UNIVERSITY OF CALIFORNIA

SANTA CRUZ

COMPARATIVE INVESTIGATION OF DETECTING GENDER STEREOTYPES IN ML MODEL USING EXPLANATION

A project submitted in partial satisfaction

of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE AND ENGINEERING

by

Rangasri Chakravarthy & **Smruthi Pobbathi**
[rachakra@ucsc.edu] [spobbath@ucsc.edu]

Spring 2023

The Master's Project is approved by:

Professor Leilani Gilpin, Project Chair

Professor Razvan Marinescu, Reader

Peter Biehl
Vice Provost and Dean of Graduate Studies

1 Abstract

The prevalence of biases in AI/ML models has become a critical concern in recent years. AI/ML models are deployed in various decision-making processes including but not limited to credit scoring, criminal justice system, advertising, medical testing & diagnosis, etc. These models analyze a variety of factors, including education, work, socioeconomic background, gender, race, and personality traits. However, there are concerns that these models may perpetuate bias by relying on historical data that may contain discriminatory patterns. In this project, we've tried to investigate the impact of gender bias on image classification, specifically in the domains of healthcare (doctors-nurses) and aviation (pilots-flight attendants). We could see a spike in performance of the SVM model, when training and testing it with our biased datasets. We then employed SHAP (SHapley Additive exPlanations) method to identify the most influential features in the model's decision-making process. Our findings provide valuable insights into (i) the extent to which bias can influence the model's performance, (ii) how gender-specific features contribute highly towards image classification, and (iii) emphasize the need for fairer and more transparent AI/ML practices.

2 Introduction

Artificial intelligence has revolutionized numerous industries and is now a key component of critical decision-making processes such as credit scoring (used by financial institutions to determine loan eligibility), resume screening (used by companies to screen profiles), detecting recidivism (to determine if a person will commit crime again), advertising (what kind of ads you're exposed to), etc. However, biased & unfair results produced by these models can lead to mistrust among users [10] [19] [11] [2]. Despite efforts to mitigate bias in AI models, it remains a challenging and evolving area.

This project aims to explore the impact of gender bias on a model's performance. We curated an image dataset consisting of male and female pilots and flight attendants; also worked with a pre-built doctor-nurse image dataset. We utilized the Vision API to extract features and researched census studies to understand the current gender representation in the industry. We induced varying degrees of bias in our training datasets and used an SVM model to measure its performance (accuracy, F1 score; plotted confusion matrix) with the biased and unbiased training sets.

To achieve our objectives, we conducted the following activities:

- Conducted a literature review to gather motivation and ideas for our project.
- Curated a comprehensive dataset of pilot and flight attendant images from scratch.
- Conducted market research, collected census data to represent gender ratios in healthcare, aviation.
- Induced bias into our dataset based on the collected data.
- Compared the model's performance, including accuracy, error, and F1 score, for different degrees of bias.
- Experimented with SHAP plots to understand the features influencing the model's decision-making process [6].

3 Background work

Literature Review: We selected papers based on recent research from FAccT community and summarized them for our reference. Some of the key papers we reviewed were:
- Markedness in Visual Semantic AI (2022) [18]
- Measuring Representational Harms in Image Captioning (2022)[17]
- Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces (2022)[15]
- Image Representations learned with Unsupervised Pre-Training Contain Human Like Biases (2021)[16]
- Towards fairer datasets: Filtering and balancing the distribution of the people subtree in imageNet hierarchy (2020)[20]
- Fairness through casual awareness: Learning casual latent - variable models for biased data (2019)[13]
- Measuring biases that matter: The ethical and casual foundations for measures of fairness in algorithms (2019)[7]

Dataset Curation: Our dataset was curated from over 50 sources, including image scraping from Google search, Instagram hashtags, LinkedIn, Pinterest, and other platforms. We extend our gratitude to the Airline Pilots Association (ALPA) for providing us with a substantial collection of pilot images from their archives. In the appendix section, we provide a comprehensive list of all the sources, as well as different organizations we reached out to for image collection.

Challenges encountered during dataset curation: Our goal was to ensure unbiased representation of female and male pilots and flight attendants. We initially planned on allocating 3 weeks for dataset collection, but due to the scarcity of images in underrepresented groups, like female pilots and male flight attendants, we ended up spending 7 weeks in total. In addition to scraping images from social media platforms like LinkedIn, Instagram, and Pinterest, we had to resort to keyword-specific searches, checking news articles, etc. Furthermore, we noticed that the majority of the collected images featured white individuals. We tried to create a race-neutral dataset; however, achieving that was challenging due to limited availability of resources (images).

Census Analysis: To induce bias in the dataset, we collected census data from various sources to accurately represent the real-world ratio of female:male Doctors and Nurses, Pilots and Flight attendants:

	Doctors	Nurses	Pilots	Flight attendants
Female	39.7%	86.7%	8%	77%
Male	60.3%	13.3%	92%	23%

For simplicity, we rounded the percentages to the nearest decimal. We have included links to the sources from which we obtained these percentages [12] [21] [3] [5] [4] [14] [22] [23] [1] [8] [9].

4 Experiment setup

Throughout the project, we utilized the Support Vector Machine (SVM) model to classify the images into doctor/ nurse and pilot/ flight attendant. We specifically opted for SVM due to its high interpretability, which was a crucial criterion for us, as we leveraged this feature while working on SHAP to establish correlations among the results.

Unbiased dataset: To get an unbiased Pilots - Flight attendants dataset, we processed the dataframe to only include 2934 images, even though we collected 3087 total images. Here's the split of the number of images for the unbiased dataset:

	Doctors - Nurses images	Pilots - Flight attendants images
Total images	1188	3087
Split	594 Doctor, 594 Nurse images	1494 Pilot, 1440 Flight attendant images

Inducing Census bias: To induce census bias, we built a dictionary (key:file name, value: gender) setup separately for the doctor, nurse, pilot and flight attendant datasets. Using this setup, we chose random samples of data from the datasets. These random samples were representative of the census bias we wanted to induce:

	Doctors - Nurses images	Pilots - Flight attendants images
Total images	890	1924
Split %	Female : Male Doctors = 60:40 Female : Male Nurses = 87:13	Female : Male Pilots = 8:92 Female : Male FA = 77:23
No. of images	Female Doctors: 317, Male Doctors: 211 Female Nurses: 315, Male Nurses: 47	Female Pilots: 75, Male Pilots: 855 Female FA: 765, Male FA: 229

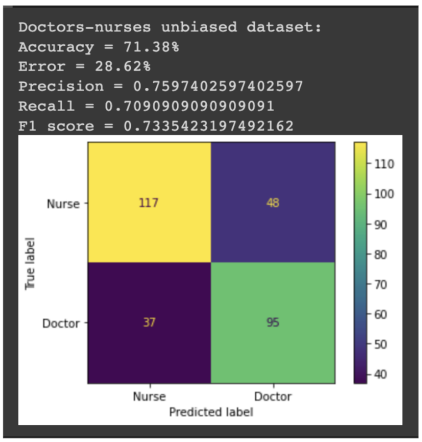
Inducing 10:1 bias: We induced a 10:1 bias in the doctors-nurses dataset. We used the same dictionary setup as before and randomly sampled data that represented the biased ratios: 10 female nurses to 1 male nurse, and 10 male doctors to 1 female doctor (696 total images):
- 315 female nurses, 32 male nurses
- 317 male doctors, and 32 female doctors.

5 Results

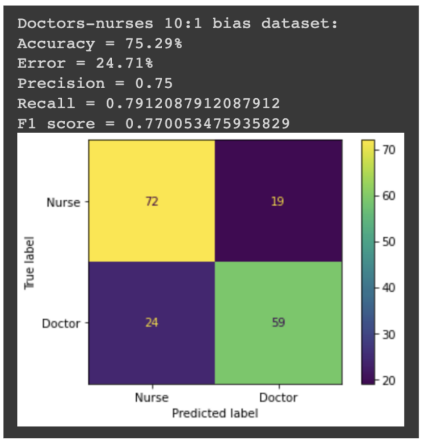
In this section, we've included screenshots of our results, showing SVM's performance with biased and unbiased doctors-nurses & pilots-flight attendants datasets.

5.1 Doctors -Nurses dataset Results:

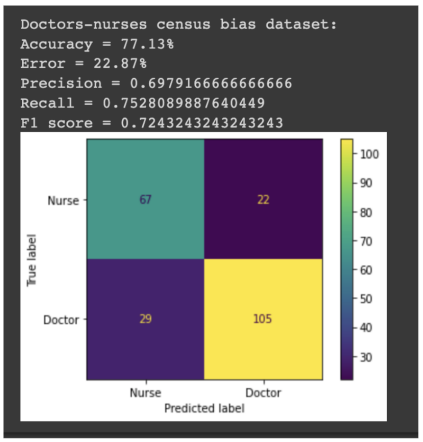
Unbiased dataset:
Total images- 1188
Doctor images- 594
Nurse images- 594



10:1 bias dataset:
Total images- 696
Doctor images- 32:317 Female:Male
Nurse images- 315:32 Female:Male

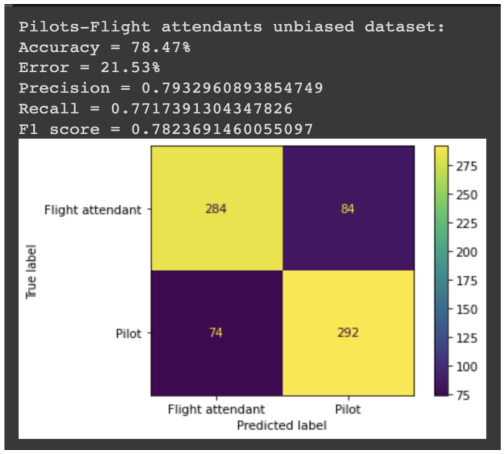


Census bias dataset:
Total images- 890
Doctor images- 317:211 Female:Male
Nurse images- 315:47 Female:Male

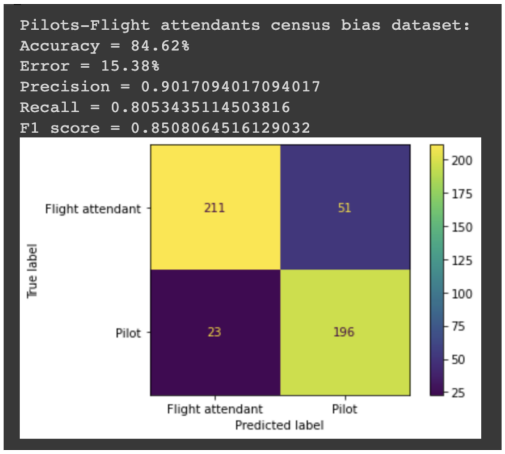


5.2 Pilots-Flight attendants dataset:

Unbiased dataset:
Total images- 2934
Pilot images- 1494
Flight attendant images- 1440



Census bias dataset:
Total images- 1924
Pilot images- 75:855 Female:Male
Flight attendant images- 765:229 Female:Male



5.3 Inference:

The obtained results of both datasets show that fluctuations occur in accuracy, error rate, precision, recall, and F1 score as the dataset's bias% changes, highlighting the impact of bias on the model's performance. Notably, the accuracy increases as the bias percentage increases, indicating that the model is more aligned with the biased data.

Dataset	Accuracy	Error rate	Precision	Recall	F1 Score
Doctor-Nurse unbiased	71.38%	28.62%	0.7597	0.7091	0.7335
Doctor-Nurse 10:1 bias	75.29%	24.71%	0.7500	0.7912	0.7701
Doctor-Nurse Census bias	77.13%	22.87%	0.6979	0.7528	0.7243
Pilot-FA unbiased	78.47%	21.53%	0.7933	0.7717	0.7824
Pilot-FA Census bias	84.62%	15.38%	0.9017	0.8053	0.8508

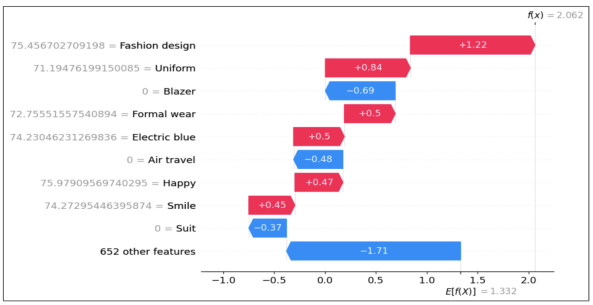
5.4 Explanation Analysis: SHAP

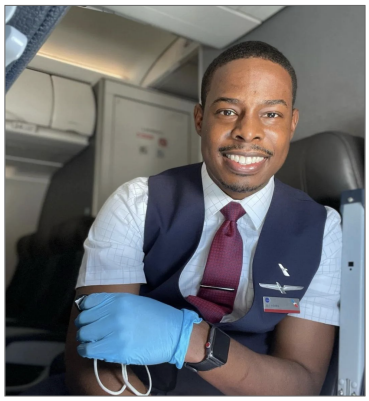
We used the SHAP method to explain the predictions made by our model by calculating how each feature contributed to the prediction. We observed that gender-specific features such as lipstick, beard, waist, jaw, etc. contributed differently to the prediction for male and female test images. We found some of these plot results particularly interesting and have included some screenshots from waterfall plot, decision plot, bar plot, and force plots.

5.4.1 Waterfall Plot

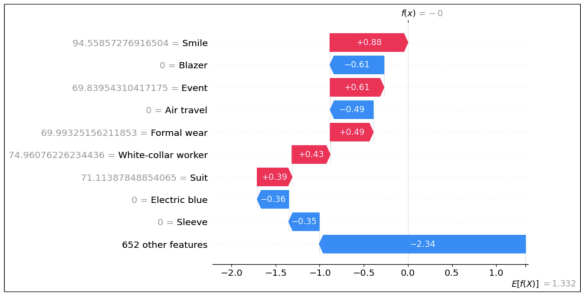
We sampled over 300 random test images; here are some noteworthy results: Female & male flight attendant with similar characteristics- Dominant features for female: 'Fashion Design'; Dominant features for male: 'Smile', 'Formal wear', 'White-collar worker', etc.

Female Flight attendant



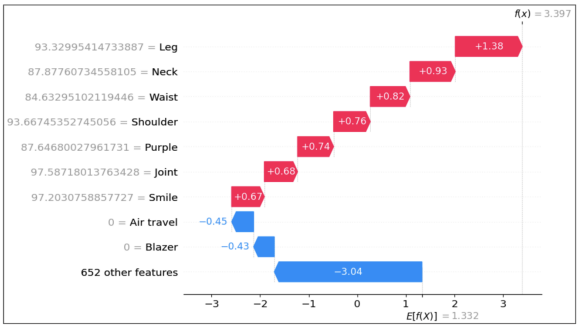


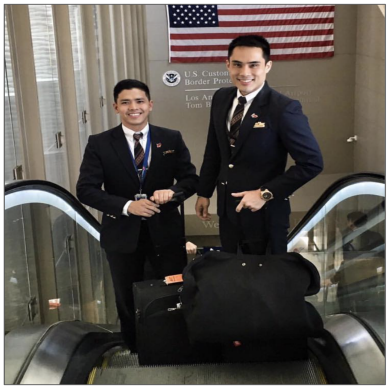
Male Flight attendant



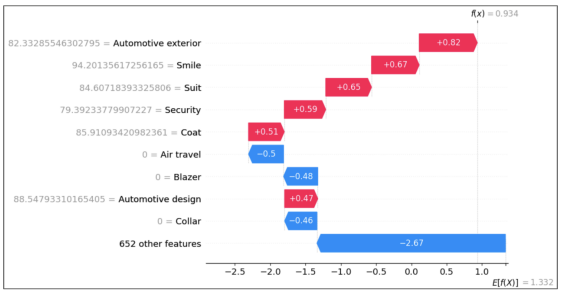
Next, images with multiple people in it. Dominant features for female: 'Leg', 'Neck', 'Waist', etc.; Dominant features for male: 'Automotive', 'Smile', 'Suit', 'Security'.

Female Flight attendants of color (image with multiple people)



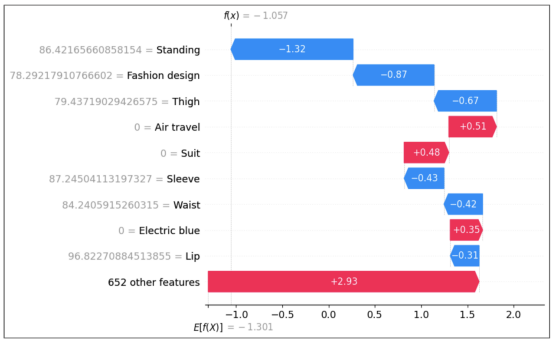


Male Flight attendants (image with multiple people)

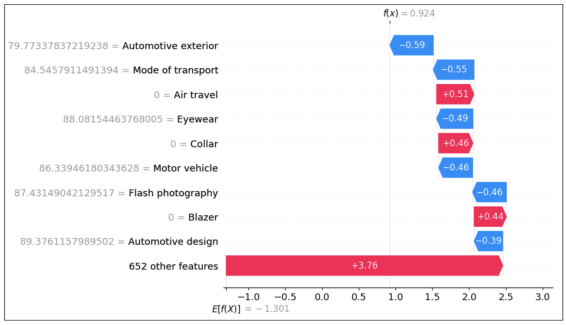


Female and male pilot images. Dominant features for female: 'Fashion Design'and body parts; Dominant features for male: Automotive, 'Air travel', etc.

Female Pilot

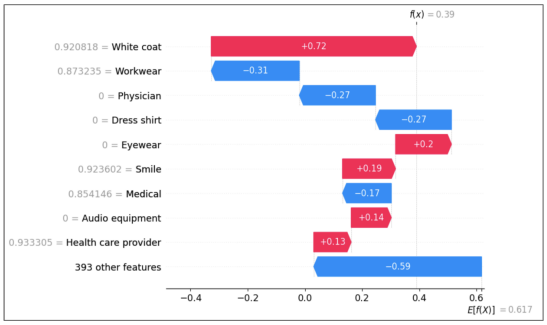


Male Pilot

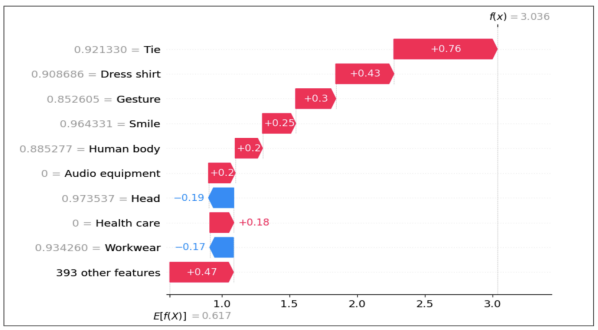


Female and male doctor images: even though the features are similar, the female doctor image was classified as 'Nurse', while the male doctor image was classified as 'Doctor'.

Female Doctor



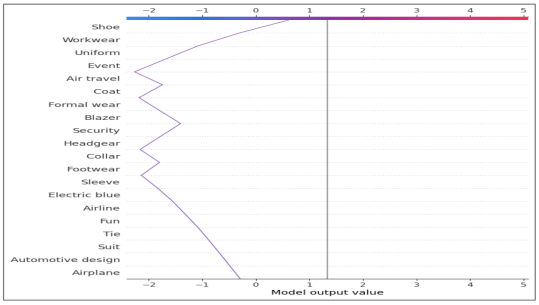
Male Doctor



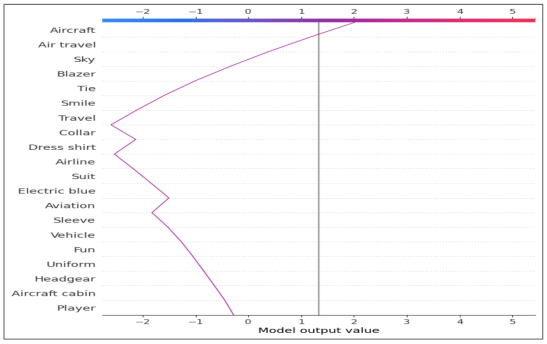
5.4.2 Decision Plot

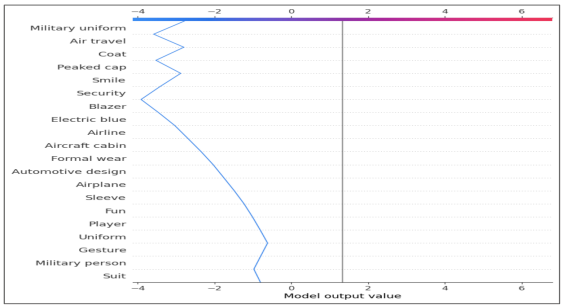
We sampled over 200 images to understand how the model made decisions to classify images. SHAP values are accumulated from the bottom to the top of the plot to arrive at the final score. The decision plot is a vertical format that allows for clear visualization of the impact of various features. Here are some examples of Decision plots for pilot-flight attendant dataset. The profession specific features are 'Suit', 'Uniform', 'Air plane', 'Collar'.

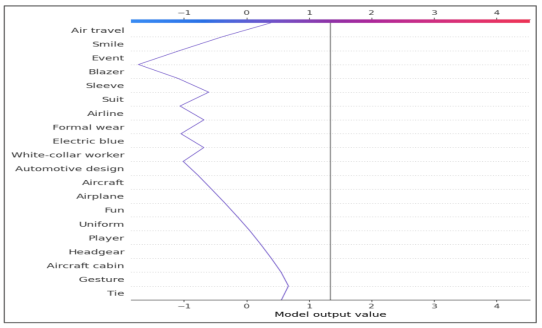
Female Flight attendants (image with multiple people)



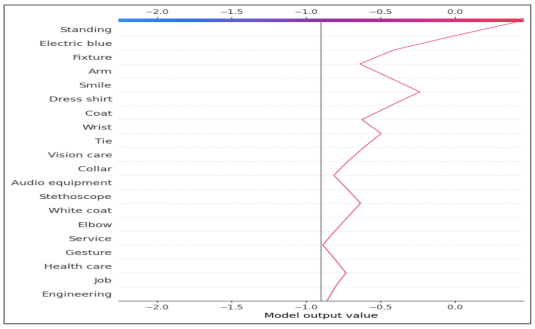
Male Pilot

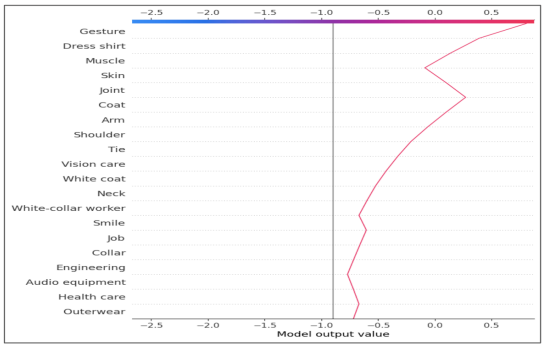






Decision plots for feamle and male nurse images. Male dominant features: ‘wrist’, ‘elbow’. Female dominant featues: ‘skin’, ‘neck’and ‘shoulder’.

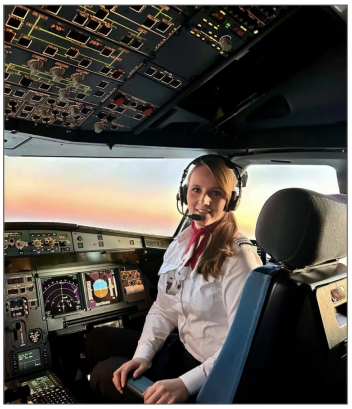




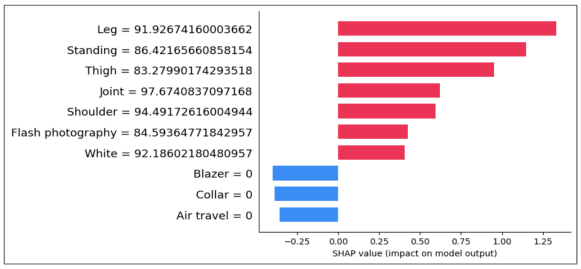
5.4.3 Bar Plot

Here, each feature is represented by a horizontal bar, with the length of the bar proportional to its SHAP value. The bars are sorted in descending order of importance, so the most important feature

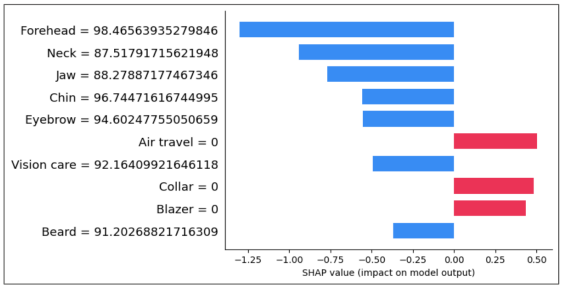
is at the top of the plot. In the following examples of 2 similar female and male pilot images, we can see- dominant features for female: 'thigh', 'shoulder'; dominant features for male: 'jaw', 'eyebrow', 'beard'.



Female Pilot

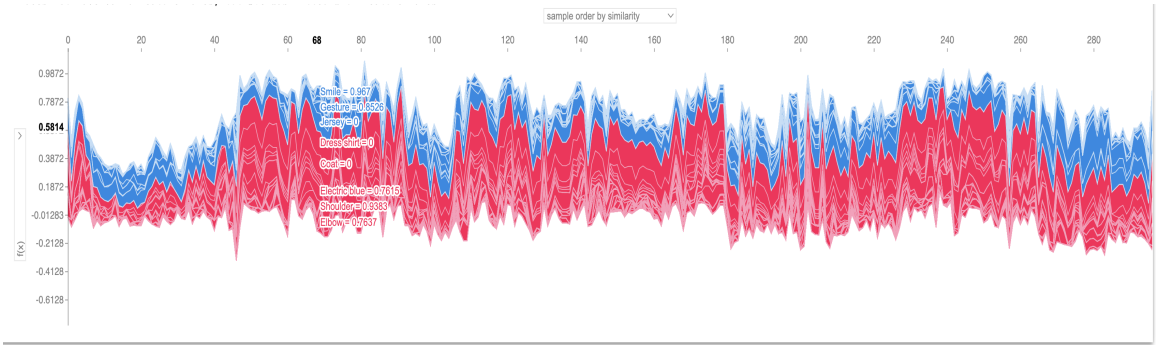


Male Pilot

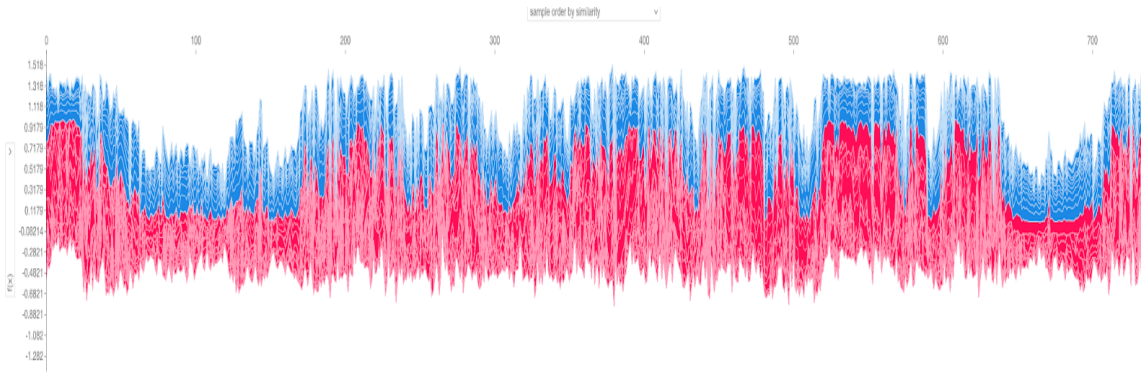


5.4.4 Force Plot

Doctors- Nurses dataset (297 test images) force plot ordered by similarity:

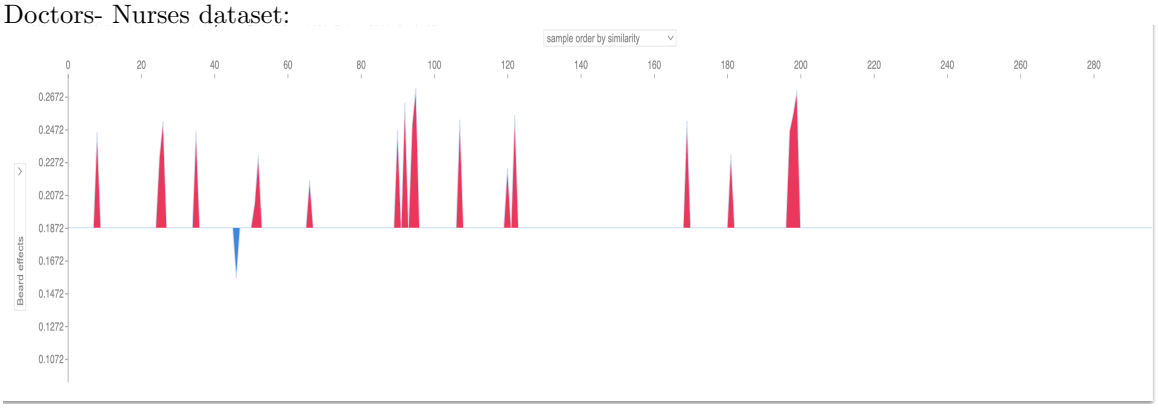


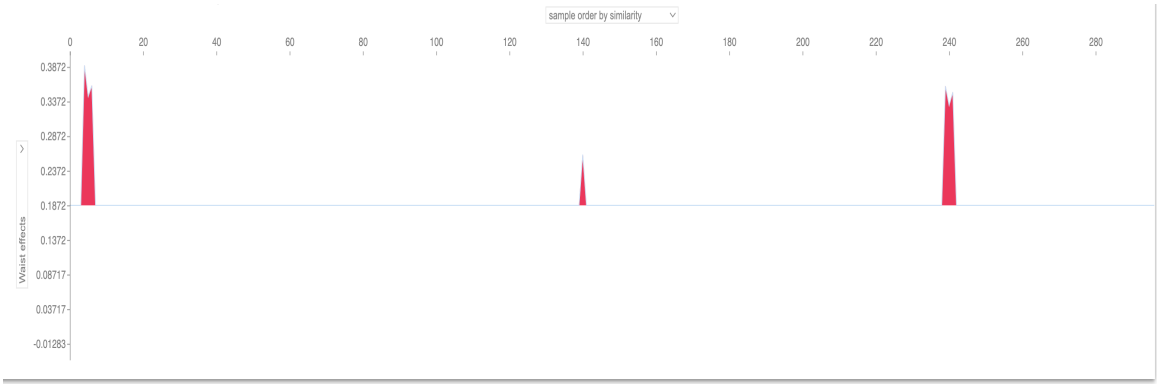
Pilots- Flight attendants dataset (734 test images) force plot ordered by similarity:



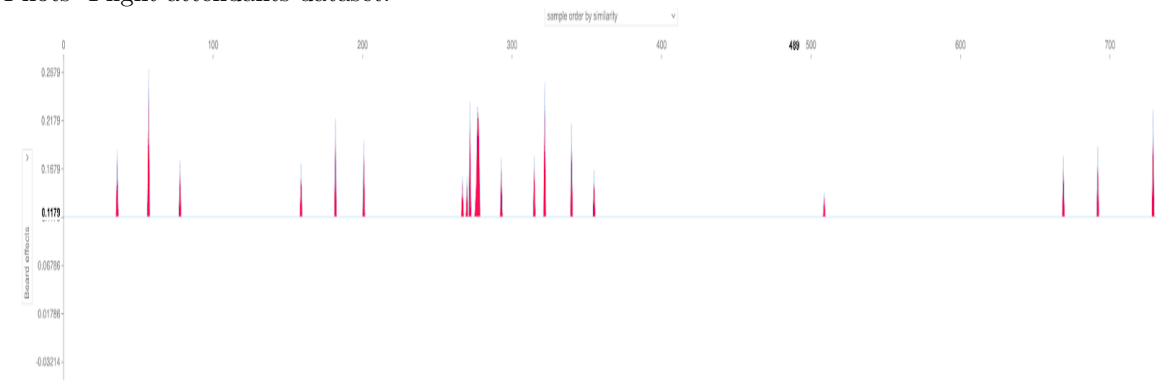
In the following force plots which show the effects of features on different datapoints, the red spikes indicate a positive contribution towards classification, while the blue spike indicates a negative contribution. We compared the effects of the features 'beard'(vs) 'waist': Beard had a positive contribution for male images irrespective of profession, while waist had a positive contribution only for female images.

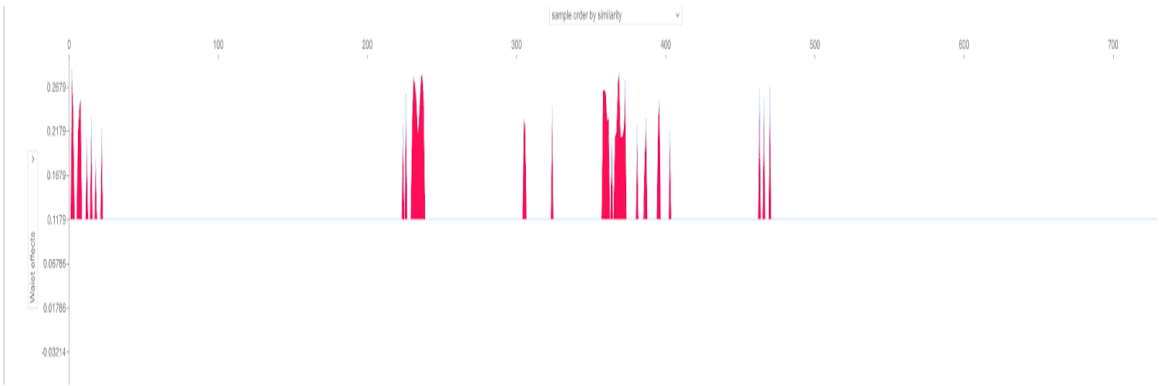
Doctors- Nurses dataset:





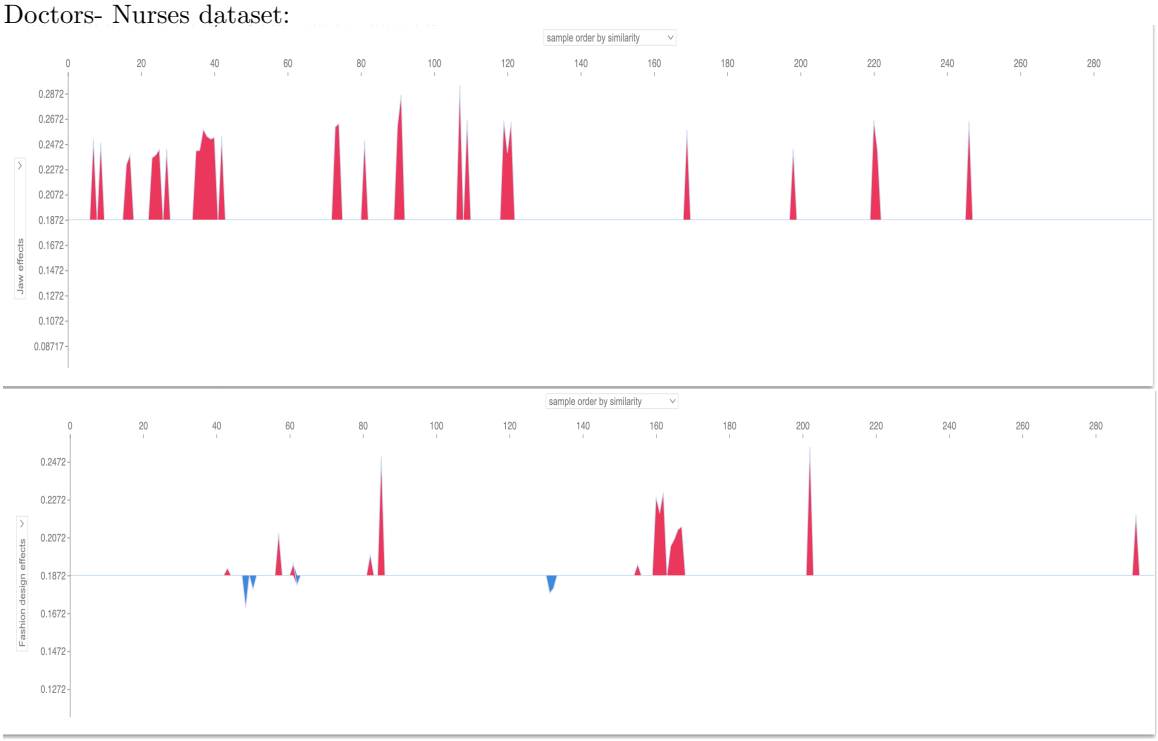
Pilots- Flight attendants dataset:



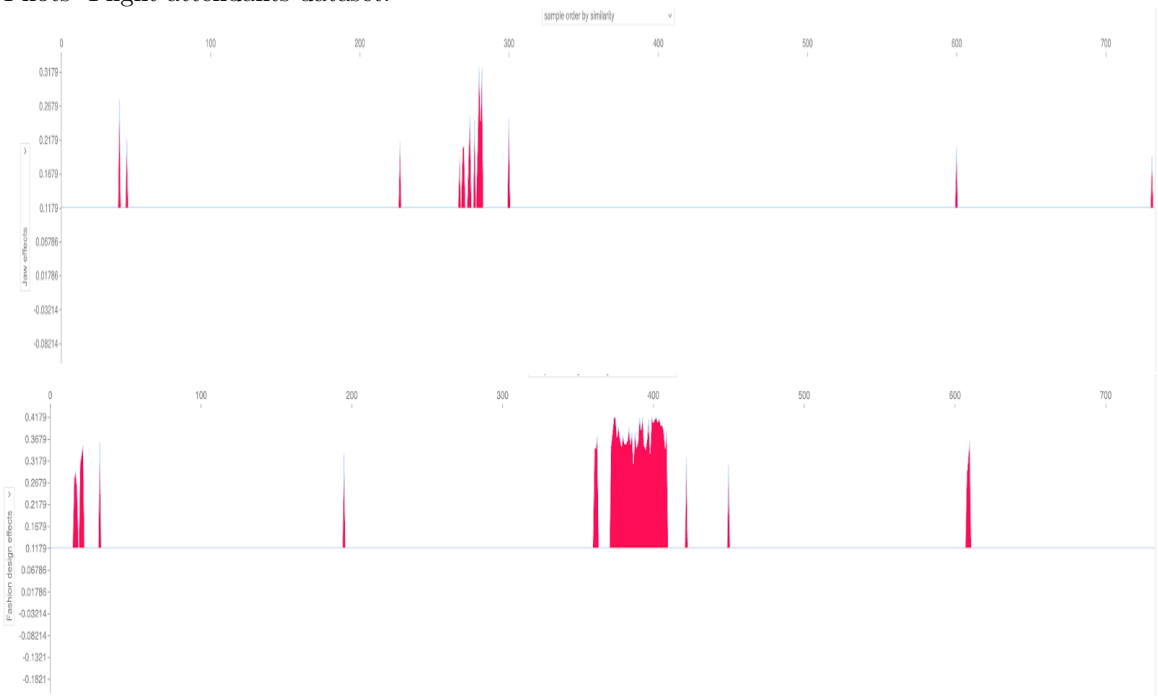


Upon comparing the effects of the features 'Jaw'(vs) 'Fashion design', Jaw had a positive contribution for male images irrespective of profession, while Fashion design not only had a positive contribution for female images, but also contributed negatively towards some of the male images.

Doctors- Nurses dataset:



Pilots- Flight attendants dataset:



6 Discussions

One of the most significant challenges we encountered in this project was assembling a gender and race-neutral dataset. From the census analysis, we can see that only 8% of the workforce is women pilots. Additionally, less than 15% of pilots come from diverse racial backgrounds, including Black, Asian, Hispanic, or other races. Similarly, the flight attendant workforce also exhibited a similar trend, with only 23% male diversity and 37% of individuals belonging to the mentioned racial groups.

Our results highlight the impact of bias on the performance of the model. It revealed that the SVM model's performance improved by 4-6% when working with biased datasets. Notably, the

model performed exceptionally well on the biased dataset representative of our census analysis, suggesting that it may better align with real-world scenarios where bias is present. It's worth noting that the dataset used in our experiment was smaller than those used in real-time decision-making processes. Additionally, we chose the SVM model due to its interpretability, whereas other models may be used in real-time. Therefore, it's essential to test how these models perform against biased datasets.

Based on the SHAP plot results included in this paper, we can clearly see how gender-stereotypical features play a very important role in image classification. Our SHAP results also showed nearly 50 other gender-stereotypical features such as lipstick, makeover, eyebrow, layered hair, thigh, jewelry, mustache, facial hair, physical fitness, and so on, that contributed highly towards classification.

These findings demonstrate the importance of creating fairer datasets and developing fairer ML models, as well as the significance of explainability. It is essential that we work towards using interpretable models to later evaluate if they perform fairly. This can increase the public's trust in large-scale AI/ML models, which are often used in decision-making processes. In conclusion, our study highlights the need for greater attention toward fairer and more transparent AI/ML models.

7 Division of work

We divided the work for this project equally between the two of us, with each of us contributing in various areas. Below is a detailed breakdown of our individual responsibilities:

Rangasri Chakravarthy:

- **Performed literature review for the following papers:** Markedness in Visual Semantic AI (2022) [18], Measuring Representational Harms in Image Captioning (2022)[17], Towards fairer datasets: Filtering and balancing the distribution of the people subtree in imageNet hierarchy (2020)[20], Fairness through casual awareness: Learning casual latent - variable models for biased data (2019)[13].
- **Doctors- nurses dataset:** Performed census analysis, Pair programmed with Smruthi to work on the following: create dictionary setup, process dataframes, split datasets with respect to bias, measure metrics of the ML model.
- **Pilots- Flight attendants dataset:** collected 855 male pilot images and 720 male flight attendant images, extracted features for these images, reached out to airline unions, performed census analysis, and worked on the dictionary setup, dataframe processing, splitting the dataset based on census bias, and measuring metrics of the ML model.
- **SHAP plots and explanations:** Implemented the (i)Waterfall plot, (ii)force plot, (iii)summary plot and (iv)dependence plots for the doctors-nurses and pilots-flight attendants dataset. Chose to use only the waterfall and force plots since they best represented all the required results.

Smruthi Pobbathi:

- **Performed literature review for the following papers:** Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces (2022)[15], Image Representations learned with Unsupervised Pre-Training Contain Human Like Biases (2021)[16], Measuring biases that matter: The ethical and casual foundations for measures of fairness in algorithms (2019)[7].
- **Doctors- nurses dataset:** Worked on the SHAPLY values plot, Pair programmed with Rangasri to work on the following: create dictionary setup, process dataframes, split datasets with respect to bias, measure metrics of the ML model.
- **Pilots- Flight attendants dataset:** collected 747 female pilot images and 765 female flight attendant images, reached out to airline unions, developed the script to extract features using API calls to Google Vision, and extracted features for these images.
- **SHAP plots and explanations:** Implemented (i)bar plot, (ii)decision plot and (iii)image plot for doctor-nurse and pilot-flight attendant datasets. Excluded image plot as the results captured by decison plot and bar plots were prominent.

References

[1] Women In Aviation (WAI). *Current Statistics of Women in Aviation Careers in U.S.* 2020. URL: https://www.wai.org/industry-stats.

[2] *Bias in computer systems.* 1996. URL: https://dl.acm.org/doi/10.1145/230538.230561.

[3] United States Census Bureau. *22 Million Employed in Health Care Fight Against COVID-19.* 2021. URL: https://www.census.gov/library/stories/2021/04/who-are-our-health-care-workers.html.

[4] DataUSA. *Aircraft Pilots and flight engineers: Detailed occupation.* 2020. URL: https://datausa.io/profile/soc/aircraft-pilots-flight-engineers.

[5] DataUSA. *Flight attendants: Detailed occupation.* 2020. URL: https://datausa.io/profile/soc/flight-attendants.

[6] Shap documetation. *Shap Documentation.* URL: https://shap.readthedocs.io/en/latest/index.html.

[7] *Measuring biases that matter: The ethical and casual foundations for measures of fairness in algorithms.* 2019. URL: https://dl.acm.org/doi/10.1145/3287560.3287573.

[8] Pilot institute. *Women Pilot Statistics: Female Representation in Aviation.* 2022. URL: https://pilotinstitute.com/women-aviation-statistics/.

[9] Pilot institute. *Women Pilot Statistics: Female Representation in Aviation.* 2022. URL: https://pilotinstitute.com/women-aviation-statistics/.

[10] *Inherent Trade-Offs in the Fair Determination of Risk Scores.* 2016. URL: https://arxiv.org/pdf/1609.05807.pdf.

[11] *It's Not the Algorithm, It's the Data.* 2017. URL: https://cacm.acm.org/magazines/2017/2/212422-its-not-the-algorithm-its-the-data/abstract.

[12] United States Department of Labor. *Labor Force Statistics from the Current Population Survey.* 2022. URL: hhttps://www.bls.gov/cps/cpsaat11.htm.

[13] *Fairness through casual awareness: Learning casual latent - variable models for biased data.* 2019. URL: https://dl.acm.org/doi/10.1145/3287560.3287564.

[14] Flying by numbers. *The TRUTH about male flight attendants!* 2022. URL: https://flyingbynumbers.com/male-flight-attendants/.

[15] *Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces.* 2022. URL: https://dl.acm.org/doi/10.1145/3531146.3533159.

[16] *Image Representations learned with Unsupervised Pre-Training Contain Human Like Biases.* 2021. URL: https://dl.acm.org/doi/10.1145/3442188.3445932.

[17] *Measuring Representational Harms in Image Captioning.* 2022. URL: https://dl.acm.org/doi/10.1145/3531146.3533099.

[18] *Markedness in Visual Semantic AI.* 2022. URL: https://dl.acm.org/doi/10.1145/3531146.3533183.

[19] *Automated Inferenceon Criminality using Face Images.* 2016. URL: https://arxiv.org/pdf/1611.04135v2.pdf.

[20] *Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy.* 2020. URL: https://dl.acm.org/doi/abs/10.1145/3351095.3375709.

[21] Aaron Young et al. *FSMB Census of Licenced Physicians in the United States.* 2020. URL: https://www.fsmb.org/siteassets/advocacy/publications/2020-physician-census.pdf.

[22] Zippia. *AIRLINE FLIGHT ATTENDANT DEMOGRAPHICS AND STATISTICS IN THE US.* 2021. URL: https://www.zippia.com/airline-flight-attendant-jobs/demographics/.

[23] Zippia. *AIRPLANE PILOT DEMOGRAPHICS AND STATISTICS IN THE US.* 2021. URL: https://www.zippia.com/airplane-pilot-jobs/demographics/.

Appendix

Short summaries of some of the papers from our literature review:

1. Markedness in Visual Semantic AI (2022) [18]

This paper evaluates state of the art visual semantics model called CLIP(Contrastive Language Image Pretraining). CLIP uses the cosine similarity between the encoded image and text to rank, retrieve classify images. The authors used FairFace dataset (relatively balanced).

What demographic information is captured by CLIP, which are assumed as the norm and left undescribed? For CLIP, the most typical concept of a "person" is: White male, aged 20-59.
Individuals more likely to be paired with a text prompt describing **gender**:

- 2 choices were given to CLIP: 1. choose a gender label described in fairface (male/ female). 2. Choose a label omitting gender (identify the photo as a person).
- Results: Given this choice between a gender label and person label, CLIP associated "person" label 26.7% times for males and 15.2% times for females. This reflects male is closer to the norm than female.

Individuals more likely to be paired with a text prompt describing **race/ethnicity**:

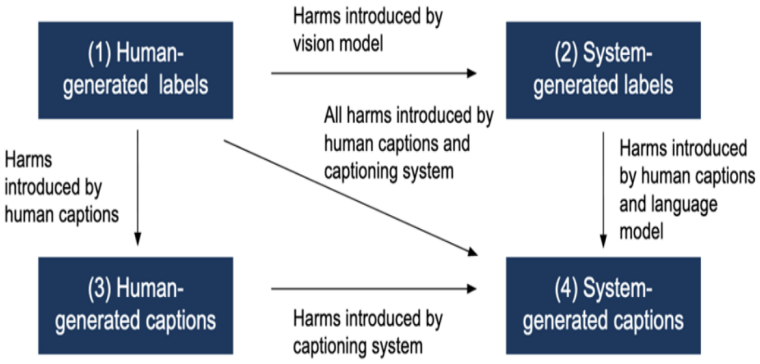
- 2 choices were given to CLIP: 1. Choose any of the 7 races described in FairFace. 2. Choose a label omitting race (identify the photo as a person).
- Results: Given this choice between a race label and person label, CLIP associated "person" label 47% times for white individuals and >2% times for East Asian/ Indian/ Southeast Asians. This reflects White is closer to the norm.

Individuals more likely to be paired with a text prompt describing **age**:

- 2 choices were given to CLIP: 1. Choose any of the 9 labels denoting age described in FairFace. 2. Choose a label omitting age (identify the photo as a person).
- Results: Given this choice between an age label and person label, CLIP associated "person" label 57.9% times for individuals aged 30-39 and >7% times for under 10 or over 70. This reflects middle age is closer to the norm.

2. Measuring Representational Harms in Image Captioning (2022)[17]
Techniques for measuring types of representational harms in image captioning systems. Types of harms discussed in this paper: Stereotype, demeaning. Image captioning system they used: VinVL. Measurement approach to find out where harm arises:

- Datasets: COCO, Conceptual captions.
- Generated labels (single words like sky, ball, etc) and captions (full sentences)
- rely on 4 sets of data annotation: Human-generated labels, System-generated labels, Human-generated captions, and System-generated labels.
- Compared the results of all 4 sets of annotations to know where bias/ harm arises.



Here, stage 1 (human-generated) labels are ground truth and stage 4 system-generated captions are measured based on stage 1.
Stereotyping: Captioning that incorrectly includes words

- Example: the word "Gun" included in the caption for an image with a black person in it racial stereotyping.
- Consider a non-imaginable concept like "Vegetarian" or "evil".

- Consider a concept that's too specific like "Field hockey stick" instead of "hockey stick".
- Consider an imageable concept that is not depicted in an image (Hallucinated imageable concept).
- Captions like this would require extra information, which would include some stereotypes.

Solution: use heuristic to rand incorrect mentions of words.
Demeaning: Captions that differ in whether people depicted in images are mentioned or not (i.e., the system omitting the presence of certain people)

- Example: people belonging to certain social groups mentioned in the captions more or less times than others ⇒ dehumanization.
- People of darker vs lighter skin tones: system generated captions yielded the largest difference in captioning them.

3. Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces (2022)[15]
In this paper, the authors focus on detection of bias for facial expressions. This is briefly what they do in the paper.

- They have posted the effective computing datasets which have a focus on Action Unit detection for missing information about metadata like age, gender, ethnicity, glasses and beards.
- Datasets used - AffWild2, BP4D, BP4D+, CK+ DISFA, DISFA+, GFT. UNBC and ERIK.
- They analyze the resulting meta data annotation distribution in the datasets.
- They evaluate susceptibility of 2 modern AU detection algorithms regarding bias - OpenFace and NISL2020 are the models.
- They evaluate the same algorithms regarding susceptibility to bias in their output for categorical emotions.

OpenFace is an opensource toolkit with different facial analysis functionalities, like eye gaze tracking or facial landmark detection. NISL2020 is a multi-task model for valence/arousal (VA) estimation, AU detection and expression classification. OpenFace is tested on - AffWild2, ERIK and GFT. But are trained on different datasets. NISL2020 is tested om AffWild2, BP4D, BP4D+, CK+, ERIK. GFT and UNBC. But trained on remaining datasets. Since both the models are trained on different datasets, the authors do not compare the results. They obtain bias for age, ethnicity, gender and glasses.

The authors observe that the bias for each taxonomy is different that is Action Units are better detected in male subjects and emotions in female subjects. However, a consistent difference in bias susceptibility was not found between the annotation styles for the data and algorithms in question. Bias itself was not a problem as long as the model is used in a suitable environment: Using a model which detects facial expressions well in middle-aged subjects in an environment with children is not advisable. At the same time it would be perfectly suited when working with university students. However, training data usually did not favor edge cases and findings correctly biased models for these groups may proof to be difficult. Either way, they conclude that for the purpose of these decisions information on performance depending on different use-case scenarios was not available. Bias susceptibility was a problem and they conslude by mentioning the changes that was necessary for the dataset itself.

4. Image Representations learned with Unsupervised Pre-Training Contain Human Like Biases (2021)[16]
In this paper, the authors aim to answer the question "Do unsupervised computer vision models automatically learn implicit patterns and embed social biases that could have harmful downstream effects? ". They develop a method to quantify biased assocuations between reperesentations of social concepts and attributes in images. They use the unsupervised models trained on Imagenet, and find that these models learn racial, gender and intersectioanl biases. They replicate 8 previously documented human biases from social psychology, from the innocuous, as with insects and flowers, to the potentially harmful, as with race and gender. Their results closely match three hypotheses about intersectional bias from social psychology. They also quantify implicit human biases about weight, disabilities, and several ethnicities. They compare their findings with statistical patterns in online image datasets, and conclude that learning models can automatically learn bias from the way

people are stereotypically portrayed on the web.

Approach: Similar to contextualized word representation task in language transformers, image transformer model iGPT generates image representation to solve the next pixel prediction task. To address the abstraction of semantic representation in image domain, they use Image Embedding Association Test (iEAT). The authors then provide different sets of stimuli as an input and train the unsupervised models iGPT and SimCLR and obtain the image embedding. These image embeddings are then tested with iEAT to obtain the results.
Results The practitioners perform experiments for widely accepted biases, racial biases, gender biases and other biases. The SimCLRv2 embeddings contain stronger biases than iGPT embeddings for widely accepted biases. They also observe that both embeddings in the case of racial bias, associate white people with tools and black people with weapons. For gender bias test, iGPT model displays significant bias in associating with male to science attributes and female to liberal arts attributes. The other bias tested was to replicate weight steryotypes. The iGPT embedding display additional bias towards thin people with pleasantness and overweight people with unpleasantness.

5. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in imageNet hierarchy (2020)[20]
The AI/ML community talks a lot about the cycle where: human bias and prejudice gets into ⇒ large-scale datasets, which are used in ⇒ AI/ML models, which are deployed for ⇒ widespread decision-making processes. In computer vision, datasets are being called out for underrepresenting specific groups, which perpetuates gender stereotypes, and for overrepresenting different countries.

This paper examines the ImageNet dataset (21,000+ labels, more than 14 million images) used in various computer vision models. ImageNet was initially created for the task of object recognition. When ImageNet was first constructed, most of the image recognition work was done on Caltech101, PASCAL VOC, etc. Because of the volume of images the ImageNet dataset contains, an automated pipeline had to be constructed involving large-scale crowdsourcing with thousands of workers.

This paper looks into the 'person' subtree of the ImageNet dataset, which has 2832 labels. All of these labels came from WordNet, a hierarchical ontology of English nouns constructed by linguists.

The 3 different problematic behaviors with the person subtree of the ImageNet dataset are discussed below:

1. Stagnant concept vocabulary: there exist some potentially problematic labels like 'racist' or 'atheist', which are considered legitimate English nouns in WordNet but become problematic when used as a label for an image in ImageNet.

Solution: identify and annotate labels defined as "unsafe", "offensive" (containing profanity, like racial slurs), or "sensitive" (might cause offense when applied inappropriately). This annotation was done by 12 grad students, which resulted in 1593 out of the 2832 labels identified as unsafe. So only 1239 labels were considered safe.

2. Non-visual concepts: some concepts might be safe but are difficult to characterize visually using images (like a vegetarian, philanthropist, etc.). While some of these terms may be positive on their own, they might lead to possible stereotypes when trying to illustrate these words with images. Labels like these were in the ImageNet dataset because the images were collected by search engines (google, yahoo, etc.) and manually verified this by asking crowd workers to select all images representing a word. This did not accurately reflect a concept like philanthropist.

Solution: counteracting the search engines and annotation artifacts, explicitly annotate the imageability of the person labels. Instead of using crowd workers, they used Amazon Mechanical Turk to annotate the imageability. From this, they found out that out of the 1239 safe labels, only 158 were identified as imageable.

3. Lack of image diversity: The images themselves may represent a biased representation of the demographics. These images were taken from search engines, which are known for retrieving biased results in terms of demographics.

Solution: annotated demographics for images from the 'safe' and 'imageable' labels. The authors included 3 specific annotations: Gender, skin color, and age. Two interesting inferences were made, where the gender of scuba divers and newborn babies were difficult to identify.

Unions we reached out to for Pilot-flight attendant image collection:

1. Airline Pilots Association (ALPA)
2. Association of Flight Attendants (AFA)
3. Allied Pilots Association (American pilots)
4. Association of Professional Flight Attendants (APFA)
5. International brotherhood of teamsters
6. Association of Flight Attendants-Communications Workers of America (AFA-CWA)
7. Coalition of Airline Pilots Association (CAPA)
8. Canadian Airline Dispatchers Association (CALDA)
9. International Association of Machinists and Aerospace Workers (IAM)
10. Australian and International Pilots Association (AIPA)
11. Syndicat National des Pilotes de Ligne (France)(SNPL)
12. Professional Air Traffic Controllers Organization (PATCO)
13. US Airways Pilots Association (USAPA)

Pilots-Flight attendants image sources:

1. Image scraping on Instagram, Facebook, LinkedIn, Pinterest
2. Google images with keyword specific search: List of airlines (`https://www.nationsonline.org/oneworld/major_airlines.htm`)
3. Getty images pilots board: (`https://www.gettyimages.com/collaboration/boards/0Afm6FBPKUO_a-qqL6MnbA`)
4. Getty images Flight attendants board: (`https://www.gettyimages.com/collaboration/boards/B8gwd_CKWkarbAPiHK6Hcg`)
5. Facebook group: (`https://www.facebook.com/LifeOfCabincrew007/photos/?ref=page_internal`)
6. Facebook group: (`https://www.facebook.com/groups/373474713289026/media`)
7. Facebook group: (`https://www.facebook.com/aflyguyslounge/photos`)
8. Facebook group: (`https://www.facebook.com/jetairways/photos/?ref=page_internal`)
9. Facebook group: (`https://www.facebook.com/groups/365245849098051/media`)
10. Facebook group: (`https://www.facebook.com/groups/664149708158542/media`)
11. Facebook group: (`https://www.facebook.com/groups/1456527894668407/media`)
12. Facebook group: (`https://www.facebook.com/aflyguyslounge`)
13. Insider article: (`https://www.insider.com/flight-attendant-uniforms-around-the-world-2018-1#aeroflot-a-russian-airline-has-bright-red-vintage-inspired-uniforms-1e`)
14. Hawaiian Airlines: (`https://www.hawaiianairlines.com/hawaii-stories/hana-hou/all-aboard`)
15. Hawaiian Airlines: (`https://www.hawaiianairlines.com/careers/flight-attendants`)
16. ALN News article: (`https://aerolatinnews.com/industria-aeronautica/female-lufthansa-pilots-are-taking-off/`)
17. Group one air website: (`https://www.grupooneair.com/female-aircraft-pilots/`)
18. Confessions of a Trolley Dolly article: (`https://confessionsofatrolleydolly.com/2021/01/08/top-ten-cabin-crew-uniforms-2020/`)
19. Pinterest: (`https://in.pinterest.com/pin/46584177374683723/`)

20. Pinterest: (https://in.pinterest.com/search/pins/?q=male\%20flight\%20attendant&rs=typed/)

21. Pinterest: (https://www.pinterest.nz/stevemarchant5/male-flight-attendants/)

22. Shutterstock Images: (https://www.shutterstock.com/search/male-flight-attendant?c3apidt=p35138749781&gclid=CjwKCAiAqt-dBhBcEiwATw-ggEsdJs56cVQptA21r7nBIADaXz-GQUqSkgZlzsCjNYJEd8r8r5gRixoC-yYQAvD_BwE&gclsrc=aw.ds&kw=\%2Bdomain+\%2Broyalty+\%2Bfree+\%2Bphotos)

23. Stock Adobe Images: (https://stock.adobe.com/search?k=flight\%20attendant\%20male)

24. Deposit photos Images: (https://depositphotos.com/stock-photos/emirates-flight-attendant.html)

25. Business Traveller Article: (https://www.businesstraveller.com/business-travel/2019/07/25/japan-airlines-unveils-new-uniforms-for-2020-and-special-livery-for-tokyo-olympics/)

26. Forbes Africa Article: (https://www.forbesafrica.com/woman/2018/12/24/in-pictures-south-africas-first-black-female-helicopter-pilot-for-saps-uplifts-young-women/)

27. ABC News Article: (https://abcnews.go.com/Politics/diversifying-flight-deck-us-pilots-black-women/story?id=72880810/)

28. USA Today News Article: (https://www.usatoday.com/story/travel/flights/todayinthesky/2019/01/16/united-airlines-new-uniforms-70-000-workers-get-new-look/2596582002/)

29. Sisters of the sky organization: (https://sistersoftheskies.org/)

30. Family on Standby article: (https://familyonstandby.com/what-is-non-rev-travel-breaking-it-down-for-newbie-standby-travelers/)

31. Better Aviation: (https://betteraviationjobs.com/job/saudia-airlines-male-cabin-crew-recruitment-february-2019/)

32. The Design Air article: (https://thedesignair.net/2020/02/07/saudia-launches-new-uniforms-trialling-on-london-and-paris-routes/)

33. The Design Air article: (https://thedesignair.net/2016/02/05/xiamenair-introduce-new-modern-classic-uniforms/#jp-carousel-10522)

34. (https://www.superadrianme.com/fashion/new-etihad-airways-uniforms/)

35. Pyok article: (https://www.paddleyourownkanoo.com/2019/04/29/etihad-becomes-first-airline-in-the-middle-east-to-allow-male-cabin-crew-with-beards/)

36. Pyok article: (https://www.paddleyourownkanoo.com/2021/10/05/the-top-10-flight-attendant-uniforms-gracing-the-skies-in-2021/)

37. South China Morning Post article:(https://www.scmp.com/news/world/europe/article/3205151/uk-require-negative-covid-test-china-travellers-media-reports-say)

38. (https://www.superadrianme.com/fashion/new-etihad-airways-uniforms/)

39. Global Times article: (https://www.superadrianme.com/fashion/new-etihad-airways-uniforms/)

40. Flicker Images: (https://www.flickr.com/search/?text=male\%20flight\%20attendant)

41. Flicker Images: (https://www.flickr.com/photos/airlinepilotsassociation/albums)

42. China Aviation Daily: (http://www.chinaaviationdaily.com/news/50/50500.html)

43. Medium article: (https://medium.com/the-isthmus/females-can-fly-too-15c41516731)

44. Russia Beyond article: (https://www.rbth.com/lifestyle/333381-russian-female-pilots)

45. CAE-Women in flight document: (https://www.cae.com/media/documents/Civil_Aviation/women_in_flight/CAE_FlightGlobal_ContentPartnership_Digital.pdfs)

46. Latin Flyer website article: (https://latinflyer.com/real-runway-mexicos-hottest-flight-attendant-uniforms/)

47. Today news article: (https://www.today.com/style/finally-female-flight-attendants-win-right-wear-pants-after-2-t73196)

48. Racked article: (https://www.racked.com/2016/7/28/12235510/flight-attendant-uniforms-history)

49. Stuff NZ article: (https://www.stuff.co.nz/travel/themes/101106077/pilot-and-flight-attendant-uniforms-the-meaning-behind-the-outfit)

50. Southern Living article: (https://www.southernliving.com/fashion-beauty/vintage-flight-attendant-uniform)

51. Simple Flying Blog: (https://simpleflying.com/boeing-777x-test-aircraft-flight-history/)