annotation, and bias measurement techniques. Thereby laying the groundwork for the research conducted throughout this 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 biases 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 issue as slight alterations to the prompt can have a major impact on model performance and output, additionally finding the appropriate prompt is a time-consuming endeavour [16]. Prompt engineering addresses these issues by altering the prompt length and wording used to effectively arrive at the required output [17]. Automated prompt engineering can further enhance this process; however, it was beyond the scope of this paper.

2.1.1 Prompt target

In accordance with the nature of this research, the prompt subjects consist of traditionally gender-biased professions, specifically doctor and nurse, as they are male and female dominated, respectively, as showcased in [18, 19]. Furthermore, this gender bias perpetuates itself online, as seen in [20], thus making it more pertinent 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 [21, 22].

2.1.2 Prompt structure

Stable Diffusion

The Stable Diffusion WebUI repository [23] outlines the tool's functionality, covering upscaling, img2img, negative prompting, face restoration, and model merging. It further specifies that Stable Diffusion accepts prompts of up to 75 tokens, with additional token chunks allotted in instances of longer prompts. However, there is an overall limit which, if surpassed, triggers a warning and prompt truncation. Despite this, the 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), and image style (hyper-realistic, fantasy) while utilising negative prompts in accordance with the Stable Diffusion prompt guide [24] to achieve the best results.

DALL-E 3

The official DALL-E 3 documentation [25] 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 [26] offers general prompting tips such as being specific and detailed, using descriptive adjectives, avoiding prompt overloading, and specifying desired styles or themes.

Midjourney

The official Midjourney documentation [27] 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. The documentation further suggests using simple, short sentences as opposed to a long list of requests and instructions to achieve 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 2.1.

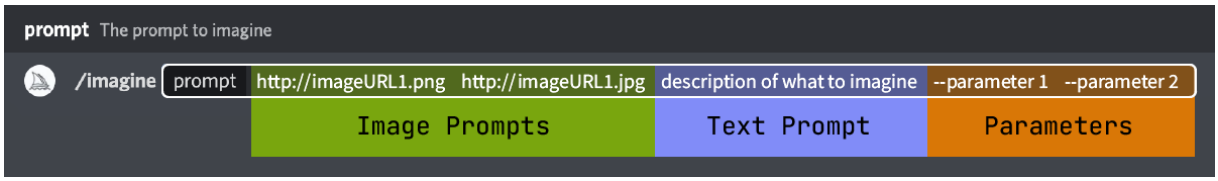


Figure 2.1 Midjourney prompt structure [27].

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, whilst longer, detailed prompts provide greater control.

Standard Prompt Structure

According to the literature reviewed, it is evident that while the models vary, their prompt features remain consistent. Thus, a standard prompt structure for image generation, applicable to the three aforementioned models, can be proposed: ***A picture of a [subject] facing forward***, where **subject** is replaced with doctor/nurse/doctor and nurse based on the images to be generated. Furthermore, ***Disfigured*** and ***Art*** serve as negative prompts for clear, realistic depictions. This prompt structure integrates various suggestions; **picture** is used to emphasise the medium and realistic nature of the required image, **subject** is used to focus on the desired aspects, **facing forward** along with the negative prompts serve to ensure facial clarity, whilst the short length of the prompt allows for model creativity whilst limiting human influence.

2.2 Diffusion Models

Generative models encompass a variety of different approaches, including generative adversarial networks (GAN), variational autoencoders (VAE), and diffusion models, with the latter offering 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 [28–30].

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 2.2.

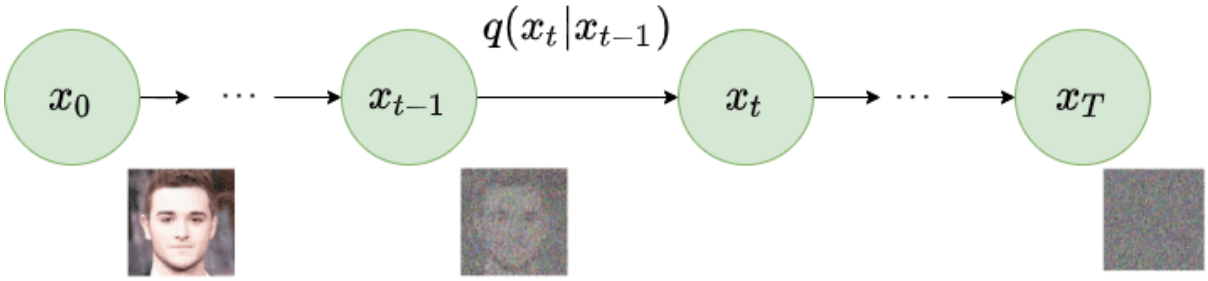


Figure 2.2 Forward diffusion process [31].

The reverse diffusion process resembles the inverse of the forward diffusion process as depicted in Figure 2.3, 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 [29, 32]. Furthermore, conditioning prompts serve to guide the diffusion process towards specific image themes or styles [32]. The U-shaped encoder-decoder network architecture (UNET) originally developed for image segmentation in biomedicine serves as the core component within this process, with the majority of generative models adopting the Residual neural network (ResNet) variant developed for computer vision, throughout their image generation pipeline.

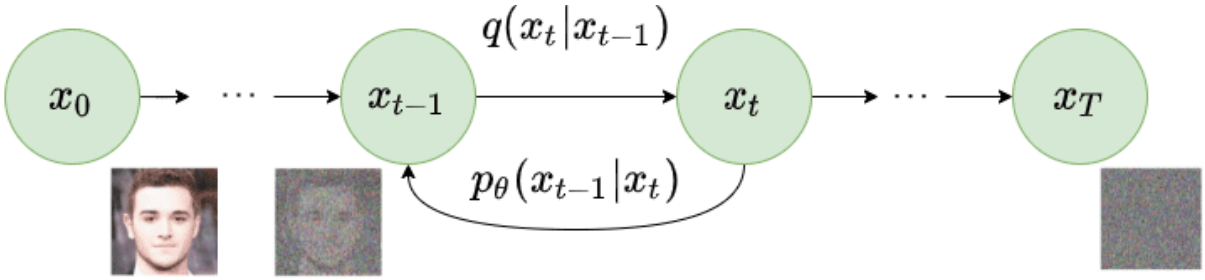


Figure 2.3 Revere diffusion process (Right to Left) [31].

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 individuals faces, age, gender, race and head pose based on the task [33]. These 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 people's attention [34], discerning an advertisements level of engagement based on an individual's attention [35], ensuring driving safety by monitoring the emotional state of the driver [36], 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 [37–39].

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 and Deep Learning based methods, with the latter being the most used. Deep Learning based methods involve the training of a Convolution Neural Network 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 application programming interface (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 and text detection, 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 model 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 2.4. 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 a mean absolute error 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. Amongst varied validation datasets and considering individuals of all races, gender classification accuracy fluctuated between 92% and 98.1%, age classification accuracy varied between 56.5% and 61.6% whilst race classification accuracy varied between 75.4% and 97% [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 2.4 DeepFace emotion model structure [45].

2.4 Image Bias

Image bias within 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 [46].

These types of biases tend to occur unintentionally 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 arrive at a conclusion based on the resultant data. These approaches usually go hand in hand as can be seen in the REVISE implementation [13] wherein the tool itself utilises 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 [12]:

- Selection bias
  - Data representativeness - do we need balanced or statistically representative data?
  - Negative set coverage - are the negative sets adequately represented?
  - Excluded groups - are their 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 the data perpetuate harmful biases?
- Label bias
  - Automated labelling biases - has the innate machine bias been taken into consideration or mitigated?
  - Annotator bias control - is their a diverse team of annotators such that human bias is mitigated?
  - Label clarity - are fuzzy labels (gender/race) being used?

2.5 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.

2.5.1 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 recognise positive and negative expressions with minimal facial expression information [47]. 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 [48]. Moreover, humans demonstrate high accuracy in specific annotation tasks, such as face and gender annotation, achieving a 96% accuracy whilst excluding contextual cues like hairstyle and makeup [49].

On the other hand, challenges include annotator bias, where cultural and experiential differences influence image annotations as is depicted in Figure 2.5 [48]. Furthermore, subjective interpretations of ambiguous visual cues result in consistency issues impacting data reliability. Scalability and cost concerns, also serve as major hurdles as manual annotation is time-consuming and expensive, particularly for large datasets, leading to the adoption of computer-assisted image annotation [50].

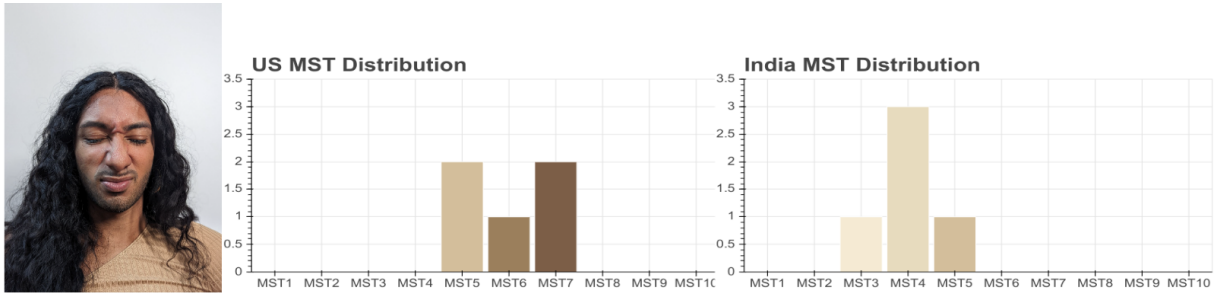


Figure 2.5 The distribution of Monk Skin Tone Scale annotations for this image from a sample of five photographers in the U.S. and five photographers in India [48].

Furthermore, considerations for effective human-based image annotation include selecting annotators from geographically diverse backgrounds to ensure accurate annotations, as advocated in [48], 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 [50].

2.5.2 Computer assisted image annotation

Computer assisted image annotation makes use of AI models such as those discussed in Section 2.3 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 such a system can annotate large volumes of images quickly without additional costs, whilst ensuring consistent annotations, thereby saving time and resources [51].

Contrarily, such systems are heavily reliant on their training dataset in order to produce accurate results, as said dataset dictates innate model biases and significantly influences system performance. Additionally, such systems are limited in understanding nuances and task specific context, leading to misinterpretations and errors, especially in complex or ambiguous scenarios [52].

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 [53].

2.6 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 Fabbrizzi et al. in [12] a total of twenty four papers and their bias detection strategies were reviewed, and categorised into four groups. The first group involves measuring bias using existing or extracted attributes and labels, treating them as if they were structured in a tabular dataset. The techniques relevant to this paper all fall within this group. Another group of techniques discovers bias by observing lower-dimensional representations of the data, while a third group uncovers bias through cross-dataset comparisons. The remaining techniques, which did not fit into these categories, were classified as ”other.”

2.6.1 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 it 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 [54] 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. Given that the dataset is suitably labelled, such a method would provide insight into selection bias along with the framing of protected attributes. However, it should be noted that this approach relies on the absence of innate bias within the recognition models. Such a requirement is rarely met and as such said approach is not fully reliable.

Yang et al. [55] similarly opted to address selection and label bias in relation to the person category of the ImageNet dataset. To address label bias, annotators were tasked with removing any images deemed offensive or sensitive such as those being sexual in nature or depicting racial slurs. Furthermore, images with ambiguous labels were also removed. Following this the annotators then labelled the remaining images in accordance with a predetermined set of categories, these being gender, age and skin colour. This was done so as to understand the bias present within the dataset. The validity of the annotation process was assessed by evaluating the agreeableness amongst annotators using a small, carefully selected set of images.

Zhao et al. [56] measured the correlation between protected attributes and the frequency of certain objects or actions. This was carried out via (2.1) wherein G was a protected attribute (e.g., gender) having values $g_1, ..., g_n$ and $O = o$ served to describe the occurrence of an object or action in the image (e.g., a sports implement or an outdoor scene). The bias score concerning protected attribute $G = g$ and object $O = o$ was denoted by $b(o, g)$ whereas $c(o, x)$ produced a count of the co-occurrences between the object/action o and the protected attribute's value x. Assuming that $b(o, g_i) > \frac{1}{n}$, would imply that the attribute g is positively related to object/action o, given that n represents the number of possible values for the protected attribute.

$$b(o, g) = \frac{c(o, g)}{\sum_{x \in \{g_1 ... g_n\}} c(o, x)} \tag{2.1}$$

Information theoretical

Merler et al. [57] presented four measurements for a balanced dataset. The Shannon entropy and Simpson Index in (2.2) and (2.3) measure diversity with the larger values depicting greater diversity. Additionally the Shannon and Simpson evenness measures in (2.4) and (2.5) denote how evenly distributed the dataset labels are across the entire dataset with the maximum value being 1. Lastly, it is worth noting that $X = x_i$ represents value x_i being assigned to attribute X, $P(X = x_i)$ is the probability that this occurs whilst n corresponds to the total number of labels.

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \cdot \ln(P(X = x_i)) \tag{2.2}$$

$$D(X) = \frac{1}{\sum_{i=1}^{n} P(X = x_i)^2} \tag{2.3}$$

$$E_{Shannon} = \frac{H(X)}{\ln(n)} \tag{2.4}$$

$$E_{Simpson} = \frac{D(X)}{n} \tag{2.5}$$

2.6.2 REVISE

The revise tool [13] 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, water, ice, and 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 (2.6) 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.

$$dist = \frac{distance\ between\ p\ and\ o\ centres}{\sqrt{area_p * area_o}} \tag{2.6}$$

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 numbers, or distances are not depicting the entire situation. This was carried out by extracting FC7 features from AlexNet on a subset of images. Scene-level features were then extracted and projected into $\sqrt{number of samples}$ dimensions to prevent over-fitting. Finally, a Linear SVM was fitted to determine if it learns the difference between images containing the same object however 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*.

2.7 Chapter Summary

This chapter provided a baseline knowledge on the key concepts and techniques required to make sense of the findings presented in this paper whilst delving deeper into the established research and techniques regarding prompt structure and target selection, facial analysis models and their inherent challenges, the two primary image annotation techniques, and an exploration of various bias measurement techniques employed in similar research areas.

3 Specification and Design

The specification and design chapter provides a detailed explanation of the steps taken to arrive at the conclusion presented in Chapter 5, going over its various components whilst detailing and justifying the decisions taken throughout. It further links the research carried out in Chapter 2 with the implementation, whilst offering a critical review of the entire process, going over the challenges encountered and their solutions. The entire process is outlined in Figure 3.1.

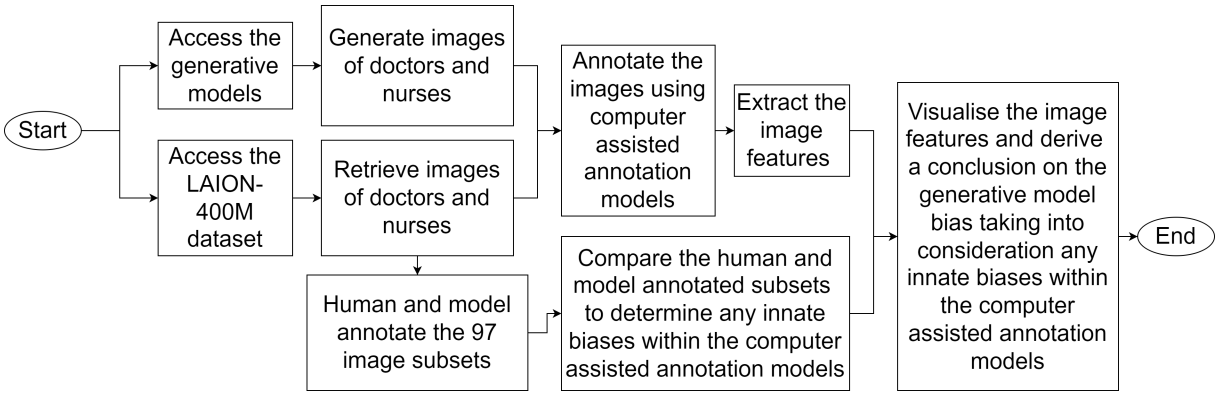


Figure 3.1 Process pipeline used throughout this research paper.

3.1 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 [14]. Although steps were taken to facilitate access to the dataset, the issue persisted. Fortunately, Kaggle still hosted access to some of LAION's datasets, in particular LAION-400M [15]. Given that said dataset preceded the LAION-5B and was curated in the same manner as the LAION-5B dataset, albeit 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.

3.2 Image Retrieval

The initial challenge in this section was deducing the image subjects, i.e., what the images will depict. In line with Section 2.1.1, *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 contrastive language-image pre-training (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 explicit 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 or 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 2.1.2, where subject was substituted for *doctor*, *nurse*, and *doctor and nurse*. The latter label was added simply to deduce if the presence of both professions in a singular prompt would affect the bias therein. The Stable Diffusion model was accessed and used via [29] whilst utilising the interface provided in [23]. 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, images were randomly selected from the doctor and nurse subsets, resulting in two subsets respectively. Furthermore, each model generated three sets of images, resulting in a total of nine subsets, with each set associated with the doctor, nurse, and doctor and nurse prompts.

3.3 Image Annotation

Following image retrieval, the next step was image annotation. This was divided into two parts: human and computer-assisted image annotation. Human annotation was carried out using two 97-image subsets, with each subset containing images of doctors or nurses, randomly selected from the LAION-400M dataset. Each subset consisted of 97 images as this minimises the time required for the annotation task whilst maintaining a confidence score of 95% with a 10% margin of error. The increase in margin of error from 5% to 10% 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. Only the image subsets from the LAION-400M dataset were human-annotated, as doing otherwise would conflict with our time constraint. Additionally, these subsets were sufficient in gauging any pre-existing bias within the DeepFace and FairFace models, as outlined in Section 5.2.

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 2.3, as it provided adequately accurate gender, age, and race classification. Emotion classification was initially considered; however, it was deemed redundant as no 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 3.2. The UTK-Face dataset [58] was used as it possessed varied images in terms of 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 whilst presenting ambiguity regarding the weighting of each component model. As such, the decision was taken to use the two models separately and interpret their results accordingly.

3.4 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 were implemented from those outlined in Section 2.6, including label count, correlation, person

prominence, Shannon entropy, Simpson index, Shannon evenness, and Simpson evenness. These were chosen for their applicability to the research and their ability to systematically identify bias and assess its severity, as outlined below:

1. The label count reveals potential bias, such as an overabundance of male-labelled images, indicating possible gender bias.
2. Correlation confirms and quantifies bias severity, with larger values indicating severe bias.
3. Shannon entropy and Simpson index both measure diversity, whilst their evenness counterparts indicate how evenly distributed the labels are in the dataset.
4. Person prominence assesses whether identified biases also manifest in image framing.

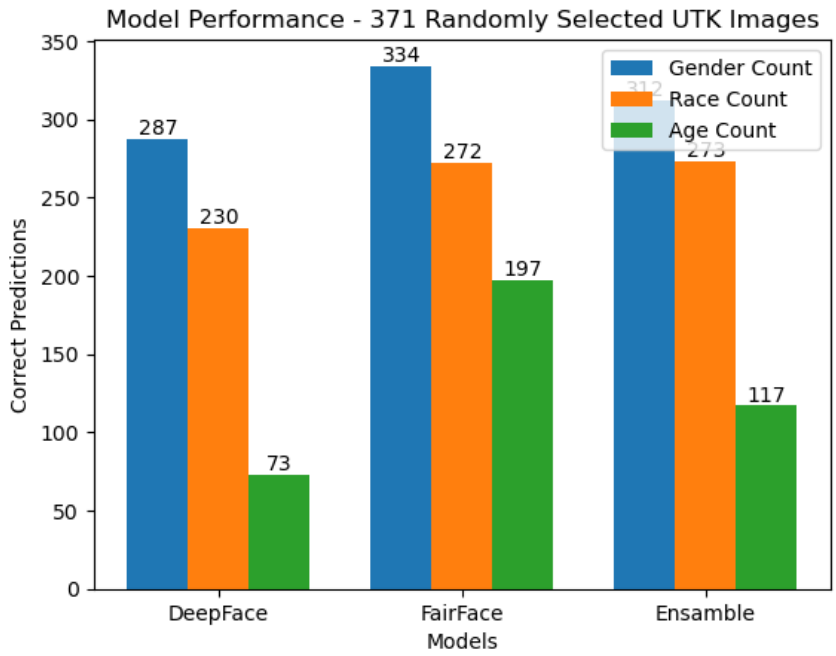


Figure 3.2 Model comparison on a subset of the UTK-Face dataset.

3.5 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, concluding with the metrics used to identify bias. The upcoming chapter will delve deeper and explain how these were fully implemented and carried out from a technical point of view.

4 Implementation

This chapter delves into the implemented program accessible via Appendix D, used for image annotation and metric extraction, detailing the primary steps whilst outlining any issues encountered throughout the implementation process.

4.1 Human Image Annotation

The LAION-400M images, distributed via google forms, were annotated by a diverse group of individuals, as outlined in Appendix A. The responses were primarily gathered 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. Initially, a weighted approach was to be used; however, seeing as therr was no clear distinction between respondent weighting, given that none of the respondents were experts, this could not be done. In addition to processing these labels, the Fleiss' Kappa [59] scale was used to identify the level of agreement between the respondents. This outlined how therr 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, as denoted by Table 4.1. These results are crucial as they support the use of the majority vote approach, seeing as agreeableness between annotators was high except amongst the age labels. Furthermore, it supports the usage for these labels to serve as baseline for identifying innate model bias, as carried out in Section 5.2. Despite the lack of agreement on age labels, it did not impede the identification of inherent age bias in the models, as the bias identified was consistent throughout all images.

Table 4.1 Average Fleiss Kappa values for the human annotated doctor and nurse subsets.

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

4.2 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 passed 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 centre is calculated. These metrics are crucial for determining person prominence. Finally, the cropped images alongside the aforementioned metrics, are returned. The flowchart for the **YOLODetectPerson** function can be seen in Figure 4.2.

Once the set of cropped images are extracted, they are passed 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. Given a list of images, these functions, extract the faces of the individuals depicted via the Multi-Task Cascaded Convolutional Network (MTCNN) model. This model was chosen due to its reliable output when tested with the DeepFace model and later utilised with the FairFace pipeline for consistent results. Furthermore, the face predictions are passed through the **FindObjWithLargestArea** function, which is used to return the index of the bounding box capturing the subjects face. This serves to filter out faces detected within the background of the image. Following this, the face is then cropped from the image and passed to the FairFace or DeepFace models, wherein they output the gender, race, and age predictions. The flowchart for the **FairFaceProcess** function can be seen in Figure 4.3, with the **DeepFaceProcess** function following a similar structure.

Finally, the image features alongside the results from the **YOLODetectPerson** function, are combined into a singular dictionary for ease of use and saved to disk. The structure of the dictionary is outlined in Figure 4.1, with each entry consisting of the cropped face image, the image labels and their confidence values. These results can then be loaded for processing via the **LoadMetricCSVFromFolder** function.

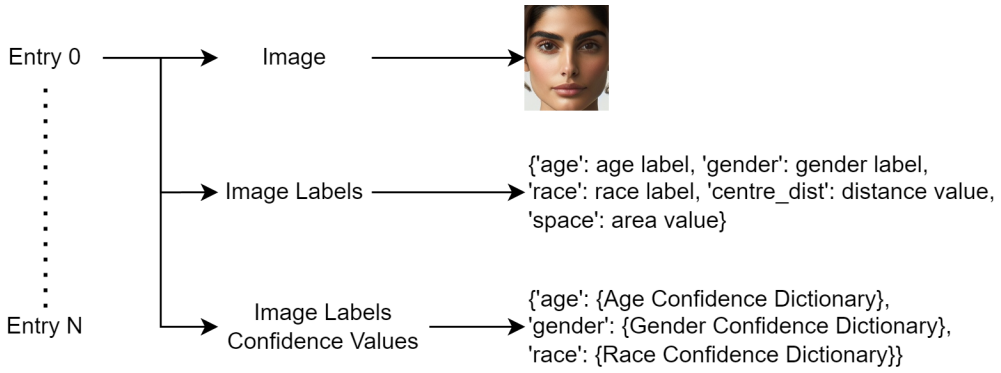


Figure 4.1 Metric dictionary structure.

4.3 Implemented Measures

4.3.1 Dulhanty and Wong measurement

The label count was implemented as outlined in [54] 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 measurements outlined below. This metric was implemented via the use of a simple tally which tracked the occurrence of each label within the image subsets.

4.3.2 Zhao et al. measurement

The correlation calculation outlined in [56] 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, and female are counted. These values are then divided by the number of labels associated with the attribute (in the case of gender, this is two as we only consider male, and female). Finally for each attribute label, we check if its value is greater than $\frac{1}{\text{no. of attribute labels}}$. If this is the case, then there is a positive correlation between profession and label, which hints at the presence of bias, with larger values indicating greater bias.

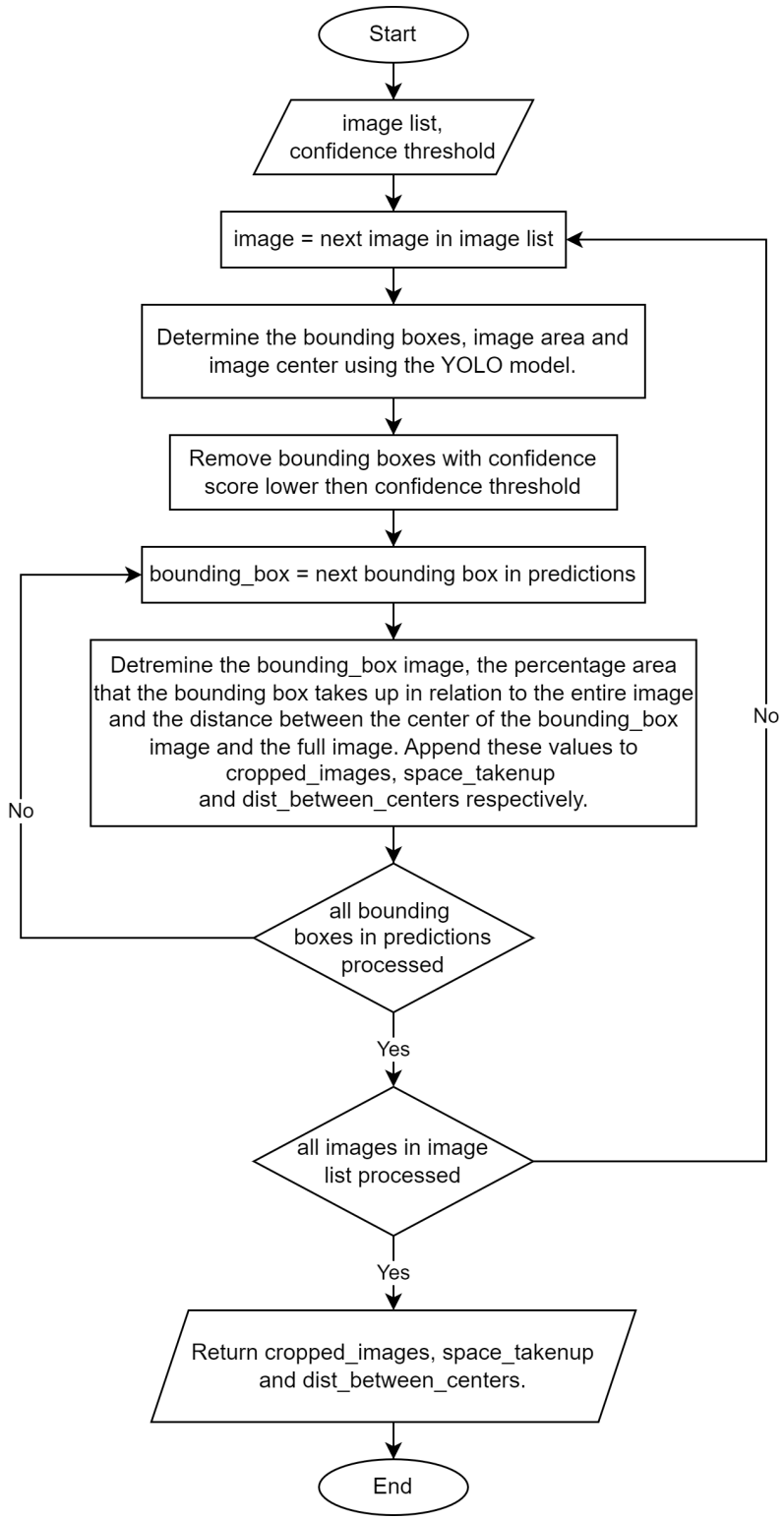


Figure 4.2 YOLODetectPerson flowchart

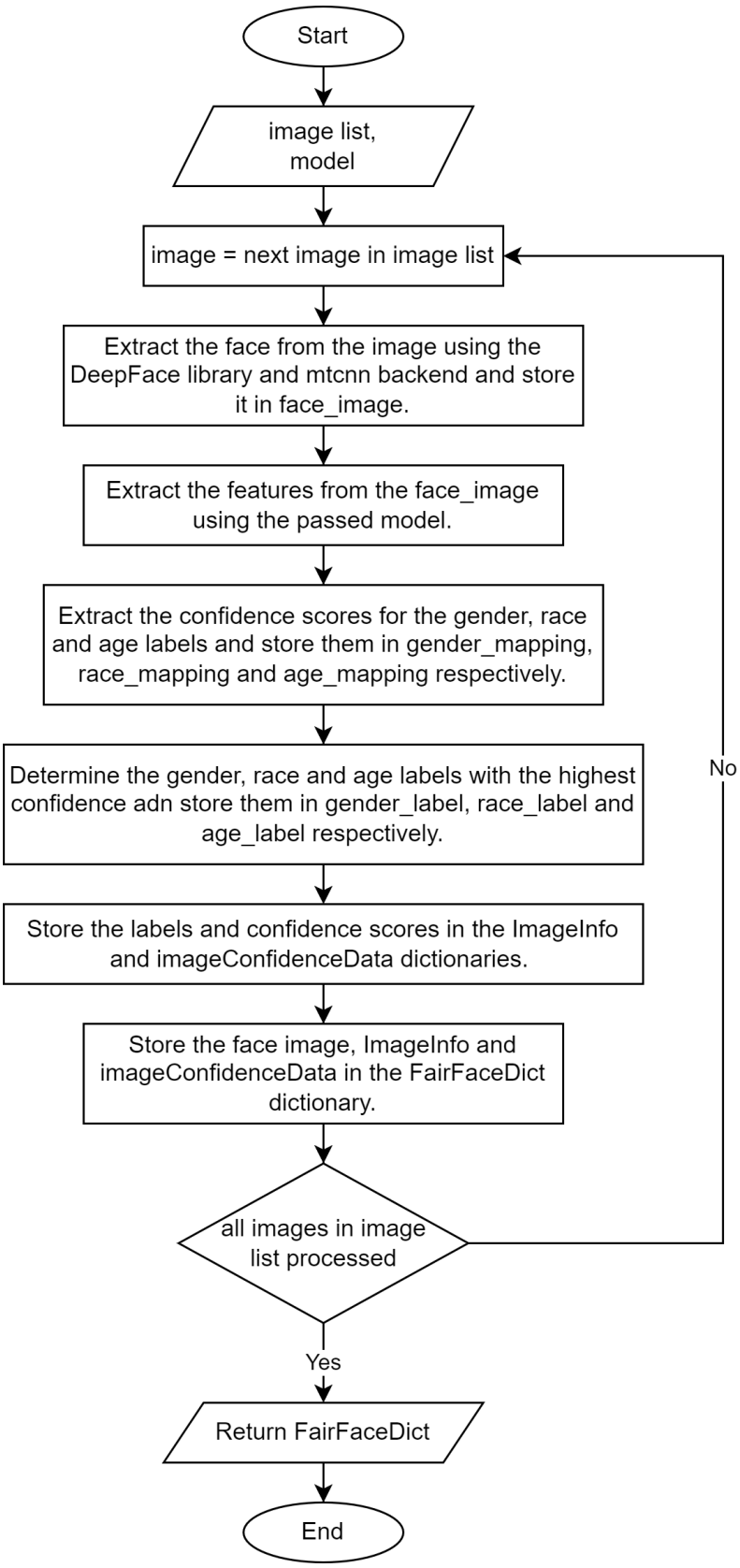


Figure 4.3 FairFaceProcess flowchart

4.3.3 Merler et al. measurements

The Shannon and Simpson calculations outlined in [57] were implemented to provide a definite measure on dataset diversity and distribution, as opposed to relying only on the label counts. These measures offer valuable insights into whether the identified bias diminishes the representation of other labels or merely skews the distribution in favor of the biased label but maintains a level of diversity. They were calculated by firstly determining the probability of each label occurring within the image set. Following this, the calculations as outlined in (2.2), (2.3), (2.4) and (2.5) 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.

4.3.4 REVISE measurement

The person prominence measure used in [13] 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 centre of the person (centre of bounding box), which allowed for the calculation of distance between image and bounding box centres and the percentage area taken up by the person in relation to the total image area.

4.4 Chapter Summary

This chapter delved into the technical implementation of the image annotation and metric calculation process. It outlined how the Google Forms were distributed and human annotation carried out, whilst emphasising the validity of the approach. Furthermore, the computer assisted annotation process was explained going over the entire pipeline along with the functions used. Finally, concluding with the implementation of the metrics, which were used to derive the conclusion presented in the following chapter.

5 Evaluation

This chapter outlines the reasoning behind the conclusions made and evaluates the effectiveness of any bias mitigation techniques present within the models. Starting with an identification of relevant real-world bias. This is then followed by the identification of bias within the LAION-400M dataset and annotation models. Concluding with a final discussion on the encountered biases, highlighting the most common ones, assessing the effectiveness of bias mitigation techniques and identifying the optimal model in terms of diversity and lack of bias.

5.1 Real World Bias

Identifying real-world bias was imperative in establishing a baseline against which the LAION-400M dataset and generative models' bias can be evaluated, given that the latter are most commonly 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 [60]. 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 [61]. Furthermore, the AAMC noted that **56.2%** of doctors depicted in their survey were white, indicating a notable racial imbalance in the medical profession [62]. Similarly, OECD reported that globally, only **34%** of doctors were aged 55 or older in 2019, although this demographic is gradually increasing [61].

Regarding the nurse demographics, a study by the World Health Organization [63] 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 [64] which revealed that globally **76.91%** of all nurses are female, with **81.62%** being younger than 55. Additionally, Rosseter highlights in [65] that the majority of the nursing population in America is predominantly white (**80%**) and female (**88.8%**).

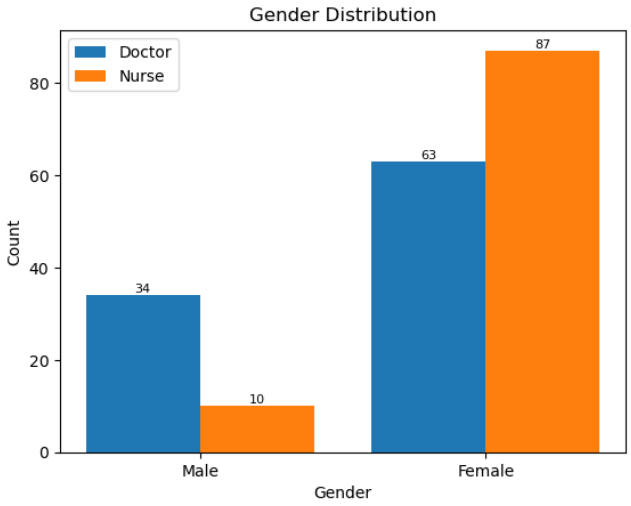
Overall, the data suggests that the majority of doctors are white males, younger than 55, with a similar demographic profile observed among nurses, except for the dominant gender being female.

5.2 Human and AI annotation comparison

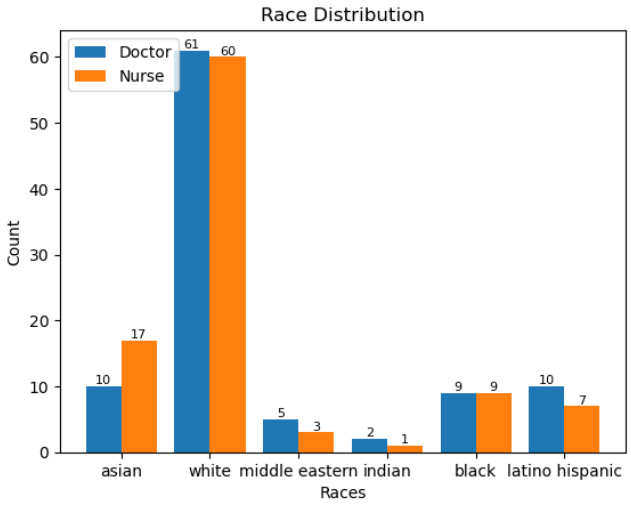
In line with Section 4.1 the LAION-400M dataset's doctor and nurse image subsets were human annotated, revealing inherent biases within the dataset. Notably, **64.95%** of doctors and **89.69%** of nurses were labelled as female. Moreover, the majority of depictions were classified as white, with **62.89%** 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 5.1.

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 female mislabelled as male. Conversely, DeepFace performed significantly worse with 62 (**31.96%**) misgendered images, all of them being female mislabelled as male. In terms of racial classification, FairFace misclassified 67 (**34.53%**) images 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, totalling 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 mislabelled as 20-29. DeepFace misclassified 89 (**45.88%**) images, 32 of which being the 20-29 age group mislabelled as 30-39 and an additional 26 being the 30-39 age group mislabelled as 20-29.

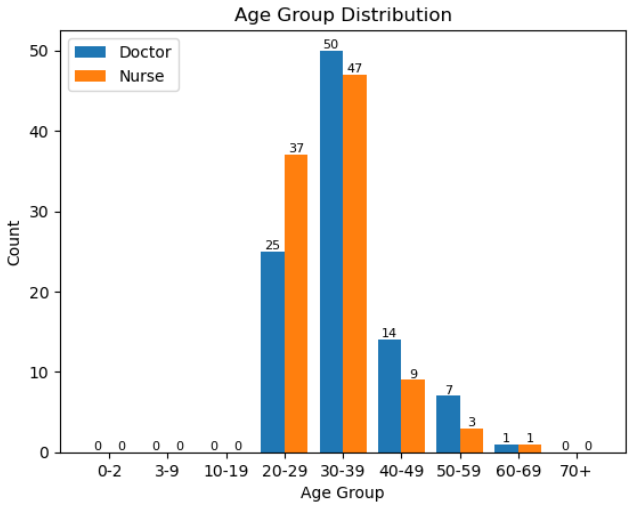
In accordance with these results, FairFace annotations were prioritised over those conducted via DeepFace, due to FairFace's minimal gender bias and predictable race and age biases relative to the human annotated baseline. Seeing as FairFace consistently mislabelled white and female individuals whilst presenting older individuals as younger, making these biases predictable and manageable when assessing generative model bias. Contrarily the DeepFace model depicted a strong consistent bias towards labelling images as male, with inconsistent race and age biases, leading to the prior decision. Furthermore, the FairFace model performed better when tested against the UTK-Face dataset as outlined in Figure 3.2.



(a) Human annotated dataset - Gender graph



(b) Human annotated dataset - Race graph



(c) Human annotated dataset - Age graph

Figure 5.1 Doctor and nurse demographics within human annotated subset.

5.3 Bias in LAION-400M Dataset

In accordance with Figures C.1, C.2, C.3 and compensating for the biases identified in Section 5.2, 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 C.2 further support the above claims, however the evenness values in Table C.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 Figures C.4, C.5, C.6, C.7 whilst excluding underrepresented labels such as 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 suggests the absence of framing bias primarily in relation to the distance from centre measure seeing as the distance was measured in pixels.

In comparing the doctor results with real life metrics outlined in Section 5.1 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 were white in comparison to the **56.2%** American statistic, furthermore **93.77%** - FairFace / **99.74%** - DeepFace of all depictions were younger than 55 as opposed to the **66%** real world metric. However, the real-world gender metric appears to be inverted with the majority of doctors depicted being female as opposed to the less than **50%** real world metric.

Observing the nurse metrics one can conclude that the gender distribution remains consistent, with female representation at **79.22%** - FairFace / **53.25%** - DeepFace when compared to real life metrics (**76.91%**). Contrarily, white representation was reduced to **59.74%** - FairFace / **64.16%** - DeepFace as opposed to the **80%** American statistic. Finally, **93.77%** - FairFace / **99.74%** - DeepFace of nurse depictions were younger than 55 as opposed to the **81.62%** real world metric.

5.4 Stable Diffusion result analysis

In accordance with Figures C.8, C.9, C.10 and the biases identified in Section 5.3, 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 nurse's 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 evenness and correlation as shown in Tables C.3 and C.4 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 Figures C.11, C.12, C.13, C.14 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 5.3 it can be concluded that the Stable Diffusion model contains increased gender bias more in line with real-world data as the majority of doctors **88.31%** - FairFace / **89.61%** - DeepFace were male whilst the majority of nurses **96.62%** - FairFace / **90.65%** - DeepFace were female. Contrarily, racial bias appears to be mostly in line with that identified within the LAION-400M dataset however, there appears to be a reduction in Asian representation accompanied with an increase in white representation as is clearly visible in Figure 5.3. Furthermore, age bias appeared to remain consistent with the LAION-400M dataset having minimal to no depictions of individuals older than 55 across both image sets.

5.5 Dall-E result analysis

In accordance with Figures C.15, C.16, C.17, Dall-E depicted a balanced gender distribution except for the doctor subset wherein the majority (**72.21%** - FairFace / **69.09%** - DeepFace) of depictions were female. Furthermore, the majority of races were sufficiently represented across all three image subsets with Asian and Indian being the most prominent overall. Contrarily, all image subsets fell 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 5.2, the 10-19 age range was likely a misclassification on behalf of the FairFace model. Considering the evenness and correlation results depicted in Tables C.5 and C.6, respectively, they supported the

claims made prior of a balanced gender and race representation. The prominence metrics in Figures C.18, C.19, C.20, C.21 denoted that both genders were relatively equally prominent; this also applied to the different race depictions, however, there were instances in which some races were marginally more prominent than others.

Given that Dall-E's training dataset was 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 was clear that the Dall-E model was primarily designed with diversity in mind, given that the percentage of male doctors was **27.79%** - FairFace / **30.91%** - DeepFace as opposed to the greater than **50%** global metric, whereas the depiction of female nurses was **56.1%** - FairFace / **53.51%** - DeepFace as opposed to the **76.91%** globally. Furthermore, the model presented a diverse distribution of races as opposed to real-world metrics wherein the majority of depictions were white (**56.2%** - doctor / **80%** - nurse), however, there appeared 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 was evident that the Dall-E model perpetuated the innate age bias present globally, as all of the models depictions were of individuals younger than 55. Given these comparisons, it was clear that such a model was far less biased than its counterparts; however, it was important to note that such results were not achieved solely via the model. Rather, Dall-E utilised a separate prompt refining model which could convert a simple prompt such as "*a picture of a doctor facing forward*" into "*Visualise 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 [25]. This begged the question of whether the reduction in bias and increased diversity was due to how the model was trained and constructed or whether it was solely the result of the prompt refining model; either way, the Dall-E model presented itself as a relatively fair and unbiased model irrelevant of the means by which this was achieved.

5.6 Midjourney result analysis

In accordance with Figures C.22, C.23, C.24, 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 resulted in a balanced joint subset for the same reason as discussed in Section 5.4. The races depicted predominately include white, black and Latino Hispanic across all three image subsets. Furthermore, although their appear to be depictions of older individuals within all three image subsets the dominant age range remained consistent with previous observations wherein 20-29 and 30-39 age ranges were dominant. The correlation metrics in Table C.8 backup the claims made above whilst the evenness results in Table C.7 denote a relatively uneven race split with the remaining results enforcing the claims presented prior. In relation to the prominence metrics in Figures C.25, C.26, C.27, C.28 all genders and races were depicted in a relatively equal degree of prominence save for minor cases with underrepresented races, such as Indian within the DeepFace annotations.

In comparing the Midjourney 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, with the doctor subset having **34%** - FairFace / **40%** - DeepFace as opposed to **56.2%** American metric and nurses having **45%** - FairFace / **57.92%** - DeepFace as opposed to the **80%** observed in America. In relation to the age demographics, it appears that although there 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.

5.7 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:

5.7.1 LAION-400M Dataset

The LAION-400M dataset displayed inverted gender bias in relation to the doctor subset, however no gender bias within the nurse subset. Furthermore, the LAION-400M dataset presented increased racial bias towards white individuals within

the doctor subset whilst presenting an inverted racial bias within the nurse subset seeing as their was a drastic reduction in white representation. Contrarily, both image subsets presented severe age bias with almost all depictions being under 55 years old. Additionally, the dataset depicted a degree of racial prominence bias opting to depict certain races in a more prominent light, however gender prominence bias was absent.

5.7.2 Stable Diffusion

The Stable Diffusion model displayed inherent gender bias, skewed towards generating male doctors and female nurses. Furthermore, it presented a slight overall increase in racial bias in comparison to the LAION-400M dataset, opting to depict a greater degree of white individuals while reducing the rate of Asian depictions. Additionally, it exhibited age bias similar to the LAION-400M dataset, 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 other races.

5.7.3 Dall-E

The Dall-E model showcased a balanced gender distribution overall, which contrasted with the established real-world bias. Similarly, it exhibited a reduction in white representation in favour 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 a slight prominence bias, opting to depict certain races in a more prominent manner based on the image subset.

5.7.4 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 remained the most depicted race. Furthermore, the model tended to generate images of older individuals, particularly doctors, yet the majority still fell within the 20-39 age range, aligning with real-world data. Finally, Midjourney presented minimal prominence bias, opting to represent all depictions in an equal manner.

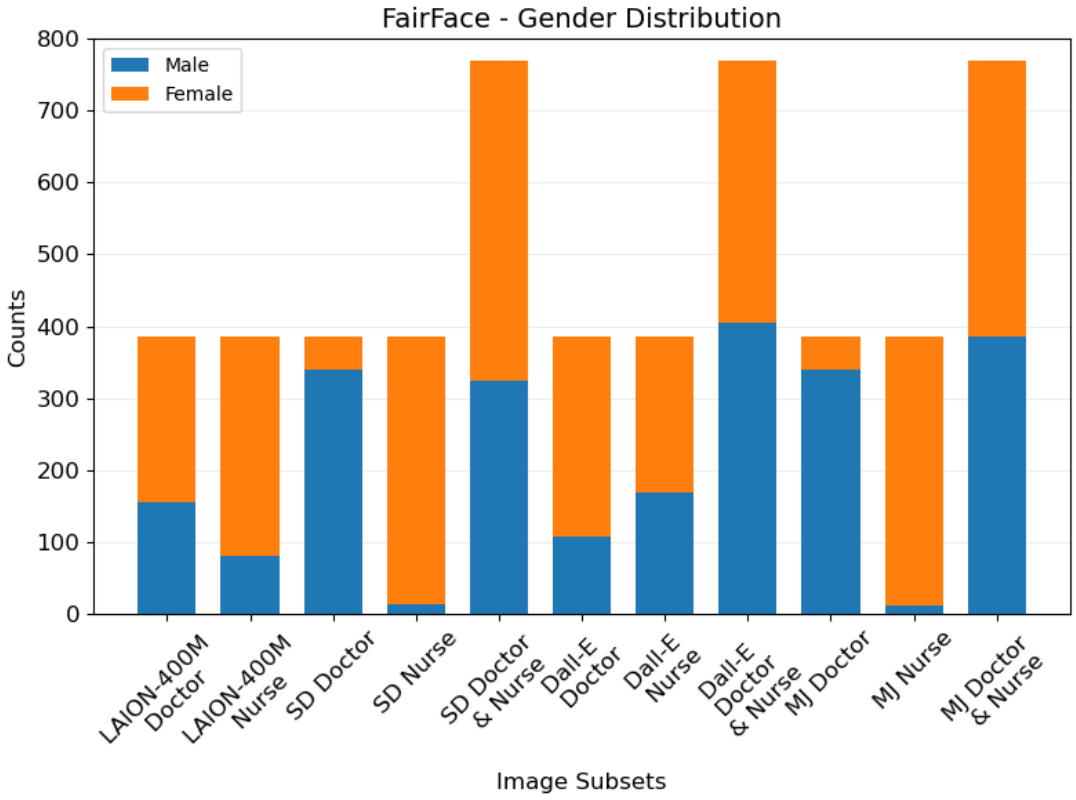
5.7.5 Final Evaluation

Among the three models, Dall-E presented the most balanced portrayal of doctors 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 biases. Given the goal of identifying the model most capable of generating a diverse set of images irrelevant of the prompt given it is clear that Dall-E is the only contender. Although the data used in training the Dall-E model influenced the images which it generates, given the results of the StableDiffusion and Midjourney models it is reasonable to conclude that the increase in diverse representations seen within the model can be largely attributed to its prompt altering model as opposed to its training dataset. These observations are in line with Figures 5.2, 5.3 and 5.4 as well as the individual demographic graphs presented in Appendix C.

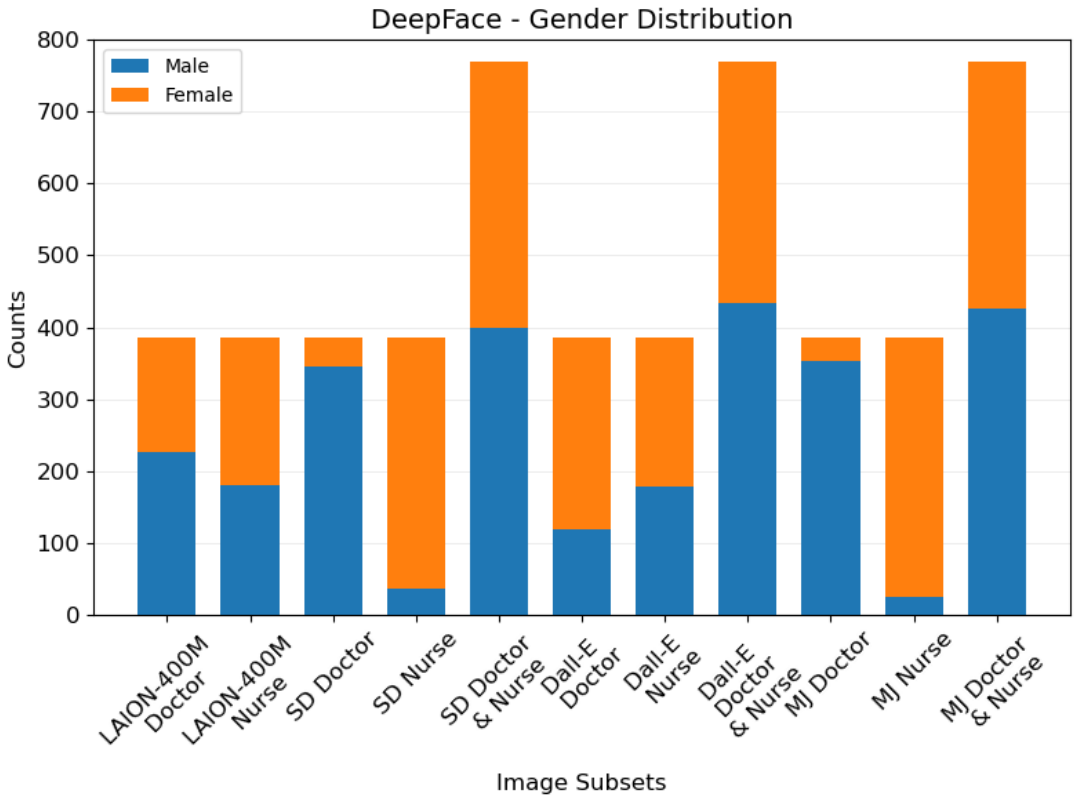
Overall, Dall-E emerges as the most promising model in terms of reduced bias and increased diversity, offering a more balanced representation of gender and race. However, further research into bias reduction techniques such as that implemented by DALL-E is required to address and mitigate biases in generative models effectively.

5.8 Chapter Summary

This chapter presented an in depth look at the biases present within the LAION-400M dataset along with the three generative models whilst presenting the varied list of metrics and graphs used to arrive to said conclusions. From the evaluation made the best model categorised by its lack of bias and diverse representation was identified to be DALL-E, furthermore its prompt altering model was presumed to be the driving factor behind its lack of bias.

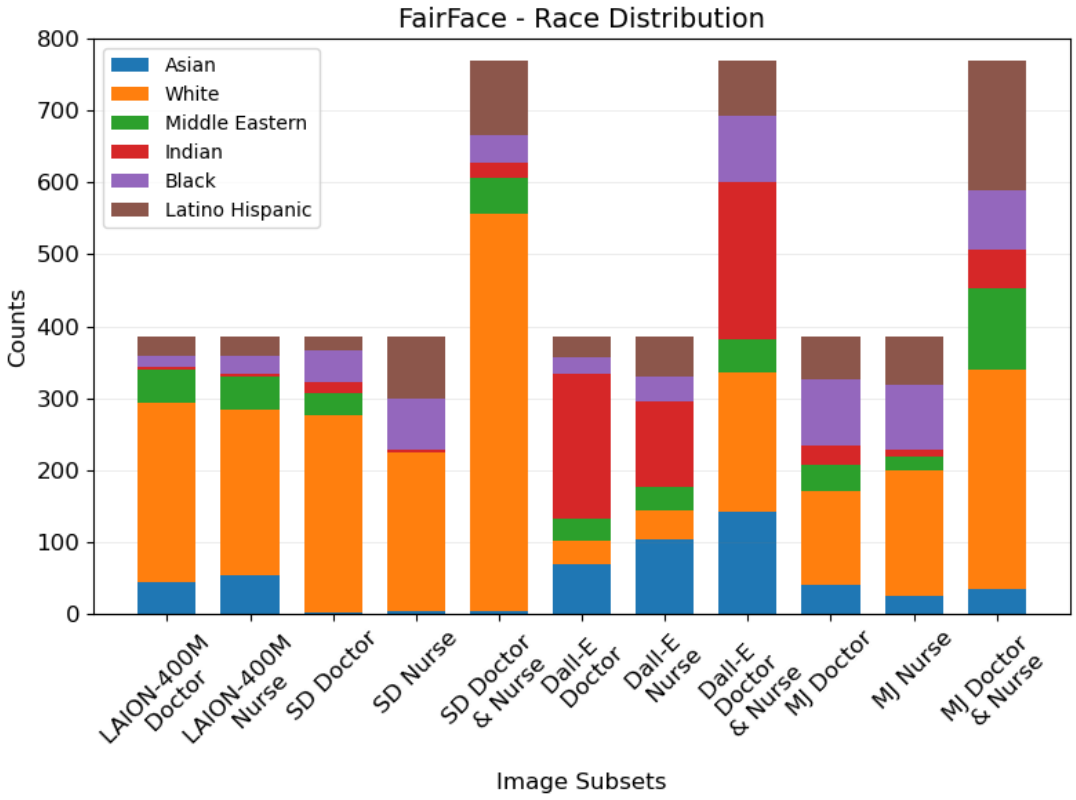


(a) FairFace gender graph

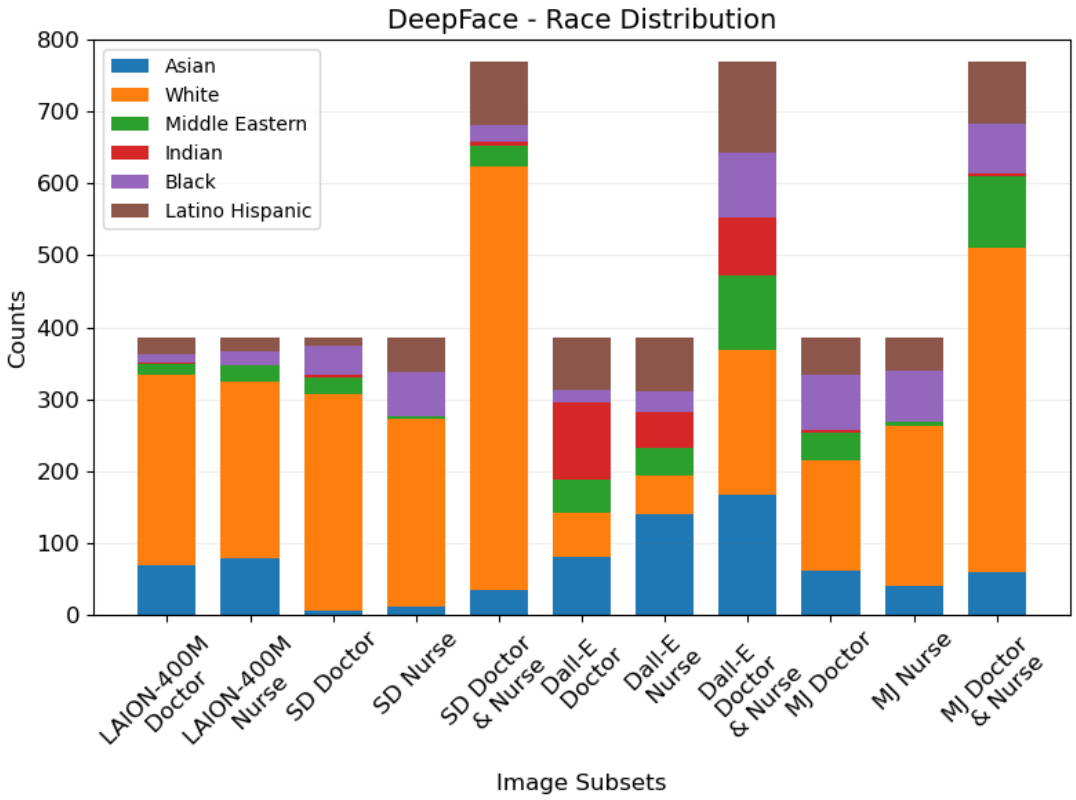


(b) DeepFace gender graph

Figure 5.2 Gender demographics across all datasets and models.

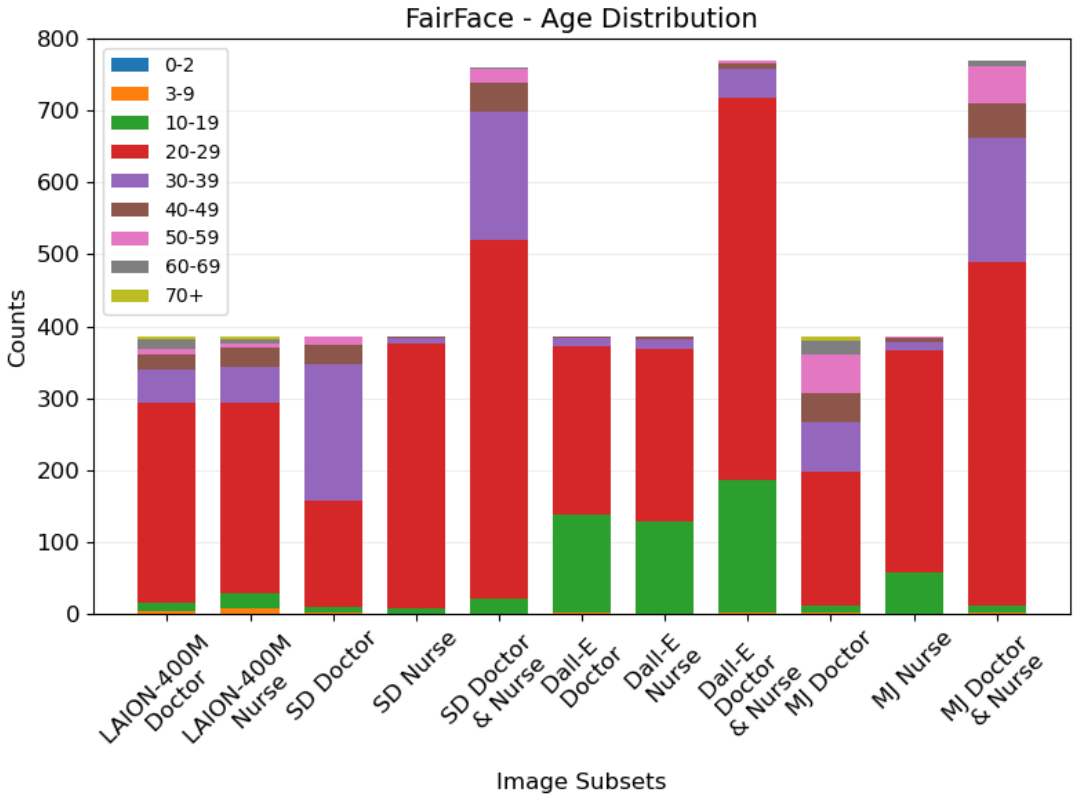


(a) FairFace race graph

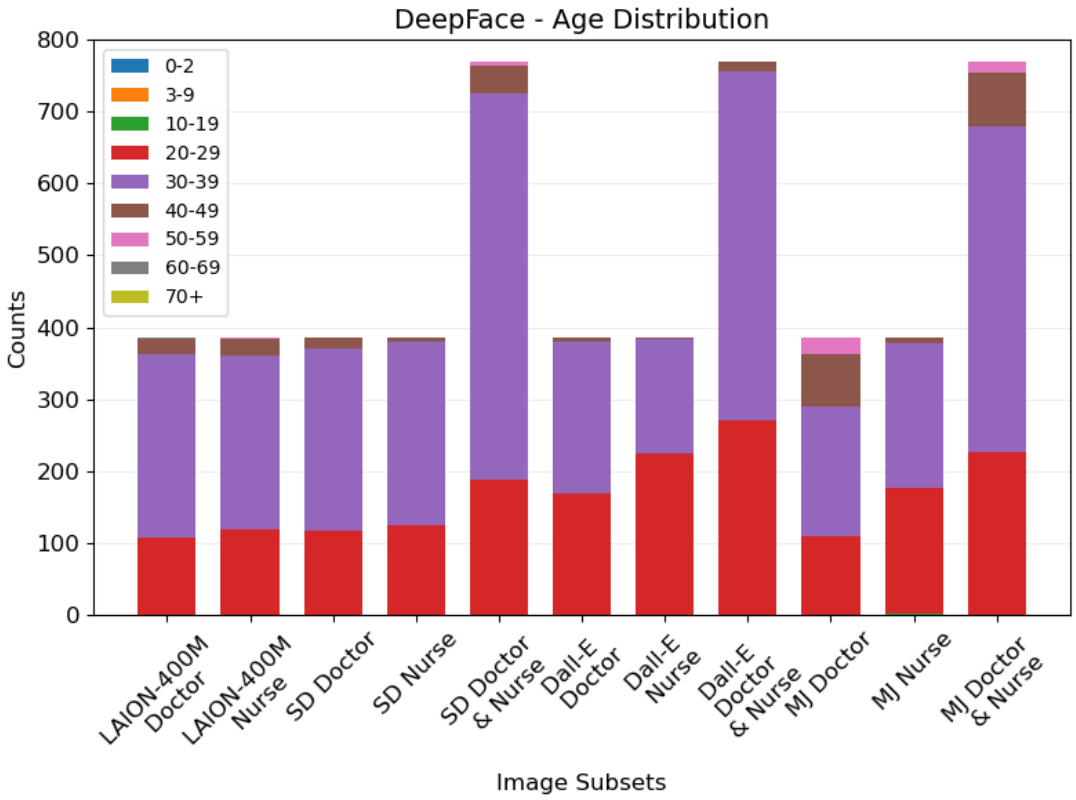


(b) DeepFace race graph

Figure 5.3 Race demographics across all datasets and models.



(a) FairFace age graph



(b) DeepFace age graph

Figure 5.4 Age demographics across all datasets and models.

6 Conclusion

6.1 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.

6.1.1 Revisiting the Aims and Objectives

The first objective was to investigate how each generative model processes their prompts and determine an optimal prompt structure, along with identifying the requirements needed to carry out valid human annotation. This was successfully achieved as outlined in Chapter 2 were it was determined that Stable Diffusion and Midjourney process their prompts in chunks whereas Dall-E automatically rewrites its prompts prior to processing. Furthermore, an optimal prompt structure in line with each models' recommendations was established: ***A picture of a [subject] facing forward***, with ***Disfigured*** and ***Art*** as negative prompts for clear, realistic depictions. Finally, the requirements for valid human annotation were established, these being geographically diverse annotators, standardised labels, and utilising annotation tools.

The second objective aimed to generate images of doctors and nurses using the Stable Diffusion, Dall-E and Midjourney models, whilst retrieving the associated images from the LAION-400M dataset. Additionally, these images were to be annotated using feature extraction models in relation to gender, race and age. Similarly human annotation was to be carried out on a subset of the LAION-400M dataset. This objective was successfully achieved as outlined in Chapters 3 and 5 were 385 images were generated for each label using each model, and retrieved from the LAION-400M dataset. Images depicting doctors and nurses simultaneously were not retrieved from the LAION-400M dataset due to a lack of search functionality rendering the task too time consuming. All images underwent annotation using both DeepFace and FairFace models. Additionally, a subset of the LAION-400M images were human-annotated in line with the valid human annotation requirements. However, the annotators were exclusively Maltese, as the distribution of the Google Forms was carried out via social media and thus, limited to Maltese participants.

The third objective was to determine the bias within the DeepFace and FairFace models via comparison with the human annotated LAION-400M subset. Additionally,

the metrics associated with the generated images were to be extracted thereby resulting in gender, race and age distributions, correlation, person prominence and Shannon/Simpson diversity and evenness measures. This objective was successfully achieved with FairFace exhibiting predictable and comparatively lower overall bias than DeepFace as outlined in Chapter 5. Furthermore, the image metrics were extracted via the DeepFace and FairFace model producing the values presented in Appendix C.

The final objective was to uncover 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. This objective was successfully achieved as outlined in Chapter 5 were the innate biases for the models and the LAION-400M dataset were established. Additionally, age bias was deemed the most common form of bias given that it was present amongst all image sets. Gender and race biases were mitigated or reduced amongst certain cases whilst prominence bias was absent or lacklustre amongst all cases. Finally, the least biased model was determined to be Dall-E seeing as it presented the least biased depiction of doctors and nurses. Additionally, the only bias mitigation technique identified was Dall-E's prompt rewriting, however its effectiveness could not be properly assessed as no other techniques were identified with which to carry out comparison. However, the lack of bias within the Dall-E model brings merit to the effectiveness of said technique.

6.1.2 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, observe the bias across the multiple image sets and arrive to a conclusion in that manner. Unfortunately, this was not plausible seeing as the generation of such a large number of images is quite costly. Additionally, along the same vein the research was limited to a small number of professions these being doctor and nurse. The addition of other professions 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 as due to time constraints the focus was limited to gender, race, age and prominence bias, in turn limiting the insights derived.

6.1.3 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 insight 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.

6.2 Final Remarks

This research paper delved into biases within generative models, identifying age bias as a common factor across all of the models whereas gender and racial biases although present amongst two of the models appear to be the current focus for bias mitigation. Furthermore, these biases appear to originate primarily from training datasets seeing as the biases identified were mostly in line with real-world metrics and training datasets tend to replicate real-world bias. This is also in line with the lack of gender and racial bias within the Dall-E model seeing as the primary discrepancy across the models was Dall-Es prompt altering capability. This suggests that curating non-biased training datasets is useful for mitigating bias however the focus ought to shift towards non-conventional means such as utilising large language models to alter prompt text thereby circumventing the innate model bias. Seeing as generative models have only recently entered the public eye, we are still in time to mitigate these biases so as to avoid perpetuating harmful gender and racial stereotypes which are likely to permeate into other fields such as advertising and media thus, negatively impacting society.

References

[1] Midjourney. “Midjourney.” Accessed on 29 October 2023. (2022), [Online]. Available: https://www.midjourney.com/home.

[2] DALL-E. “Dall-e 2.” Accessed 13 February 2024. (), [Online]. Available: https://openai.com/dall-e-2.

[3] Stability.ai. “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. Ntoutsi *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://doi.org/10.1002/widm.1356.

[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. “Machine bias.” Accessed on 29 October 2023, ProPublica. (May 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, Jun. 2019. DOI: 10.3386/w25943. [Online]. Available: http://www.nber.org/papers/w25943.

[11] B. Christian, “The dark side of embeddings,” in *The Alignment Problem: Machine Learning and Human Values*, W. W. Norton & Company, 2020, pp. 37–39.

[12] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, 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.

[13] 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.

[14] F. Ng. "Large ai training data set removed after study finds child abuse material." Accessed: 28 February 2024. (Dec. 2023), [Online]. Available: https://cointelegraph.com/news/laion-5b-ai-data-set-removed-child-sexual-abuse-material.

[15] B. Romain. "Laion-400m." Accessed on 9 March 2024. (Sep. 2021), [Online]. Available: https://www.kaggle.com/datasets/romainbeaumont/laion400m.

[16] 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, Sep. 2022, ISSN: 1573-1405. DOI: 10.1007/s11263-022-01653-1. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01653-1.

[17] O. AI. "Prompt engineering." Accessed on 28 February 2024. (), [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering.

[18] 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, ISSN: 1660-4601. DOI: 10.3390/ijerph19137640. [Online]. Available: https://www.mdpi.com/1660-4601/19/13/7640.

[19] L. A. Boge, C. Dos Santos, L. A. Moreno-Walton, L. X. Cubeddu, and D. A. Farcy, "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, pp. 961–964, 2019. DOI: 10.1089/jwh.2018.7571. [Online]. Available: https://doi.org/10.1089/jwh.2018.7571.

[20] 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.

[21] R. Beaumont. “Laion-5b: A new era of open large-scale multi-modal datasets.” (Mar. 2022), [Online]. Available: https://laion.ai/blog/laion-5b/.

[22] C. Crawl. “Frequently asked questions.” Accessed on 23 February 2024. (), [Online]. Available: https://commoncrawl.org/faq.

[23] AUTOMATIC1111, *Stable diffusion webui*, 2023. [Online]. Available: https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[24] Andrew. “Stable diffusion prompt: A definitive guide.” Accessed on 23 February 2024. (2023), [Online]. Available: https://stable-diffusion-art.com/prompt-guide/.

[25] OpenAI. “Image generation.” Accessed on 23 February 23 2024. (), [Online]. Available: https://platform.openai.com/docs/guides/images/introduction?context=node.

[26] PaulBellow. “Dalle3 prompt tips and tricks thread.” Published on OpenAI Forum. (Sep. 2023), [Online]. Available: https://community.openai.com/t/dalle3-prompt-tips-and-tricks-thread/498040.

[27] Midjourney. “Prompts.” Accessed on 28 February 2024. (), [Online]. Available: https://docs.midjourney.com/docs/prompts-2.

[28] J. Betker *et al.*, “Improving image generation with better captions,” 2023. [Online]. Available: https://cdn.openai.com/papers/dall-e-3.pdf.

[29] 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)*, 2022. [Online]. Available: https://arxiv.org/abs/2112.10752.

[30] L. Melchor. “Midjourney vs stable diffusion: Which tool should you use?” Accessed 20 February 2024. (Nov. 2023), [Online]. Available: https://www.pickfu.com/blog/midjourney-vs-stable-diffusion.

[31] K. Ahirwar. “A very short introduction to diffusion models.” Accessed on 28 February 2024. (Sep. 2023), [Online]. Available: https://kailashahirwar.medium.com/a-very-short-introduction-to-diffusion-models-a84235e4e9ae.

[32] A. W. Services. “What is stable diffusion?” Accessed 20 February 2024. (), [Online]. Available: https://aws.amazon.com/what-is/stable-diffusion/.

[33] 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.

[34] B. Benfold and I. Reid, “Guiding visual surveillance by tracking human attention,” Jan. 2009. DOI: 10.5244/C.23.14. [Online]. Available: https://www.researchgate.net/publication/221260001_Guiding_Visual_Surveillance_by_Tracking_Human_Attention.

[35] 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.

[36] 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.

[37] I. Marques, “Face recognition algorithms,” *Master’s thesis in Computer Science, Universidad Euskal Herriko*, vol. 1, Jun. 2010. [Online]. Available: https://www.ehu.eus/ccwintco/uploads/d/d2/PFC-IonMarqu%C3%A9s.pdf.

[38] H. Wang, Y. Wang, and Y. Cao, “Video-based face recognition: A survey,” *World Academy of Science, Engineering and Technology*, vol. 60, pp. 293–302, Dec. 2011.

[39] 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.

[40] R. Lakshmanan. “Facebook to shut down facial recognition system and delete billions of records.” Accessed on 24 February 2024. (Nov. 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. (Feb. 2019), [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. (Nov. 2019), [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. (Jan. 2018), [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. [Online]. Available: https://github.com/joojs/fairface?tab=readme-ov-file.

[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] 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.

[47] K. Xu and T. Matsuka, "Conscious observational behavior in recognizing landmarks in facial expressions," *PLoS ONE*, vol. 18, no. 10, pp. 1–21, Oct. 2023. DOI: 10.1371/journal.pone.0291735.

[48] C. Schumann and G. Olanubi. "Consensus and subjectivity of skin tone annotation for ml fairness." Accessed on 19 November 2023, Google Research. Google. (May 2023), [Online]. Available: https://blog.research.google/2023/05/consensus-and-subjectivity-of-skin-tone_15.html.

[49] 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.

[50] 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/.

[51] Tasq. "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/.

[52] A. Mehra. “Image annotation: Challenges & their solutions.” Accessed on 25 February 2024, Labellerr. (Nov. 2023), [Online]. Available: https://www.labellerr.com/blog/challenges-and-solutions-in-image-annotation/.

[53] H. Sajid. “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.

[54] 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), [Online]. Available: https://arxiv.org/abs/1905.01347.

[55] 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.

[56] 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.

[57] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith. “Diversity in faces.” arXiv: 1901.10436 [cs.CV]. (2019), [Online]. Available: https://arxiv.org/abs/1901.10436.

[58] Z. Zhang and Y. Song. “Utkface.” Accessed on 12 March 2024. (), [Online]. Available: https://susanqq.github.io/UTKFace/.

[59] D. Team. “Fleiss kappa.” Accessed on 28 March 2024. (2024), [Online]. Available: https://datatab.net/tutorial/fleiss-kappa.

[60] “Figure 19.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.

[61] "Doctors (by age, sex and category)." 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.

[62] "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.

[63] J. Yang and N. 30. "Distribution of nurses across regions by gender worldwide 2008 to 2018, by region." Accessed on 23 March 2024, Statista. (Nov. 2023), [Online]. Available: https://www.statista.com/statistics/1099804/distribution-of-nurses-across-regions-worldwide-by-gender/.

[64] 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, May 2023. DOI: 10.1186/s12912-023-01313-w. [Online]. Available: https://doi.org/10.1186/s12912-023-01313-w.

[65] R. Rosseter. "Nursing workforce fact sheet," American Association of Colleges of Nursing (AACN). (2024), [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 Figures A.1, A.2 and A.3. The structure of the forms is denoted in Figure A.4 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.

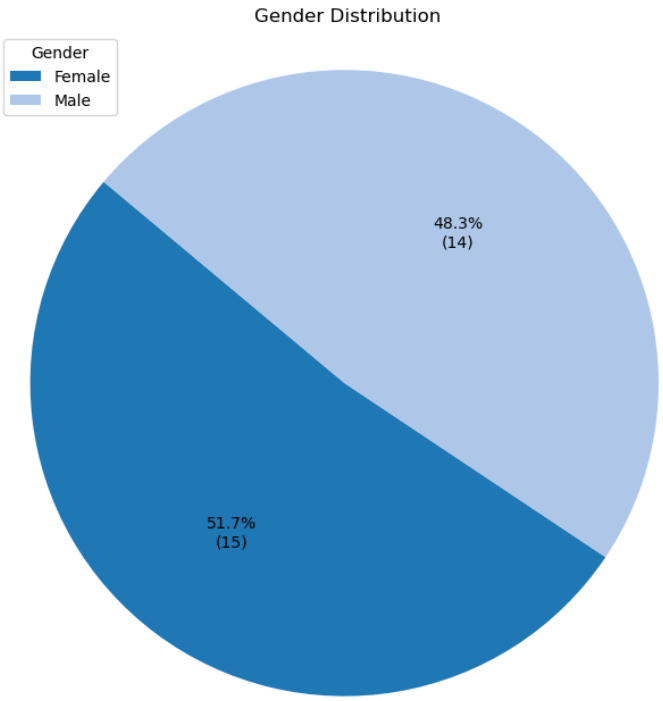


Figure A.1 Pie chart depicting the gender of the human annotators.

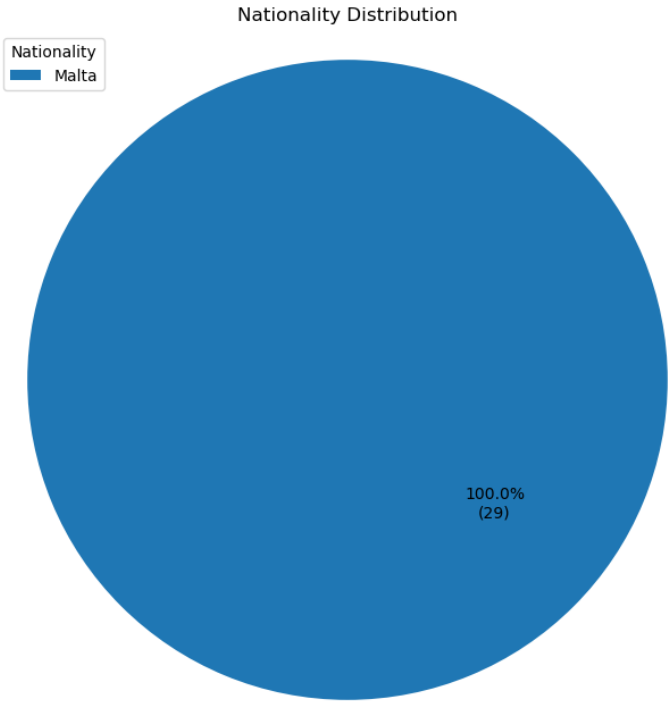


Figure A.2 Pie chart depicting the nationality of the human annotators.

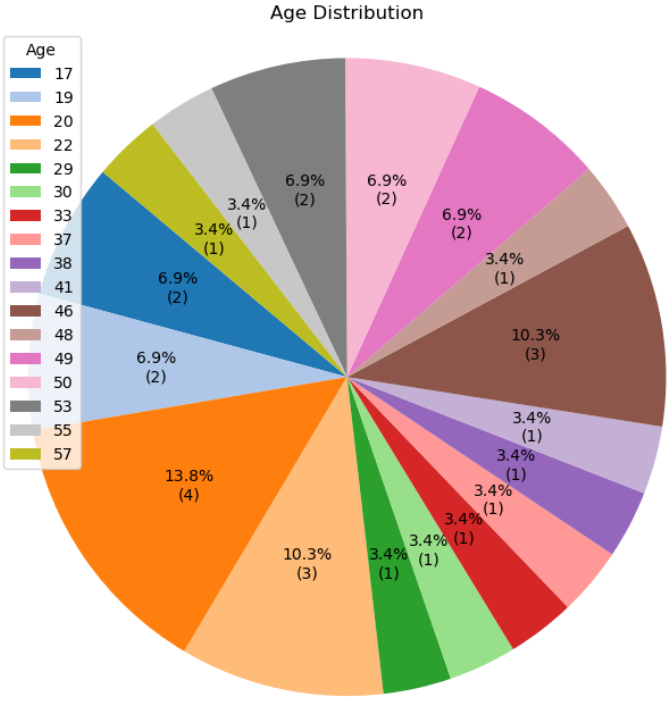


Figure A.3 Pie chart depicting the age of the human annotators.

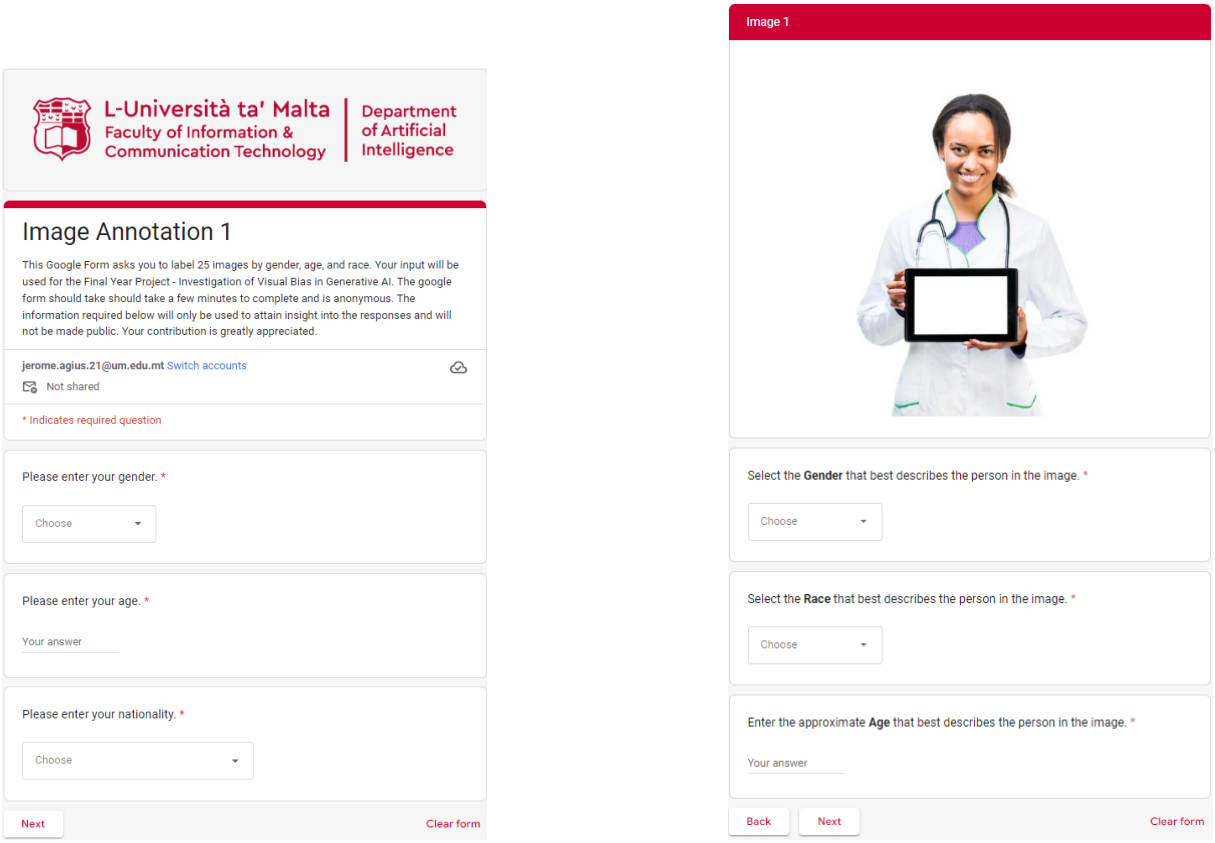


Figure A.4 The first (left) and second (right) pages of one of the distributed google forms.

Appendix B Sample images

A sample of the images retrieved from the LAION-400M dataset is depicted in Figure B.1 with Figures B.2, B.3, and B.4 depicting sample images generated from the StableDiffusion, Dall-E and Midjourney models respectively. The remaining images used throughout this research paper can be accessed via the GitHub repository outlined in D.

Figure B.1 LAION-400M Sample Images

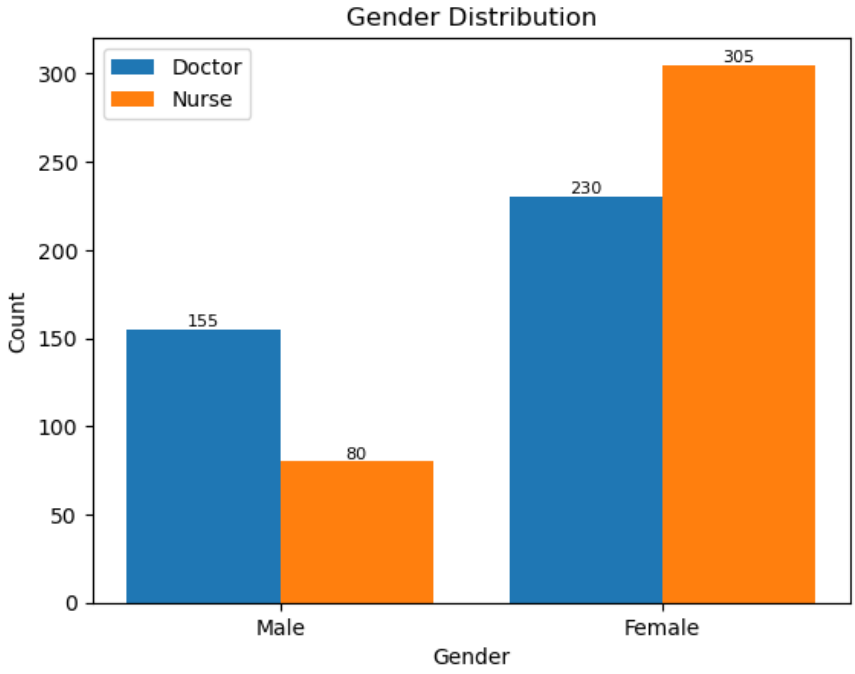
Figure B.2 Stable Diffusion Sample Images

Figure B.3 Dall-E Sample Images

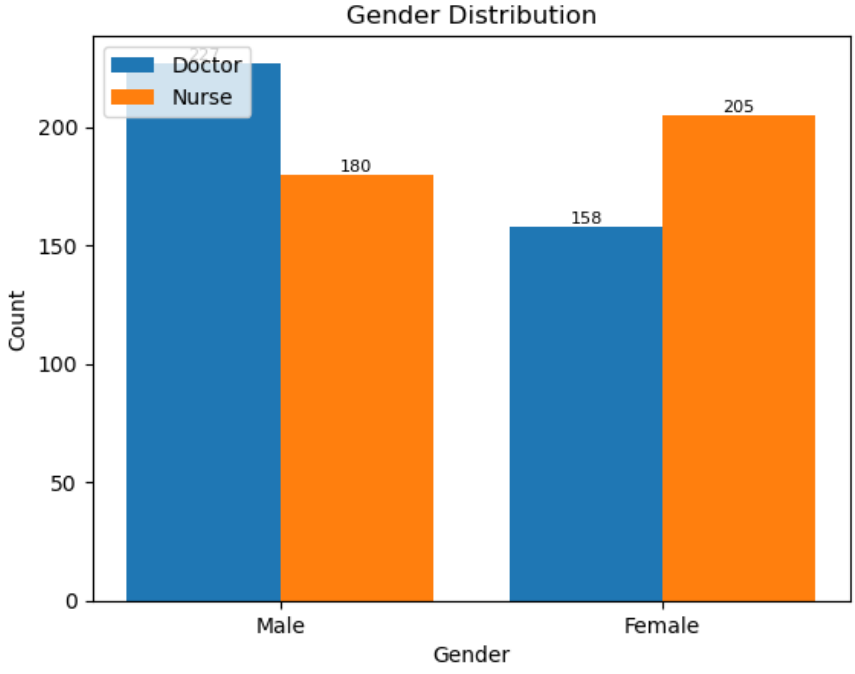
Figure B.4 Midjourney Sample Images

Appendix C Image Metrics

C.1 LAION-400M

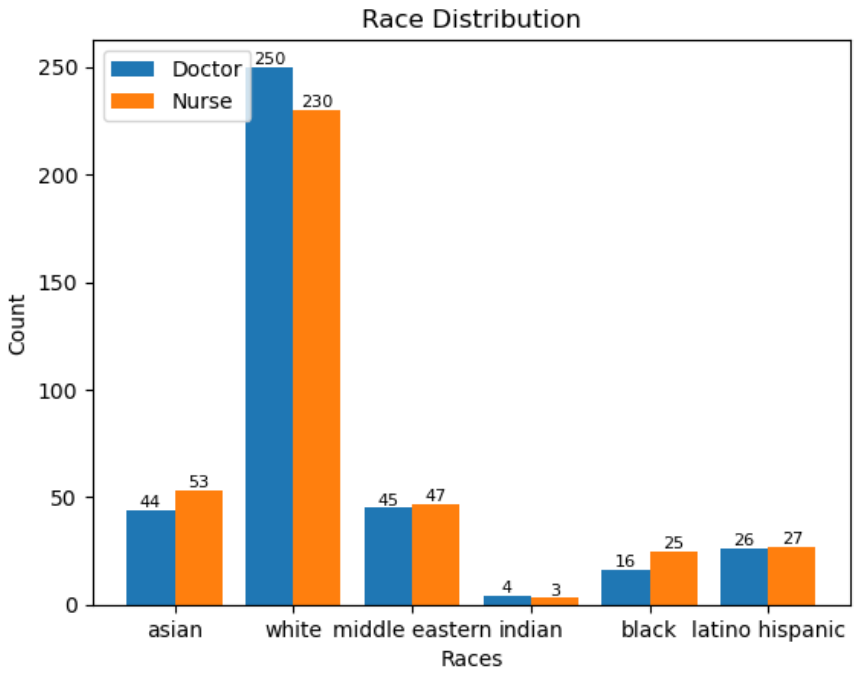


(a) FairFace Gender Graph

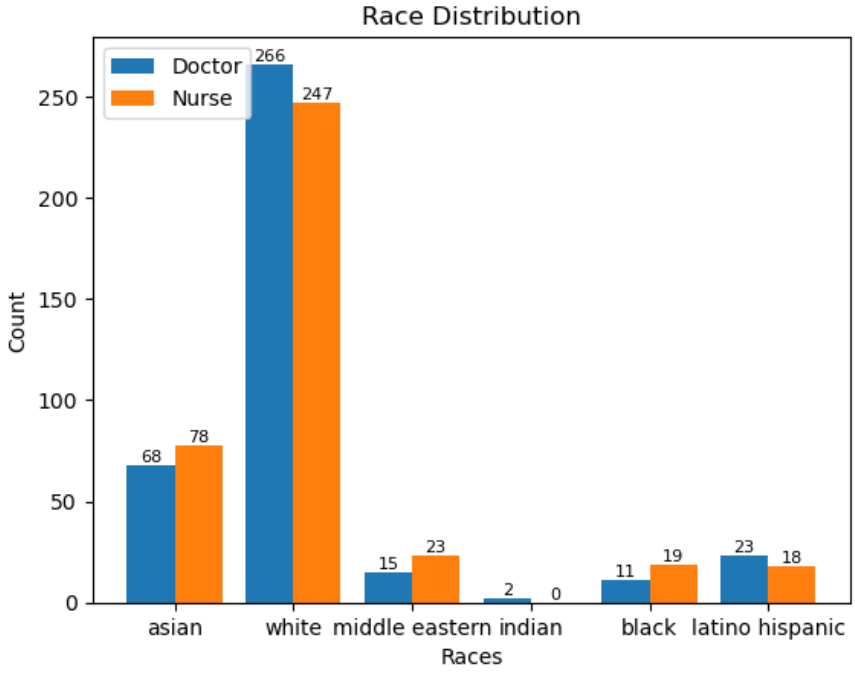


(b) DeepFace Gender Graph

Figure C.1 LAION-400M Gender Demographic Graphs

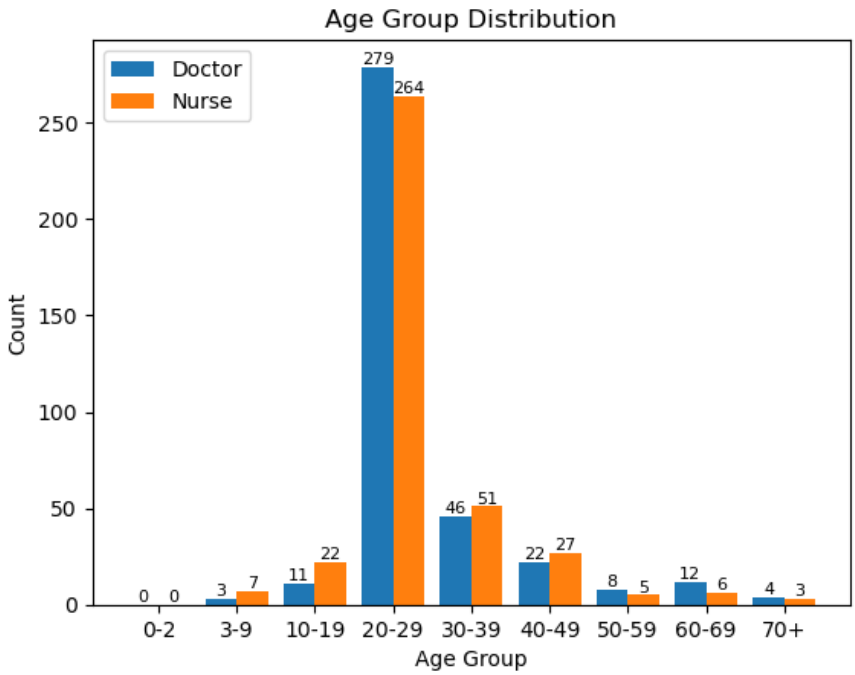


(a) FairFace Race Graph

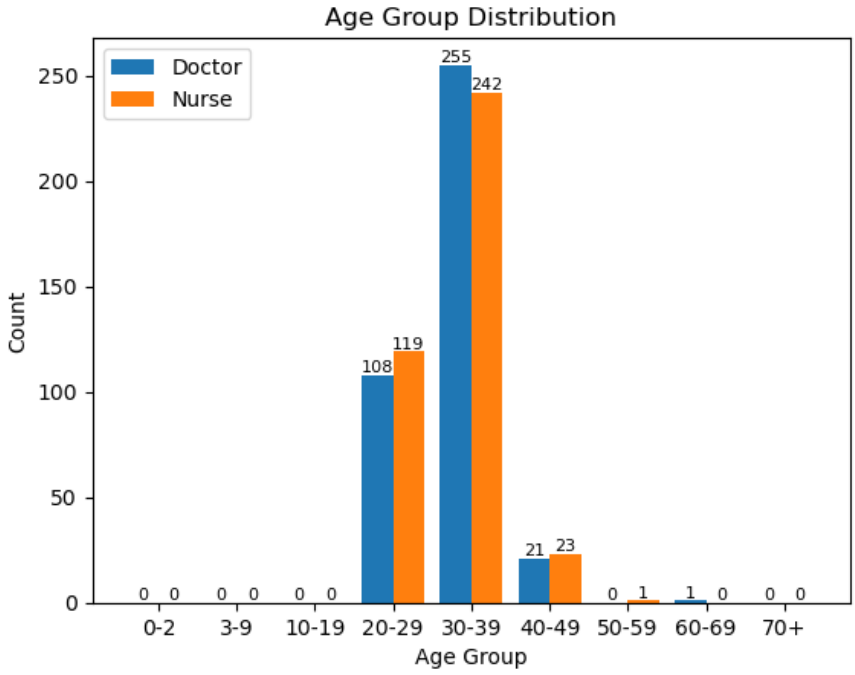


(b) DeepFace Race Graph

Figure C.2 LAION-400M Race Demographic Graphs

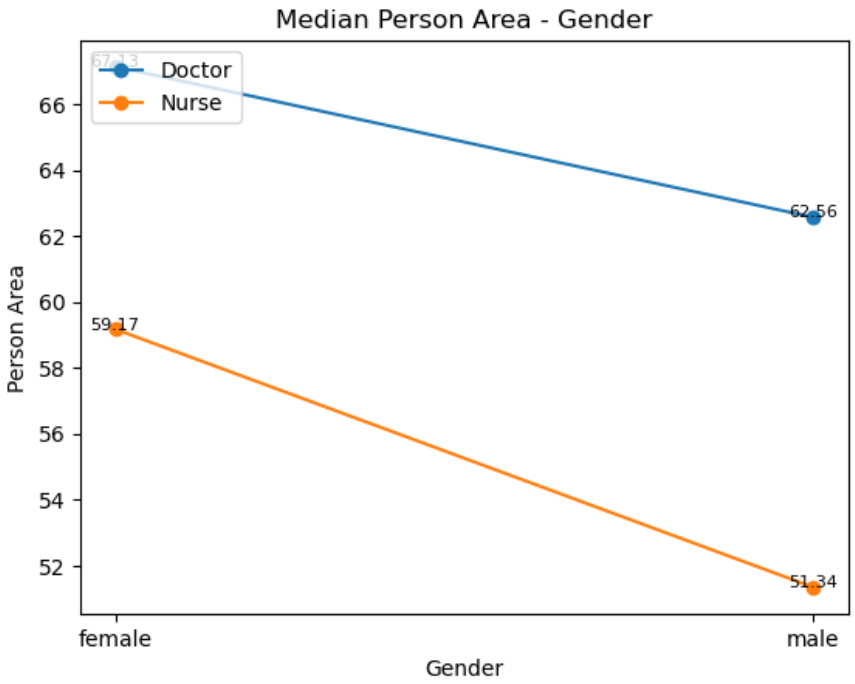


(a) FairFace Age Graph

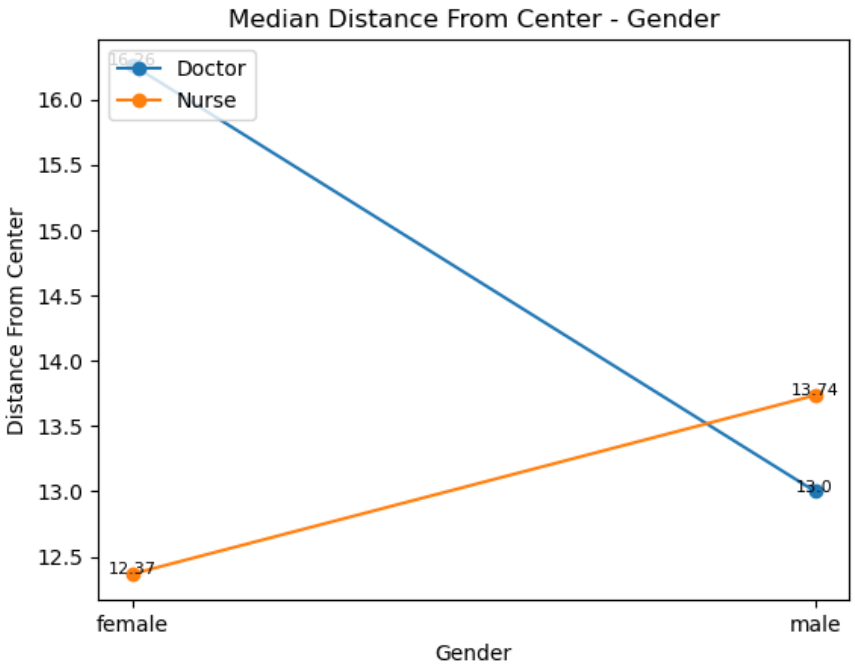


(b) DeepFace Age Graph

Figure C.3 LAION-400M Age Demographic Graphs

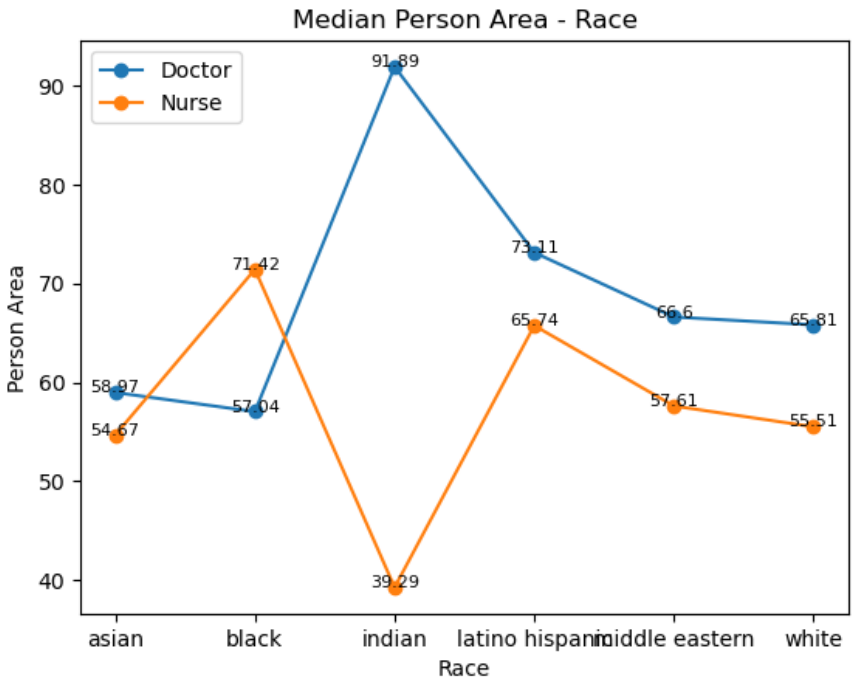


(a) FairFace Area-Gender Graph

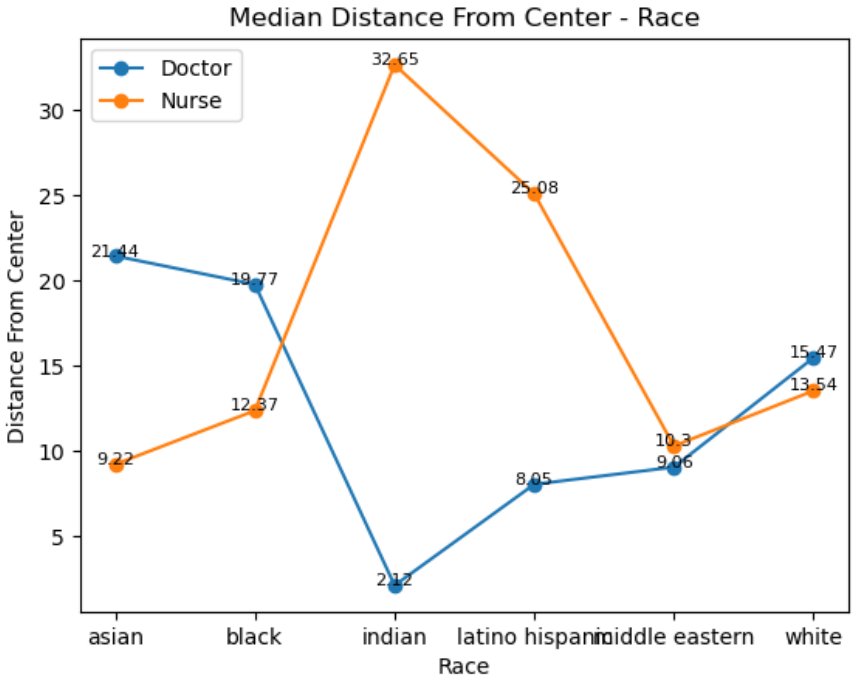


(b) FairFace Center-Gender Graph

Figure C.4 LAION-400M FairFace Prominence Graphs (1/2)

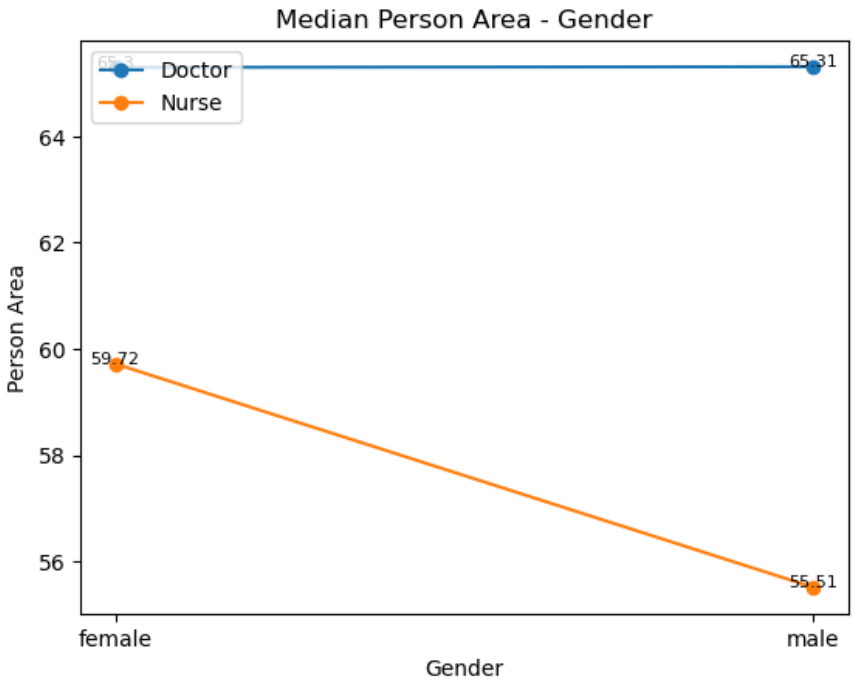


(a) FairFace Area-Race Graph

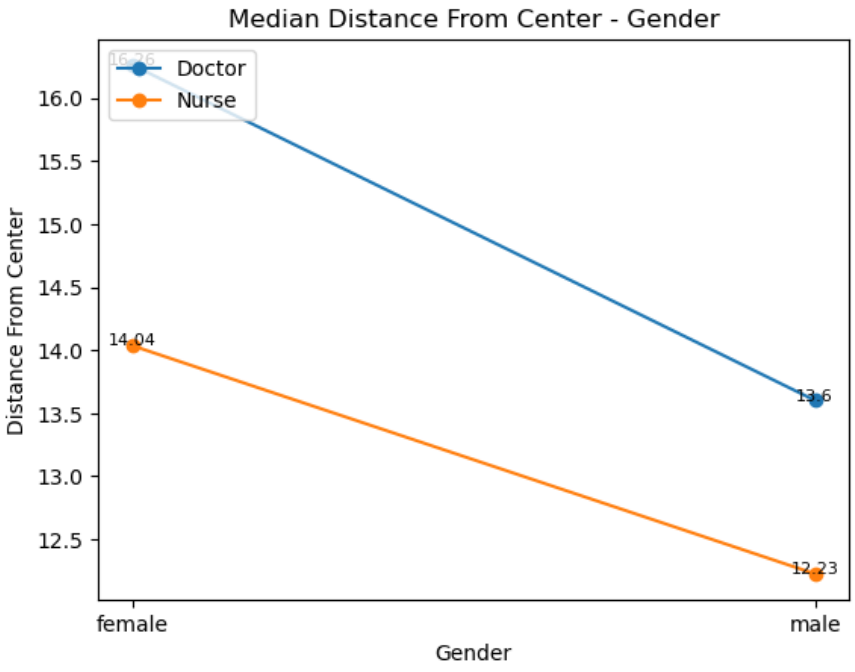


(b) FairFace Center-Race Graph

Figure C.5 LAION-400M FairFace Prominence Graphs (2/2)

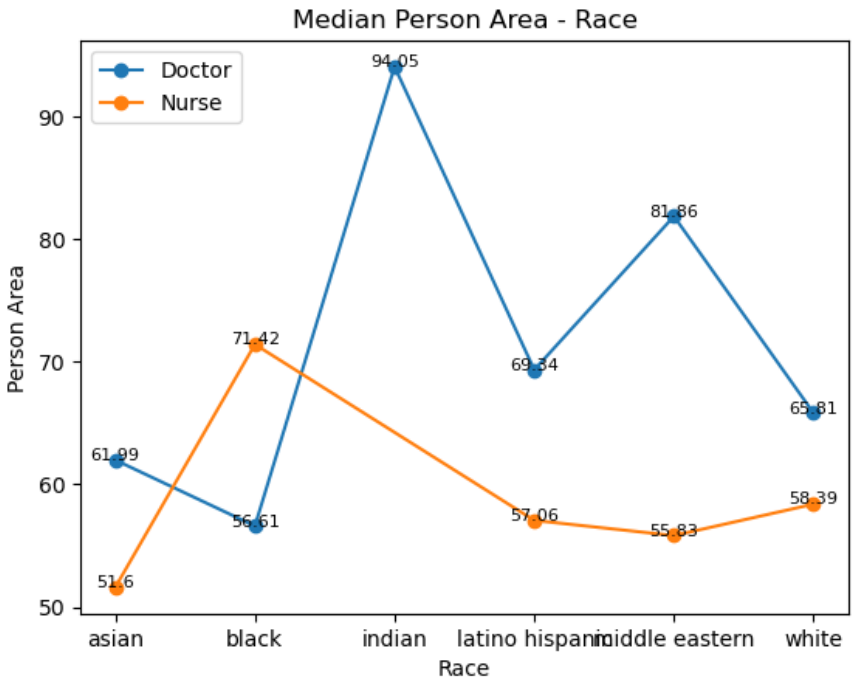


(a) DeepFace Area-Gender Graph

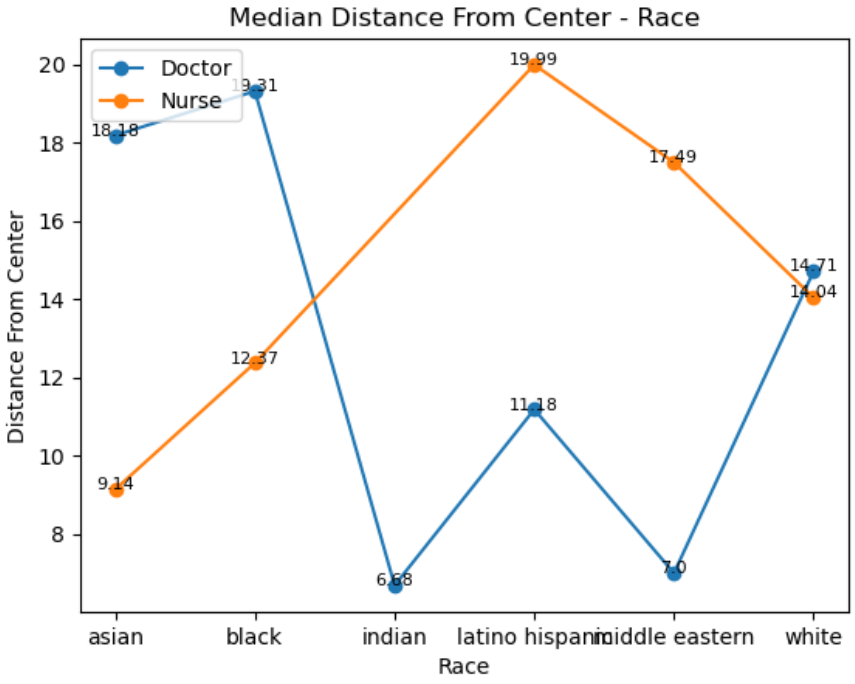


(b) DeepFace Center-Gender Graph

Figure C.6 LAION-400M DeepFace Prominence Graphs (1/2)



(a) DeepFace Area-Race Graph



(b) DeepFace Center-Race Graph

Figure C.7 LAION-400M DeepFace Prominence Graphs (2/2)

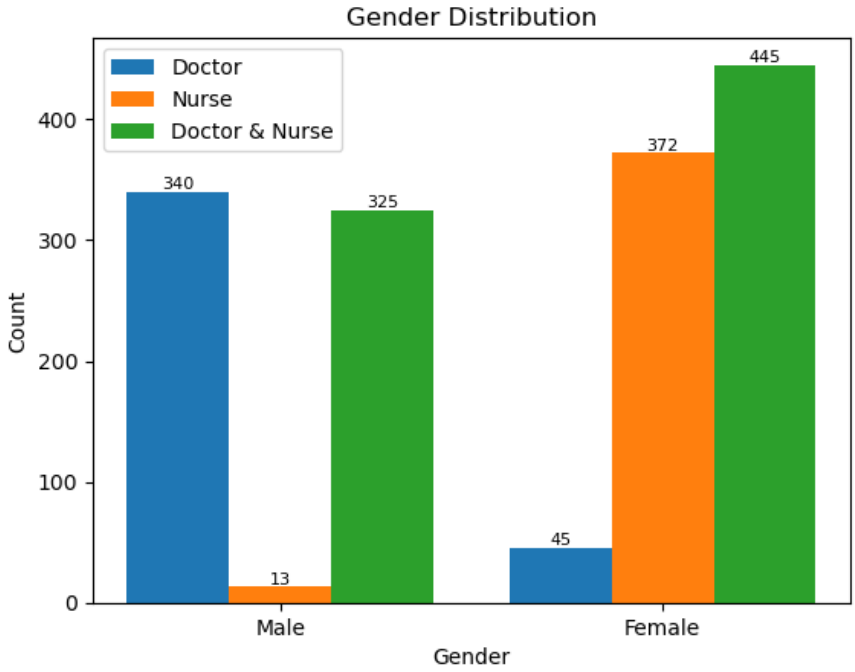
Table C.1 LAION-400M Shannon & Simpson Measurements

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

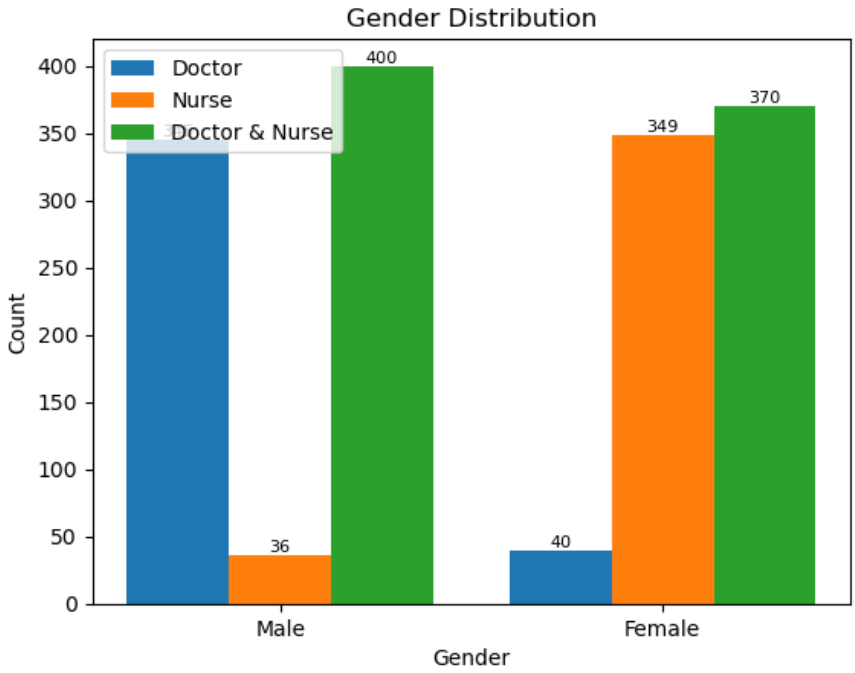
Table C.2 LAION-400M Positive Correlation Measurements

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

C.2 Stable Diffusion

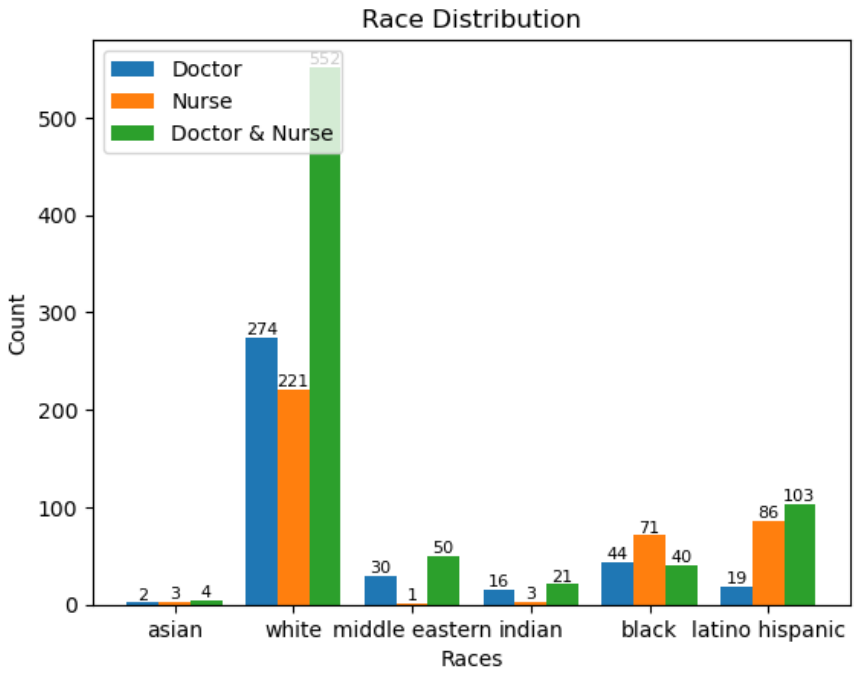


(a) FairFace Gender Graph

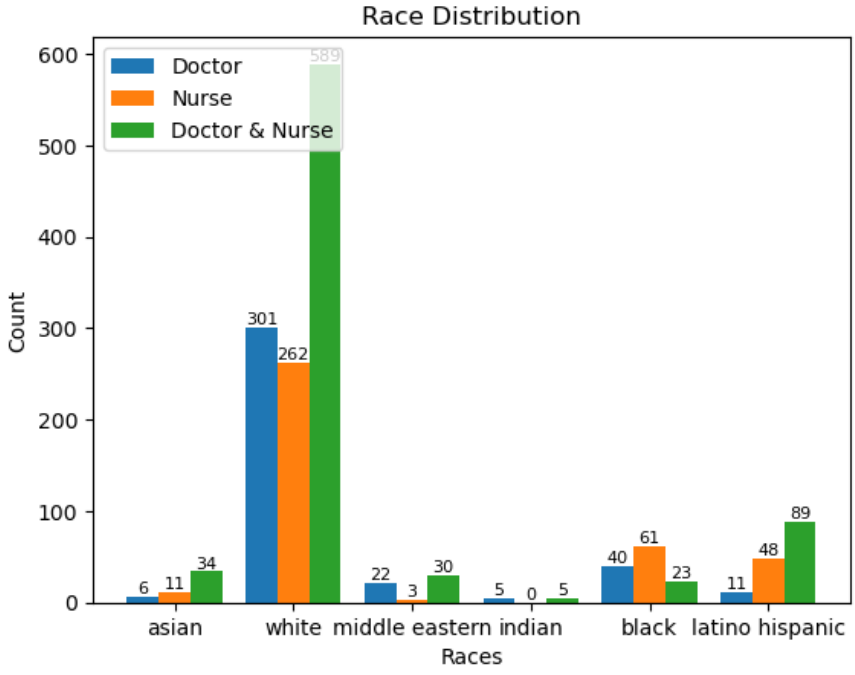


(b) DeepFace Gender Graph

Figure C.8 StableDiffusion Gender Demographic Graphs

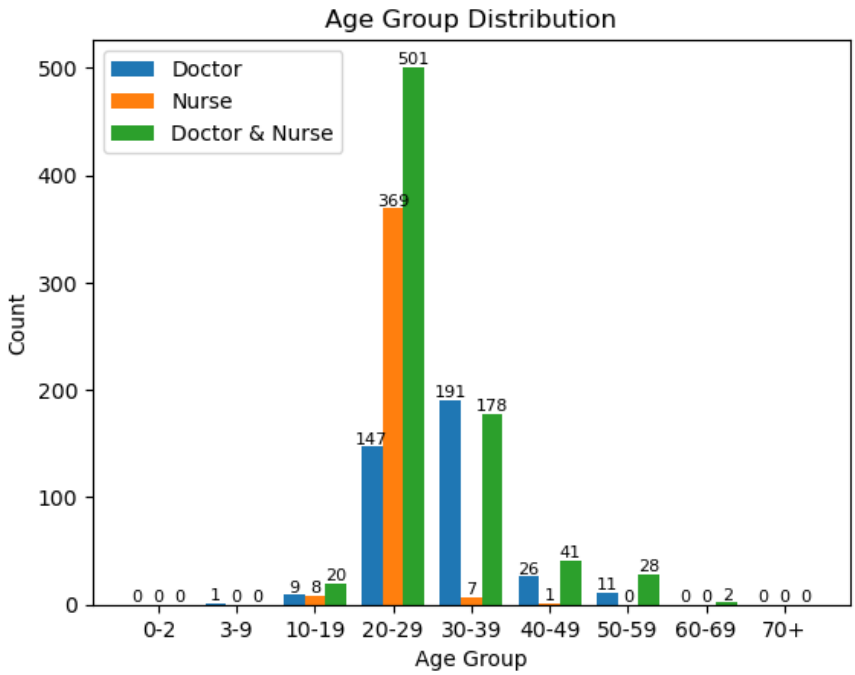


(a) FairFace Race Graph

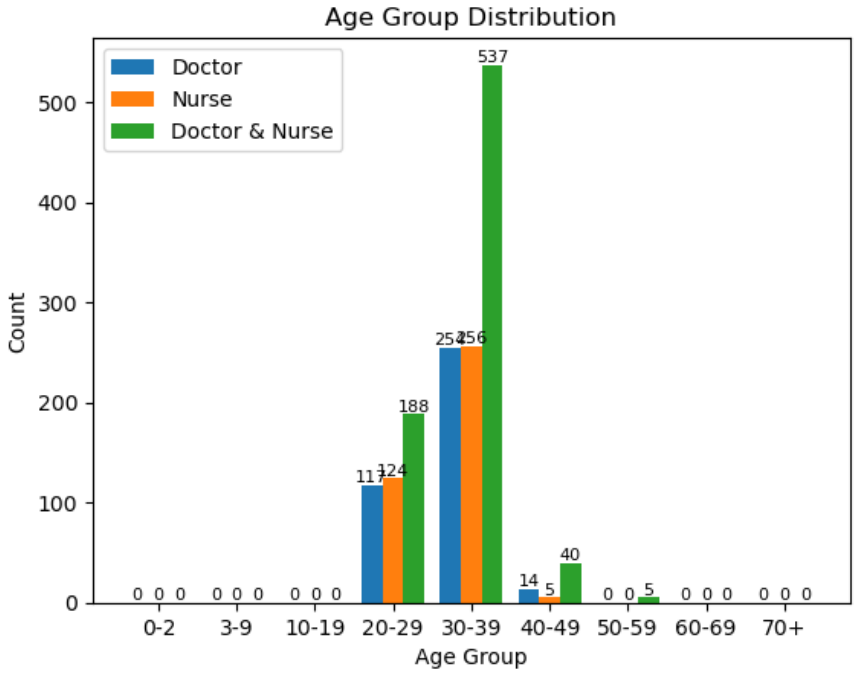


(b) DeepFace Race Graph

Figure C.9 StableDiffusion Race Demographic Graphs

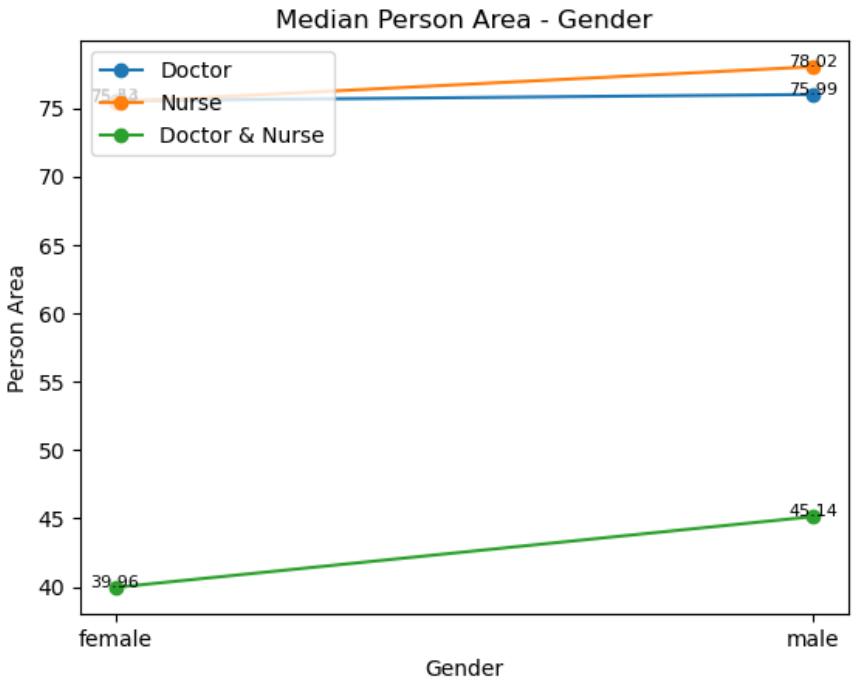


(a) FairFace Age Graph

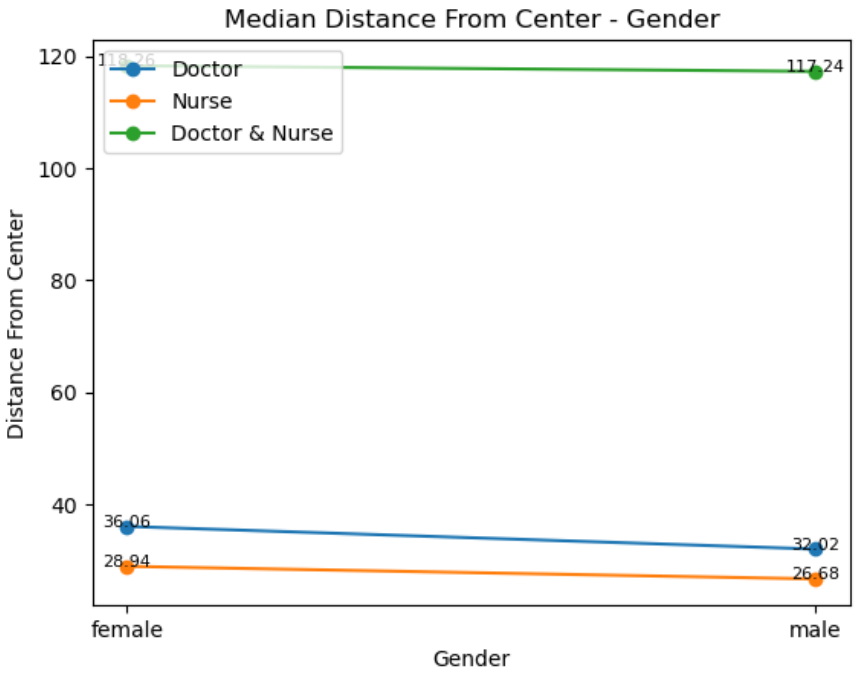


(b) DeepFace Age Graph

Figure C.10 StableDiffusion Age Demographic Graphs

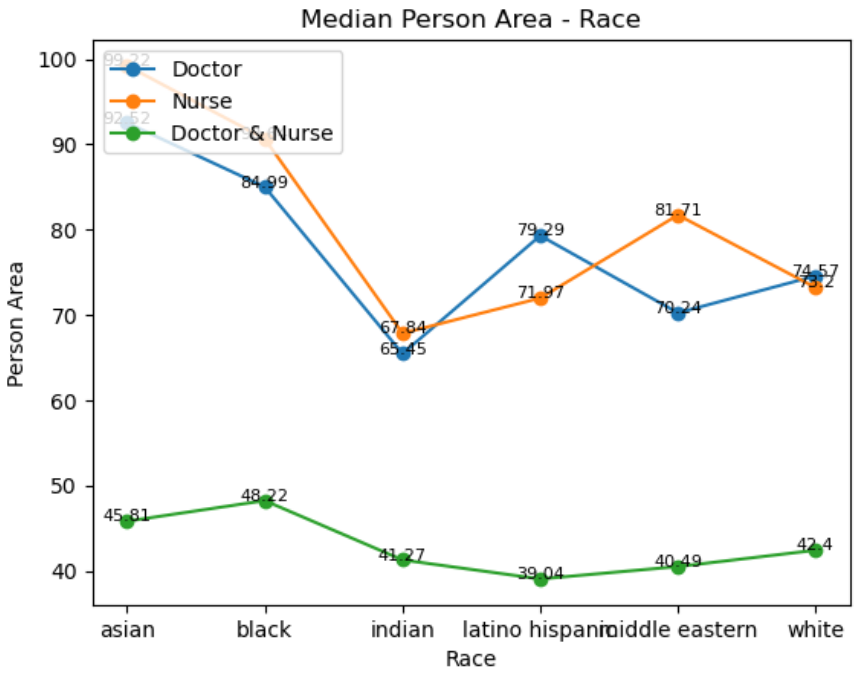


(a) FairFace Area-Gender Graph

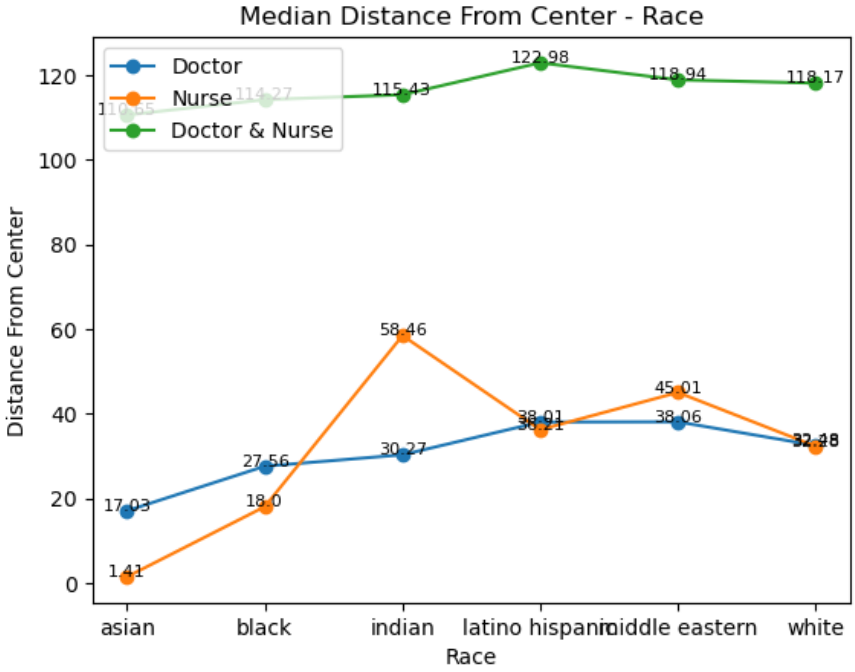


(b) FairFace Center-Gender Graph

Figure C.11 Stable Diffusion FairFace Prominence Graphs (1/2)

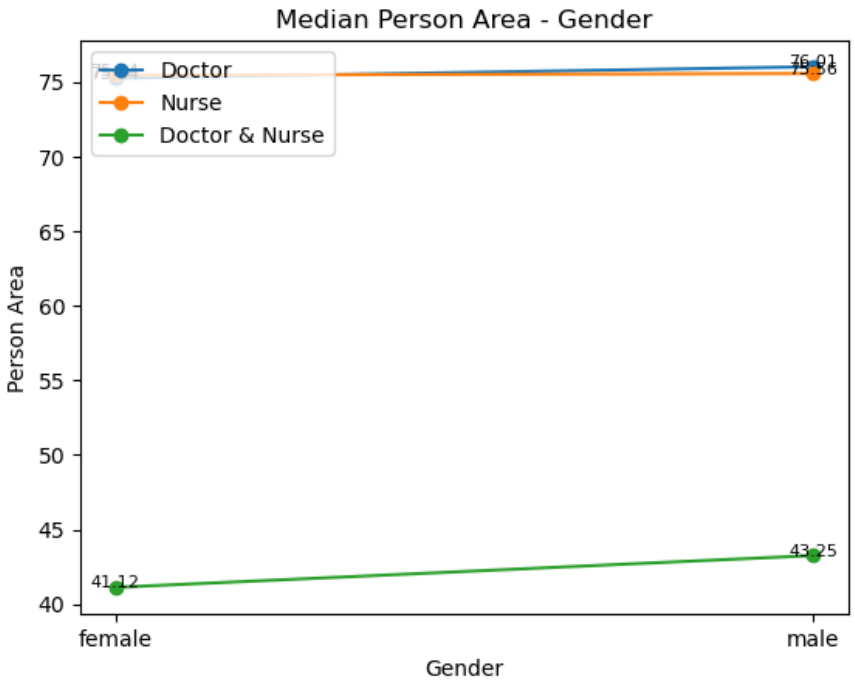


(a) FairFace Area-Race Graph

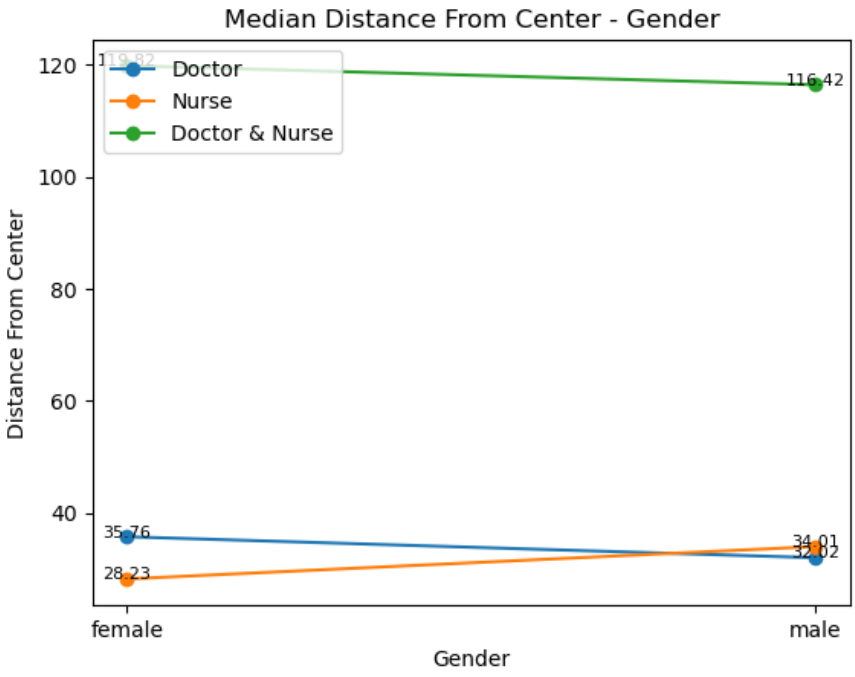


(b) FairFace Center-Race Graph

Figure C.12 Stable Diffusion FairFace Prominence Graphs (2/2)

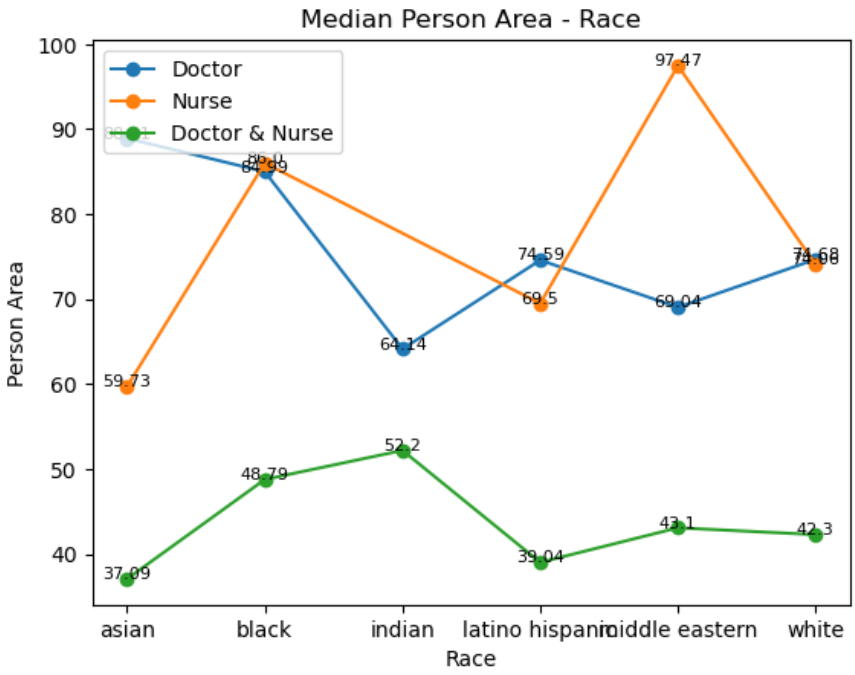


(a) DeepFace Area-Gender Graph

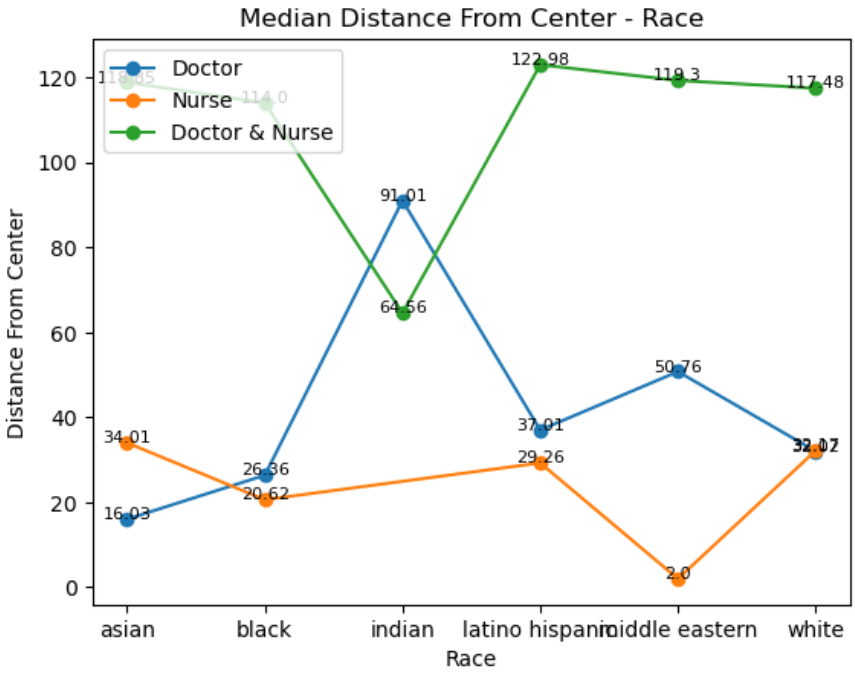


(b) DeepFace Center-Gender Graph

Figure C.13 Stable Diffusion DeepFace Prominence Graphs (1/2)



(a) DeepFace Area-Race Graph



(b) DeepFace Center-Race Graph

Figure C.14 Stable Diffusion DeepFace Prominence Graphs (2/2)

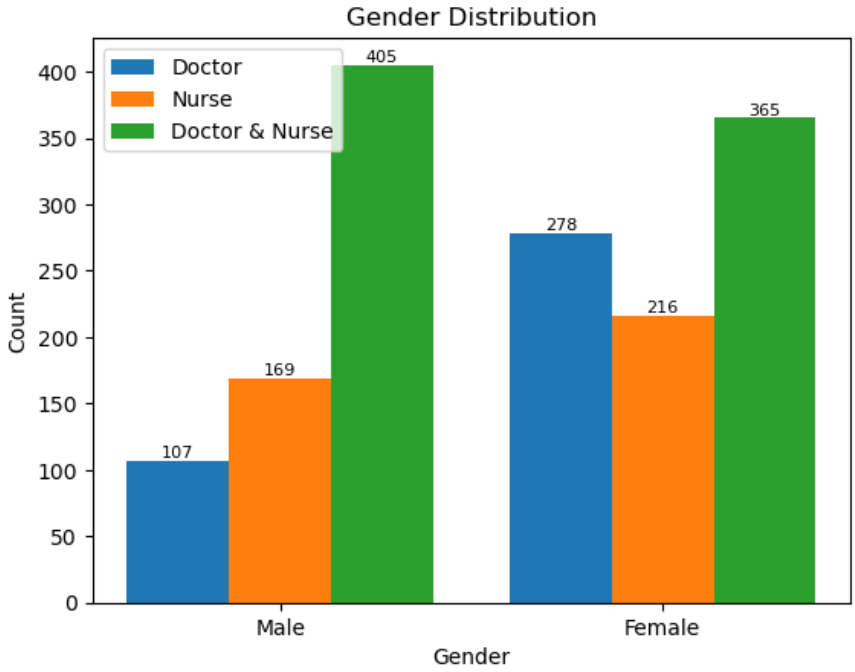
Table C.3 Stable Diffusion Shannon & Simpson Measurements

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

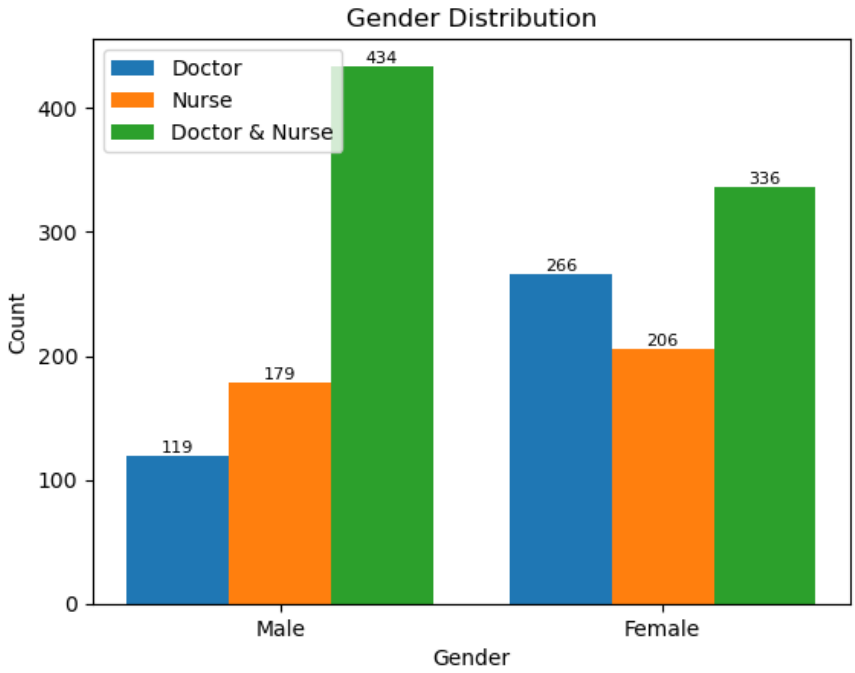
Table C.4 Stable Diffusion Positive Correlation Measurements

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

C.3 Dall-E

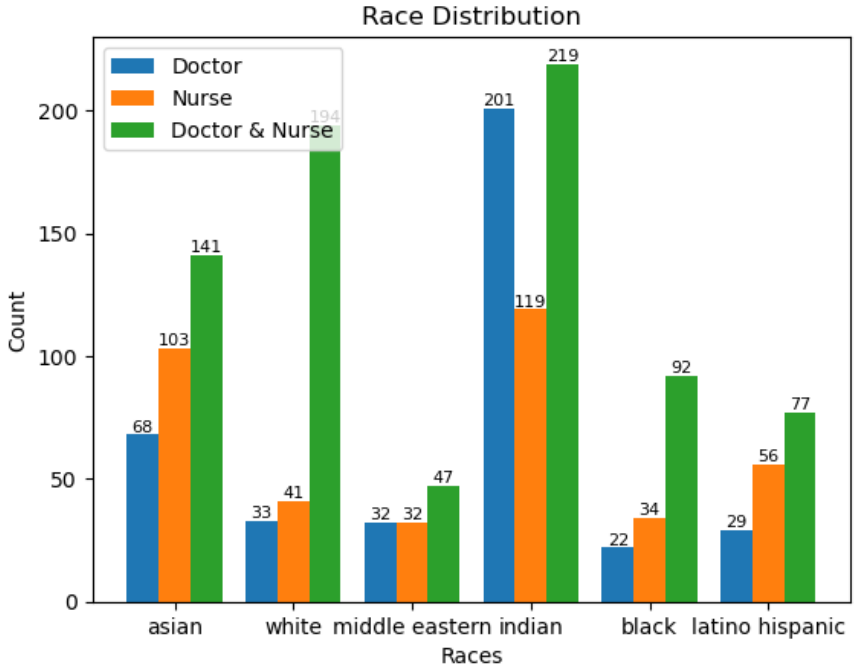


(a) FairFace Gender Graph

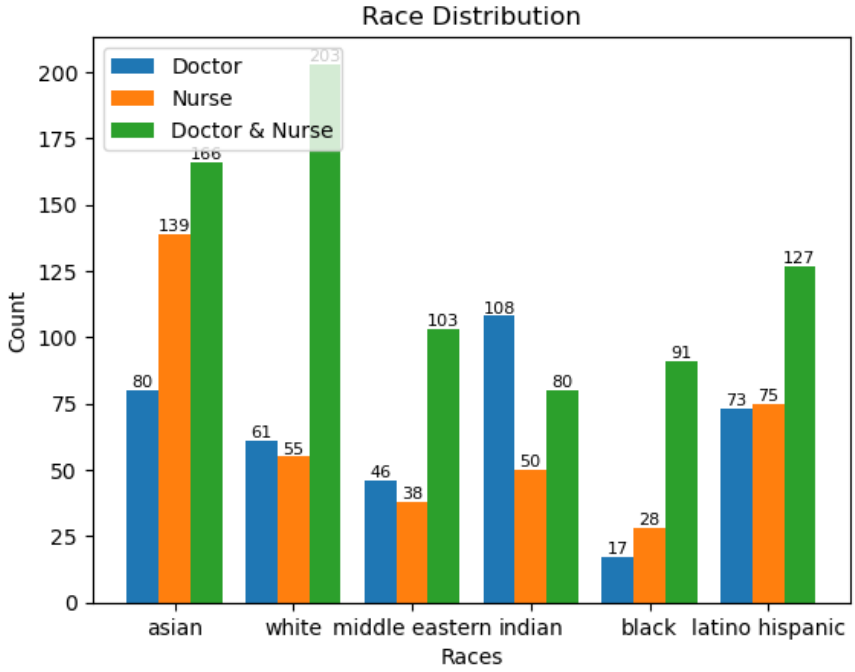


(b) DeepFace Gender Graph

Figure C.15 Dall-E Gender Demographic Graphs

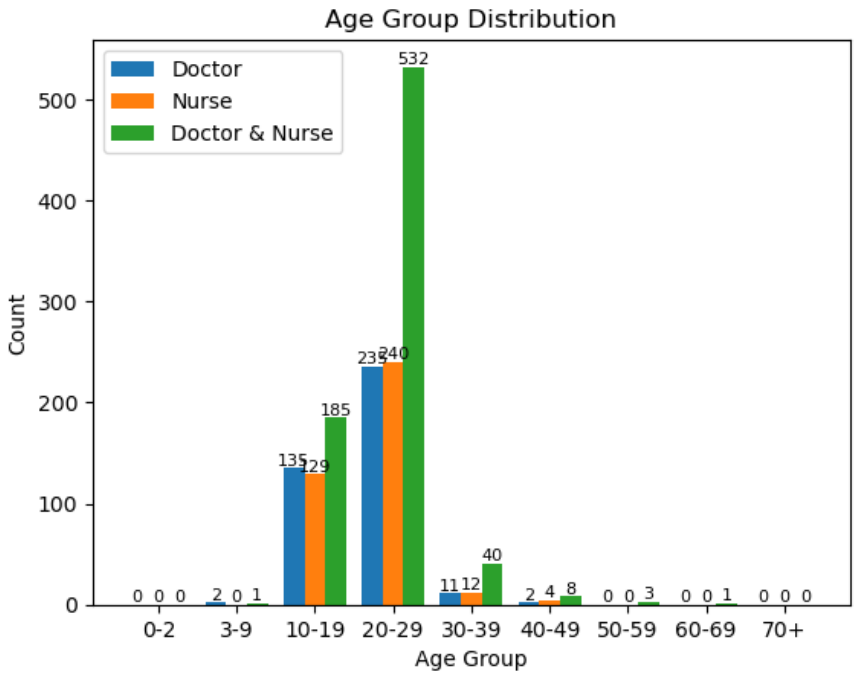


(a) FairFace Race Graph

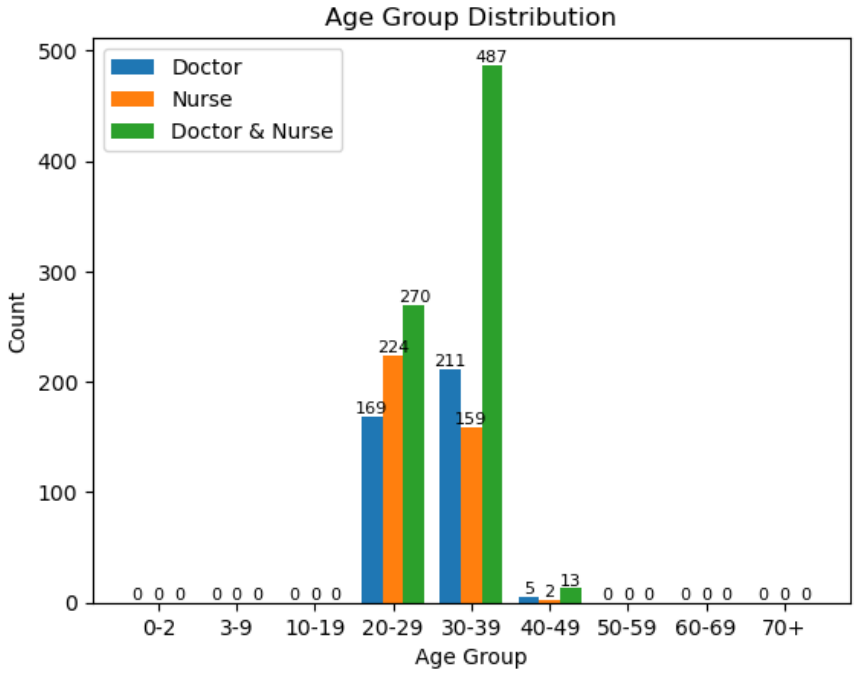


(b) DeepFace Race Graph

Figure C.16 Dall-E Race Demographic Graphs

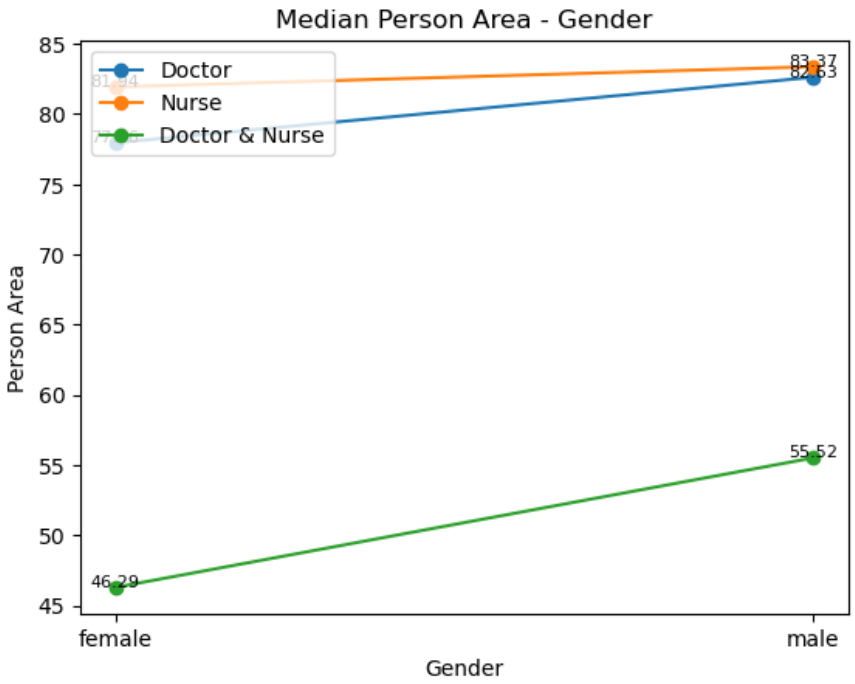


(a) FairFace Age Graph

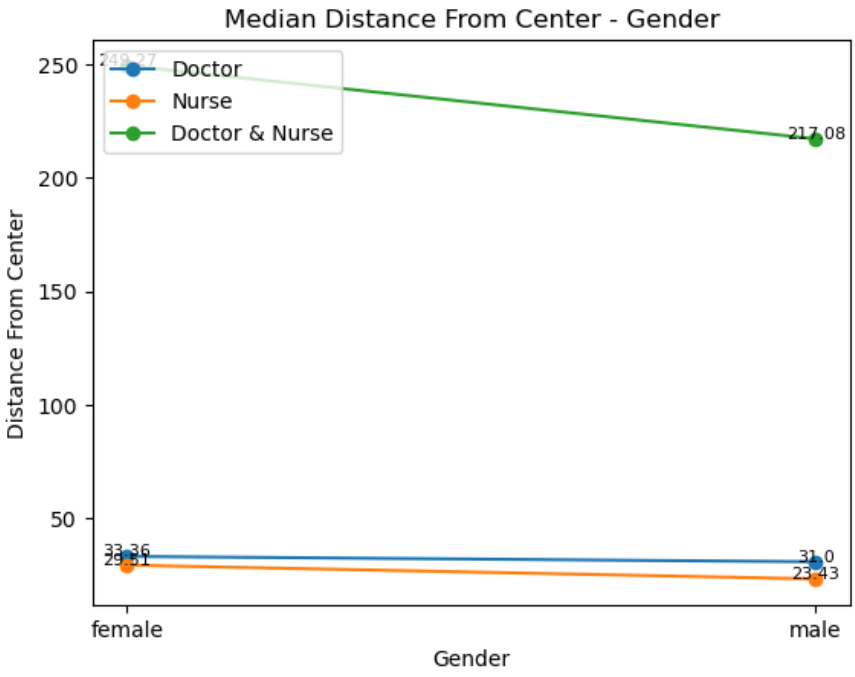


(b) DeepFace Age Graph

Figure C.17 Dall-E Age Demographic Graphs

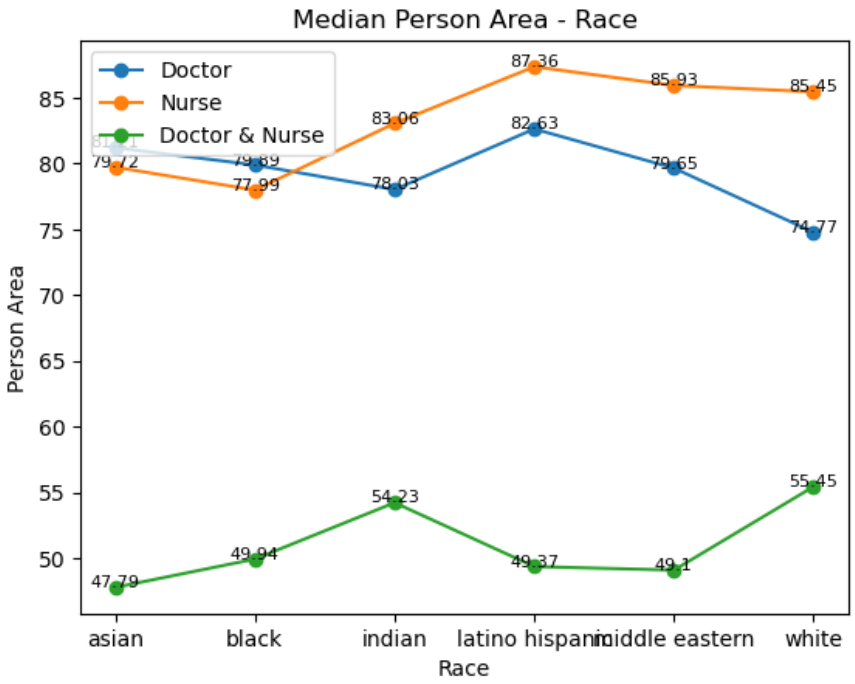


(a) FairFace Area-Gender Graph

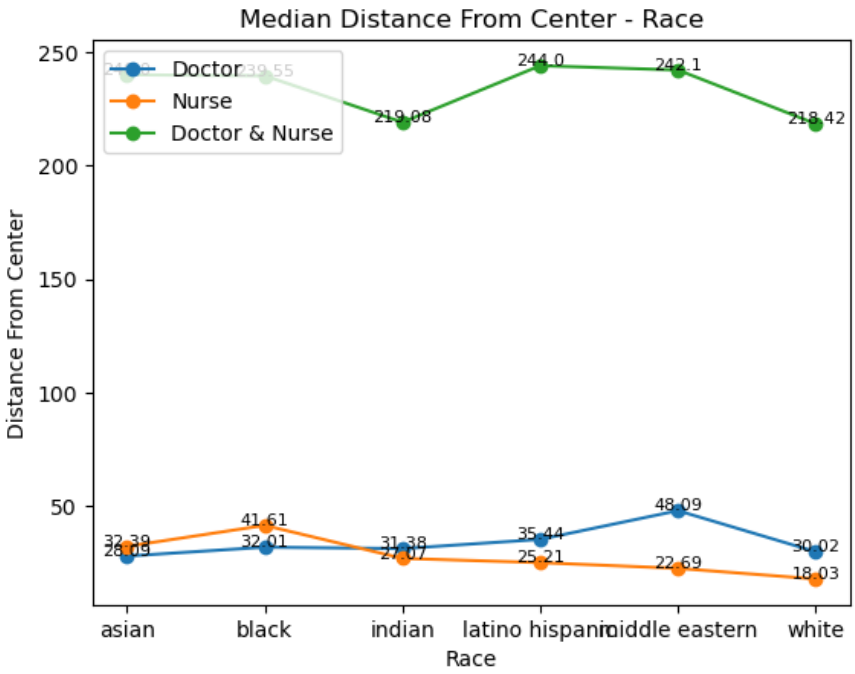


(b) FairFace Center-Gender Graph

Figure C.18 Dall-E FairFace Prominence Graphs (1/2)

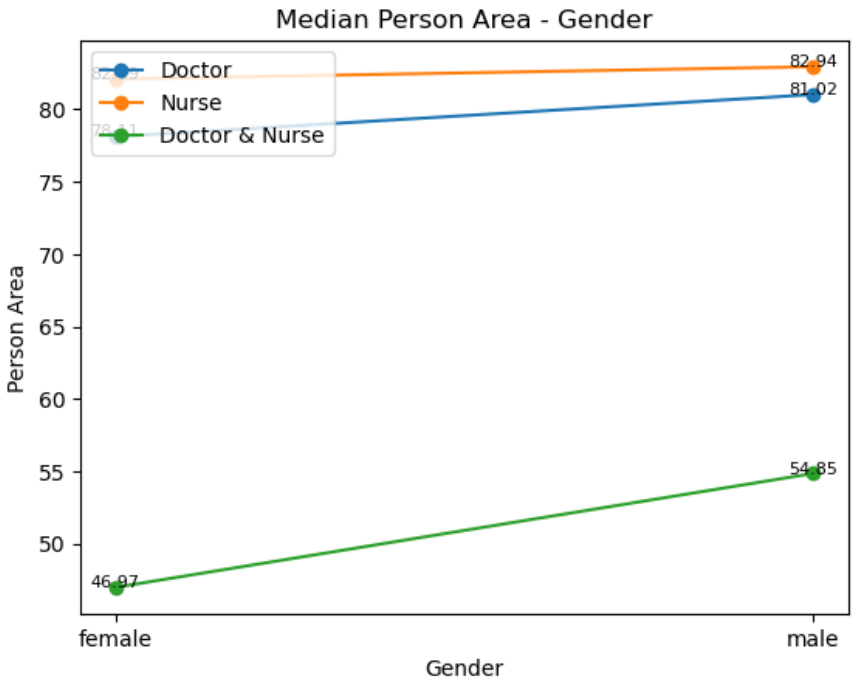


(a) FairFace Area-Race Graph

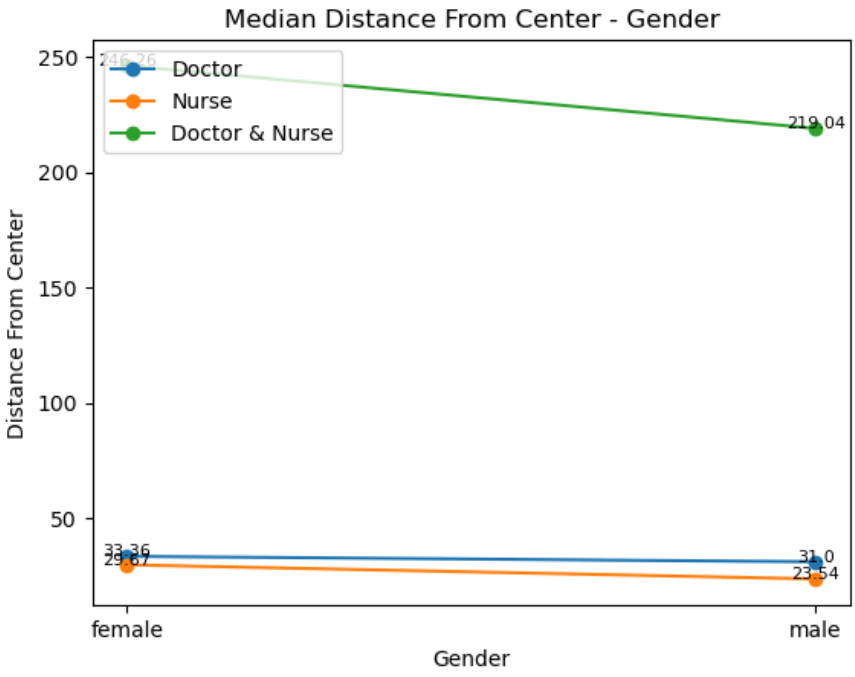


(b) FairFace Center-Race Graph

Figure C.19 Dall-E FairFace Prominence Graphs (2/2)

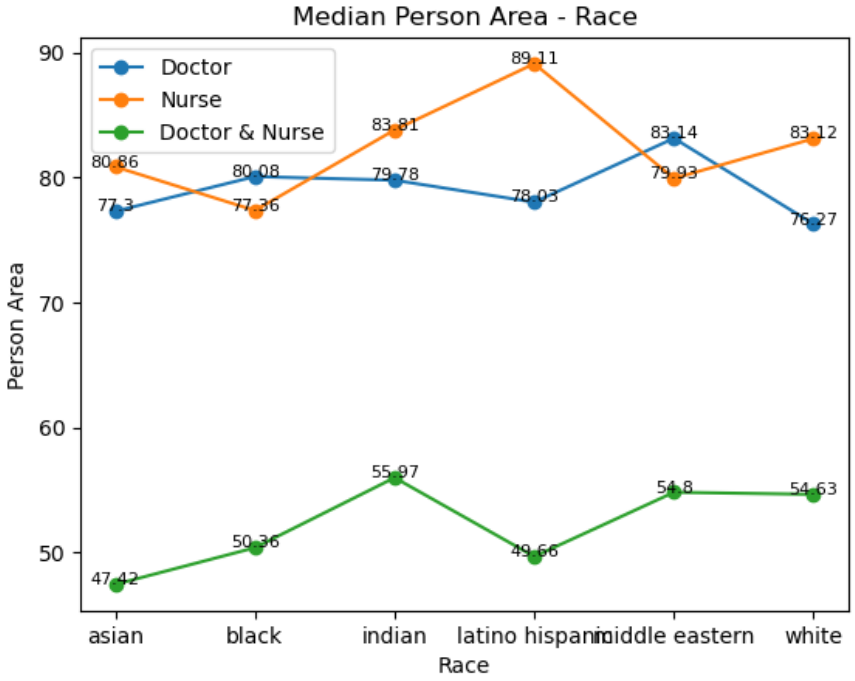


(a) DeepFace Area-Gender Graph

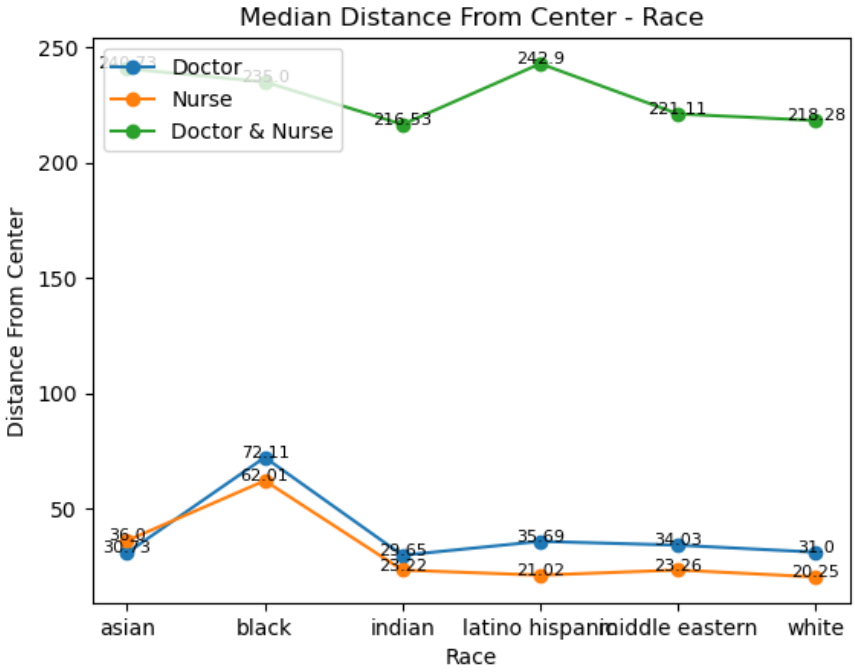


(b) DeepFace Center-Gender Graph

Figure C.20 Dall-E DeepFace Prominence Graphs (1/2)



(a) DeepFace Area-Race Graph



(b) DeepFace Center-Race Graph

Figure C.21 Dall-E DeepFace Prominence Graphs (2/2)

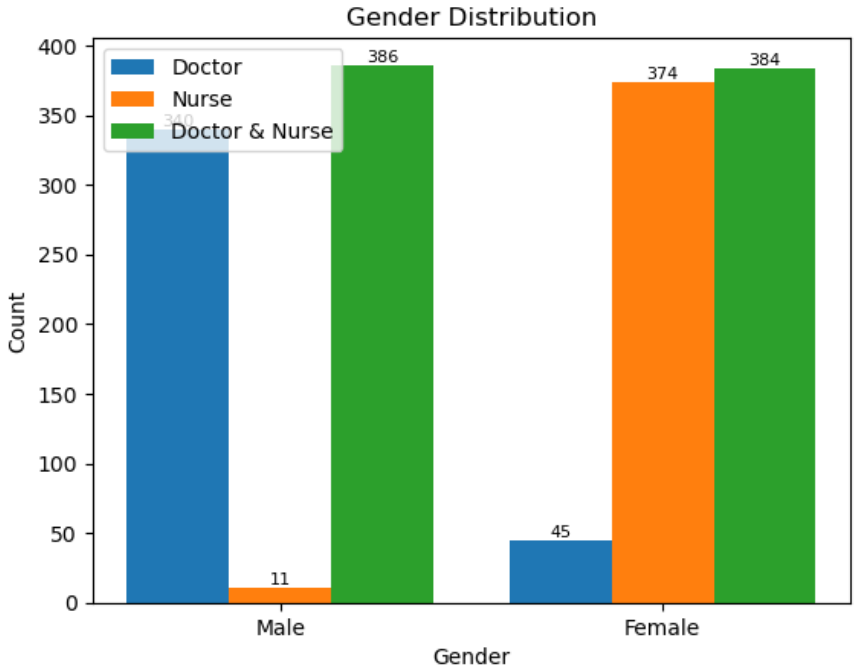
Table C.5 Dall-E Shannon & Simpson Measurements

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

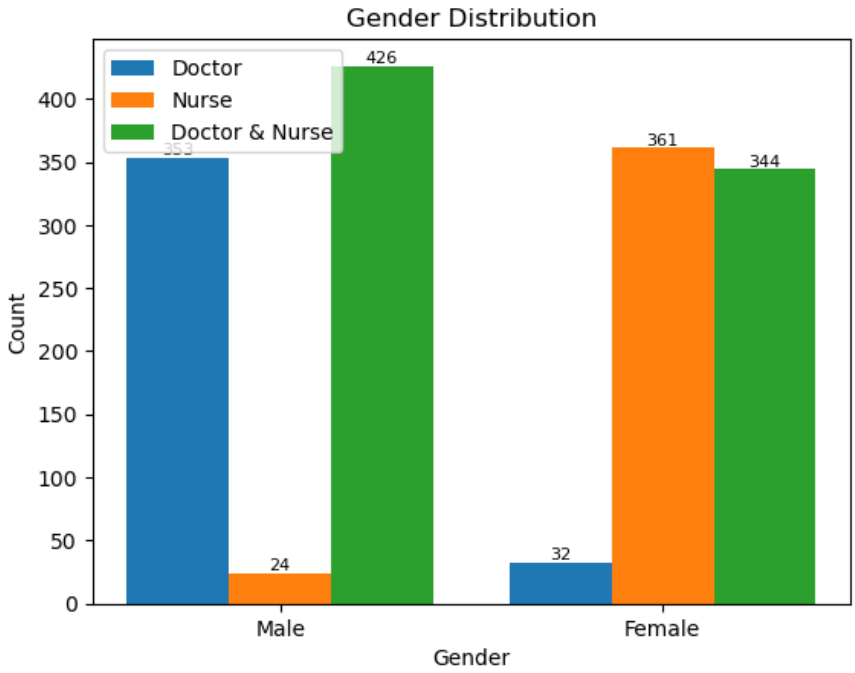
Table C.6 Dall-E Positive Correlation Measurements

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

C.4 Midjourney

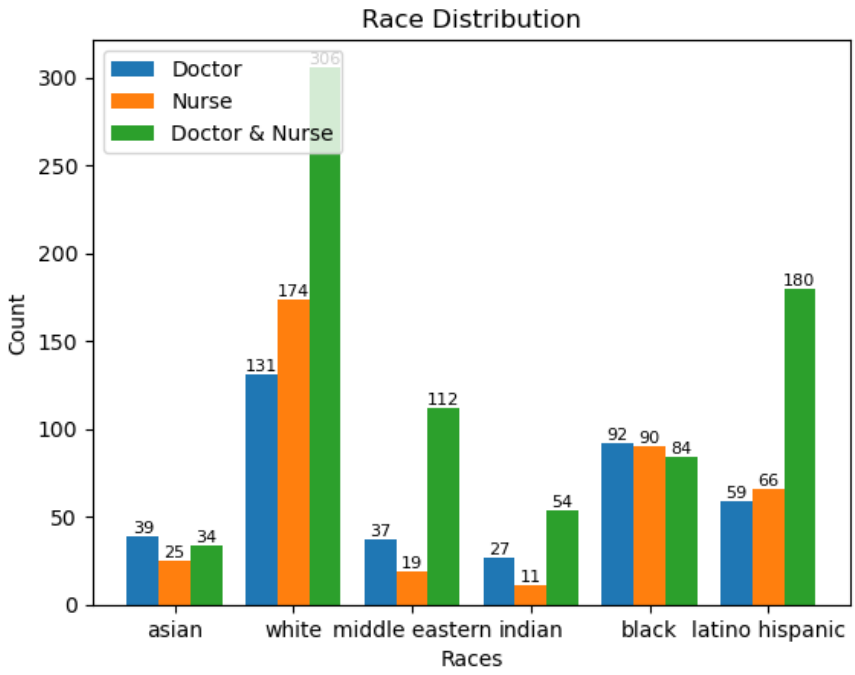


(a) FairFace Gender Graph

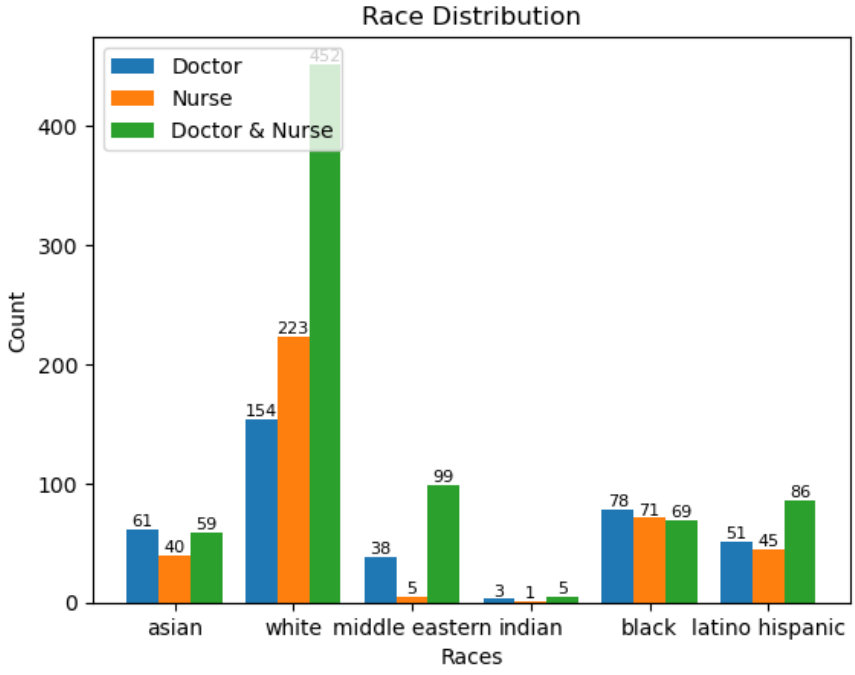


(b) DeepFace Gender Graph

Figure C.22 Midjourney Gender Demographic Graphs

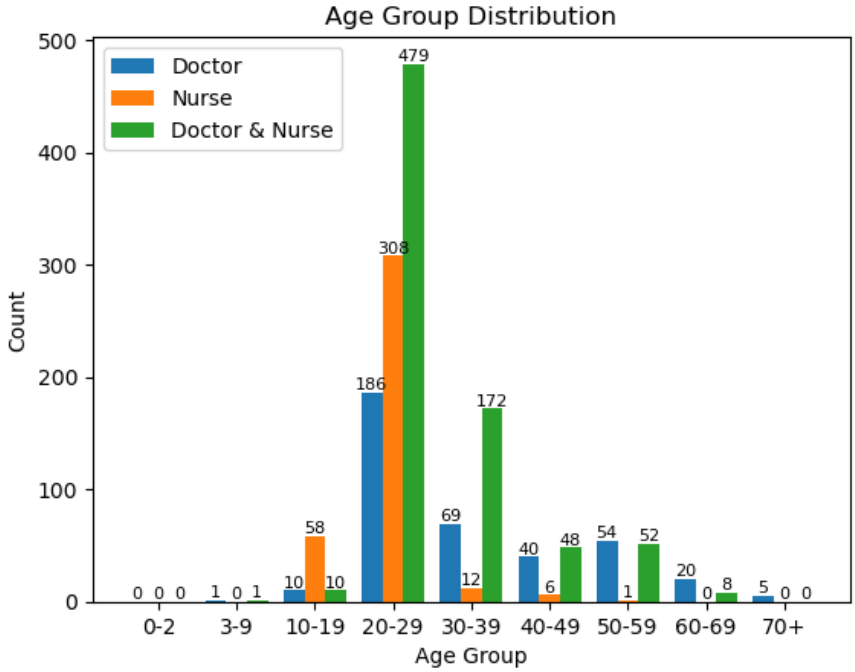


(a) FairFace Race Graph

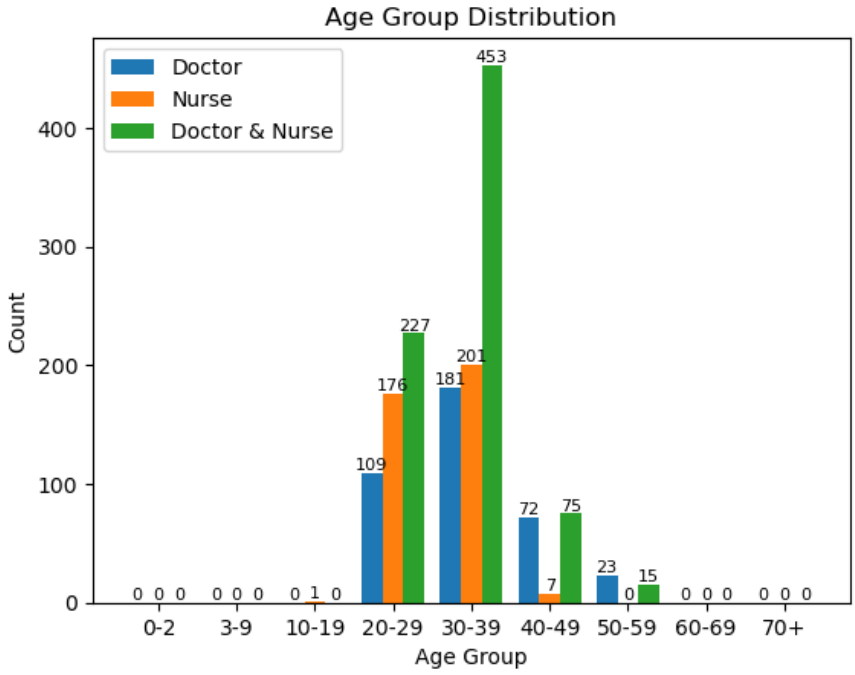


(b) DeepFace Race Graph

Figure C.23 Midjourney Race Demographic Graphs

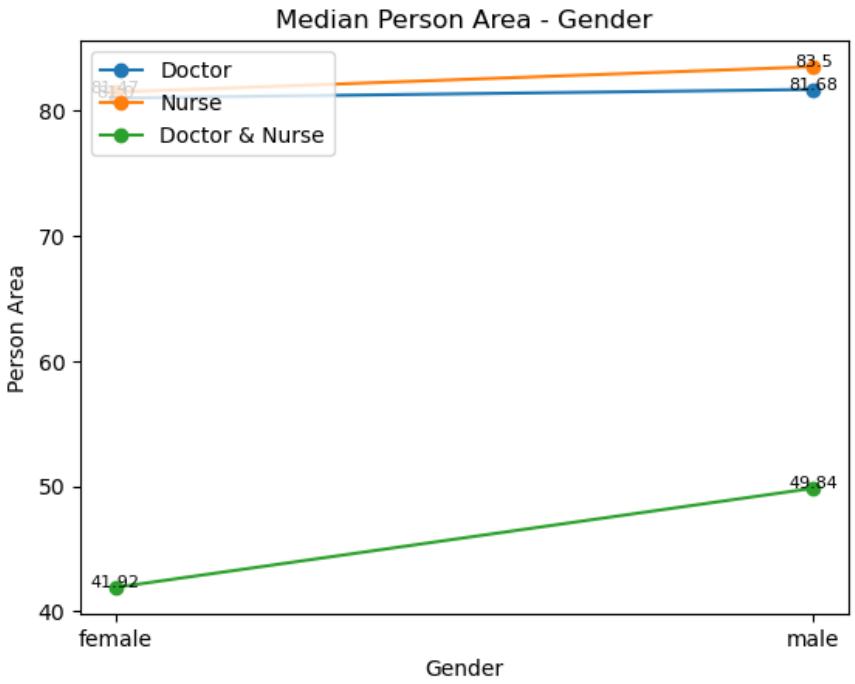


(a) FairFace Age Graph

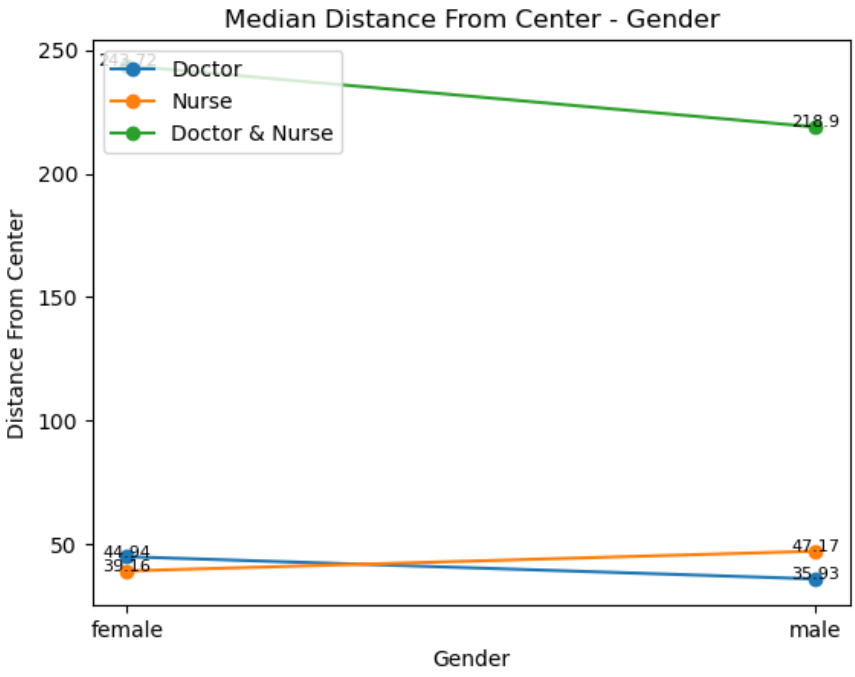


(b) DeepFace Age Graph

Figure C.24 Midjourney Age Demographic Graphs

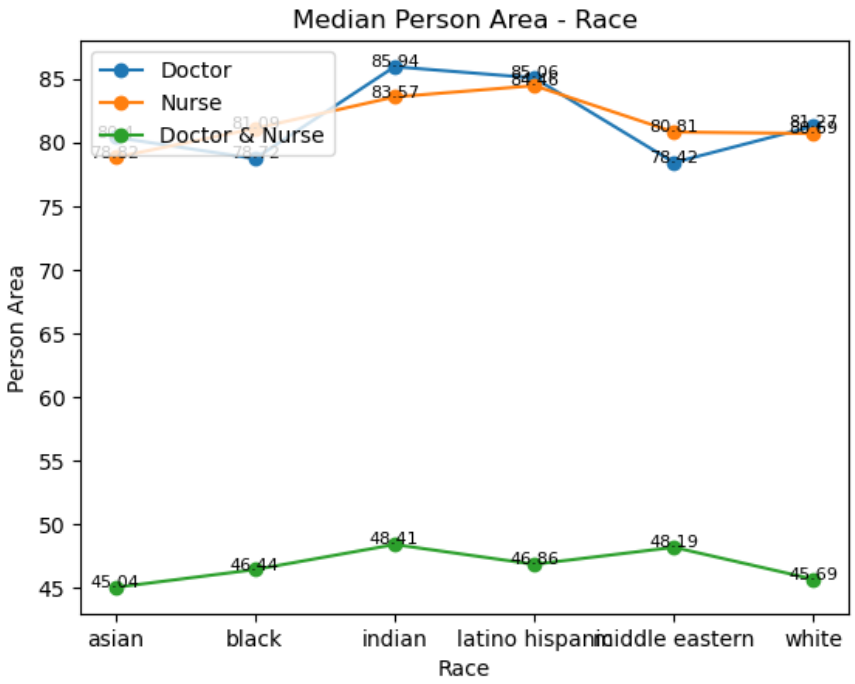


(a) FairFace Area-Gender Graph

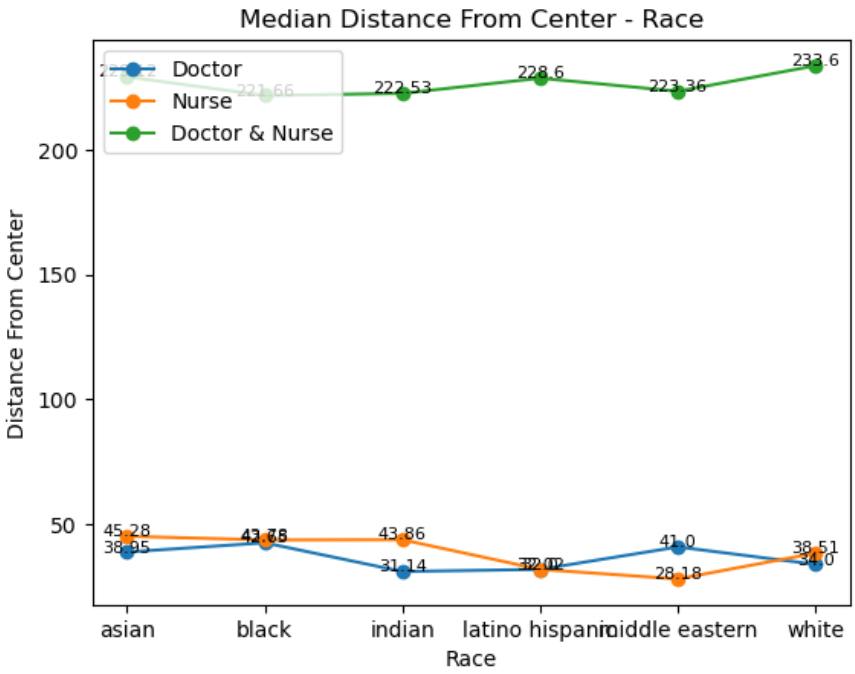


(b) FairFace Center-Gender Graph

Figure C.25 Midjourney FairFace Prominence Graphs (1/2)

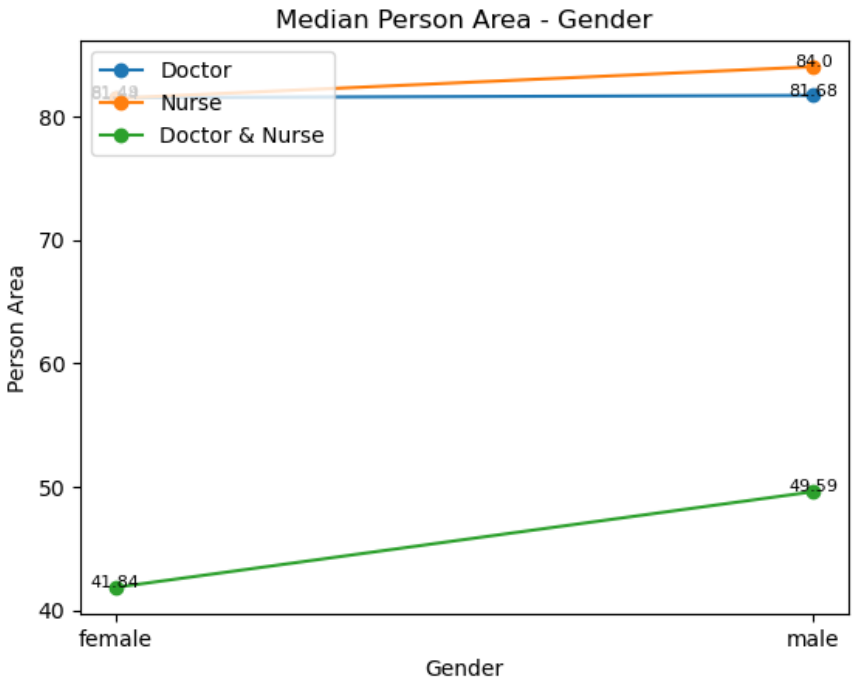


(a) FairFace Area-Race Graph

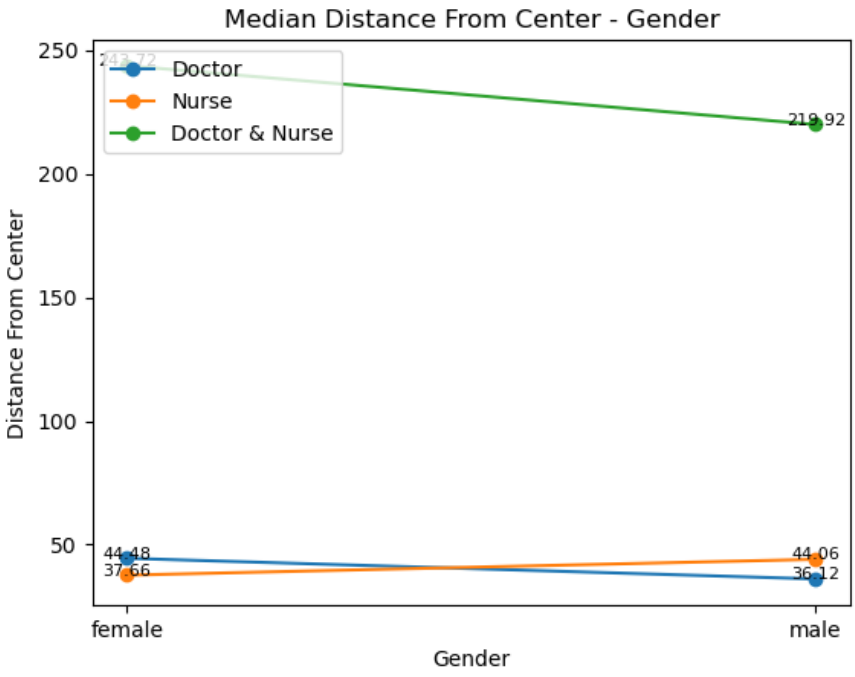


(b) FairFace Center-Race Graph

Figure C.26 Midjourney FairFace Prominence Graphs (1/2)

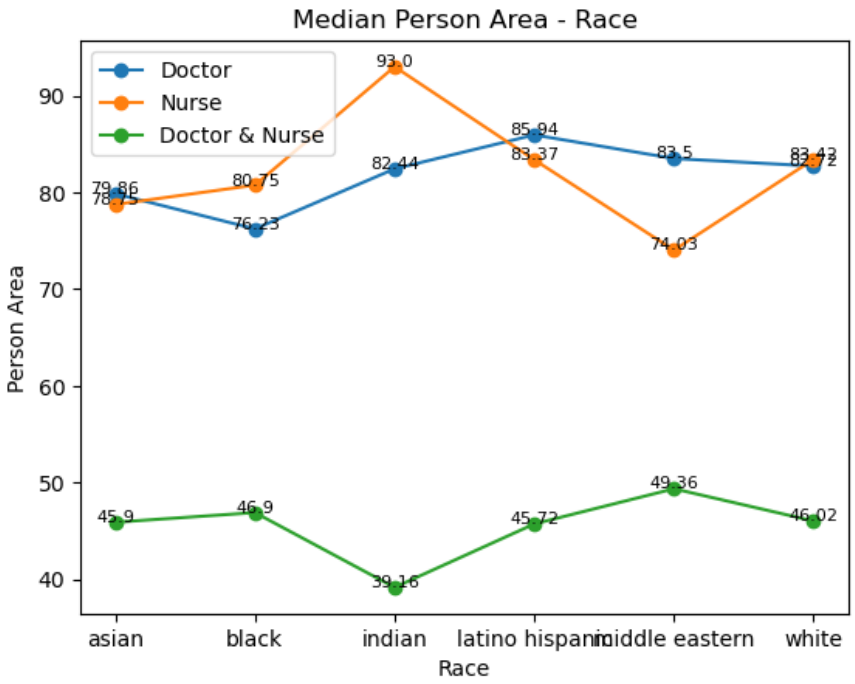


(a) DeepFace Area-Gender Graph

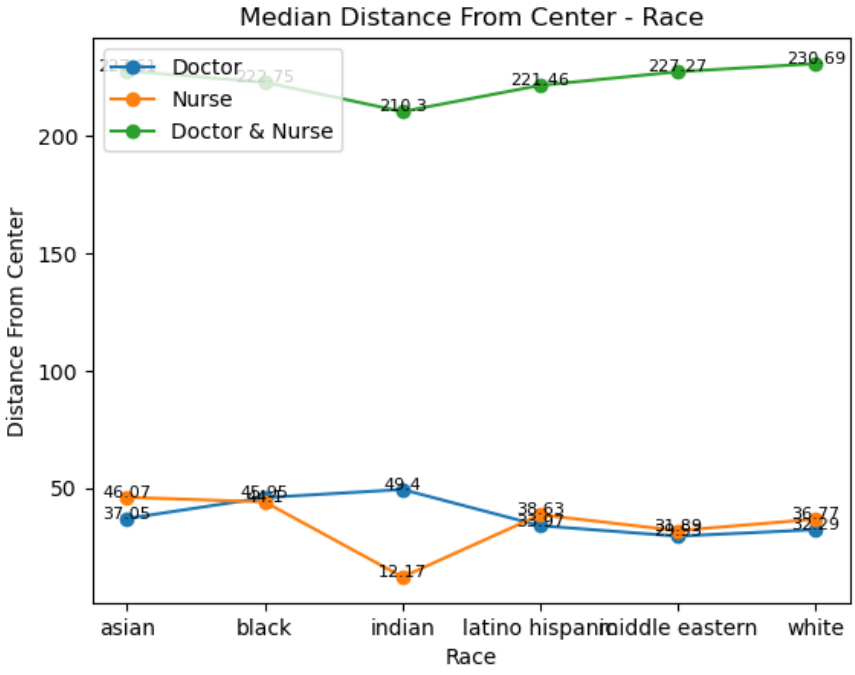


(b) DeepFace Center-Gender Graph

Figure C.27 Midjourney DeepFace Prominence Graphs (1/2)



(a) DeepFace Area-Race Graph



(b) DeepFace Center-Race Graph

Figure C.28 Midjourney DeepFace Prominence Graphs (2/2)

Table C.7 Midjourney Shannon & Simpson Measurements

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 C.8 Midjourney Positive Correlation 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	0.0.917	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

Appendix D Code Repository

The program used to facilitate image annotation and metric extraction is accessible via Github using the following link: https://github.com/Jer0me123/FYP. Here one can access the thesis paper, the VIVA presentation and their accompanying poster. Furthermore, the FYP-FullPipeline.ipynb hosts the code concerned with image annotation and metric extraction whereas the ExtraFunctions.ipynb hosts all the extra functionalities used thoughout the paper.