Mahek Patel

Data In Context

April 28th, 2024

Final Project: Auditing an algorithm

Part II: Project Report

Betaface API Facial Recognition Service

In this report, I cover the findings from my group's chosen Betaface API, facial recognition service that allows users to upload images for analysis, identifying attributes such as gender, race, and additional metadata for each image. In the project, my group used the Betaface facial recognition API online demo that allows users to detect faces with “2 basic and 101 pro facial points [,] supports functions such as uploading image files [,] retrieving image and face metadata[,] comparing single faces or groups of faces[,] transforming face images[, and] performing fast searches in large face collections”(Betaface Demo). My group used the AI-powered random face generator tool to generate and collect 90 diverse images encompassing a wide variety of races, ages, and genders. For the purpose of our analysis, we chose to focus on three variables, race, gender, and age, in the detection results. With our 90 randomly generator images uploaded to Betaface, the API analysed each image, classified “gender, age, ethnicity, and emotion (smile/neutral)” (Betaface Demo), and provided us with a match percentage comparing the detected attributes to the original ones assigned by the generator. In the further sections of this report, I will delve deeper into the performance evaluation of Betaface's facial recognition service, identify points of concern, and suggest future directions for improvement.

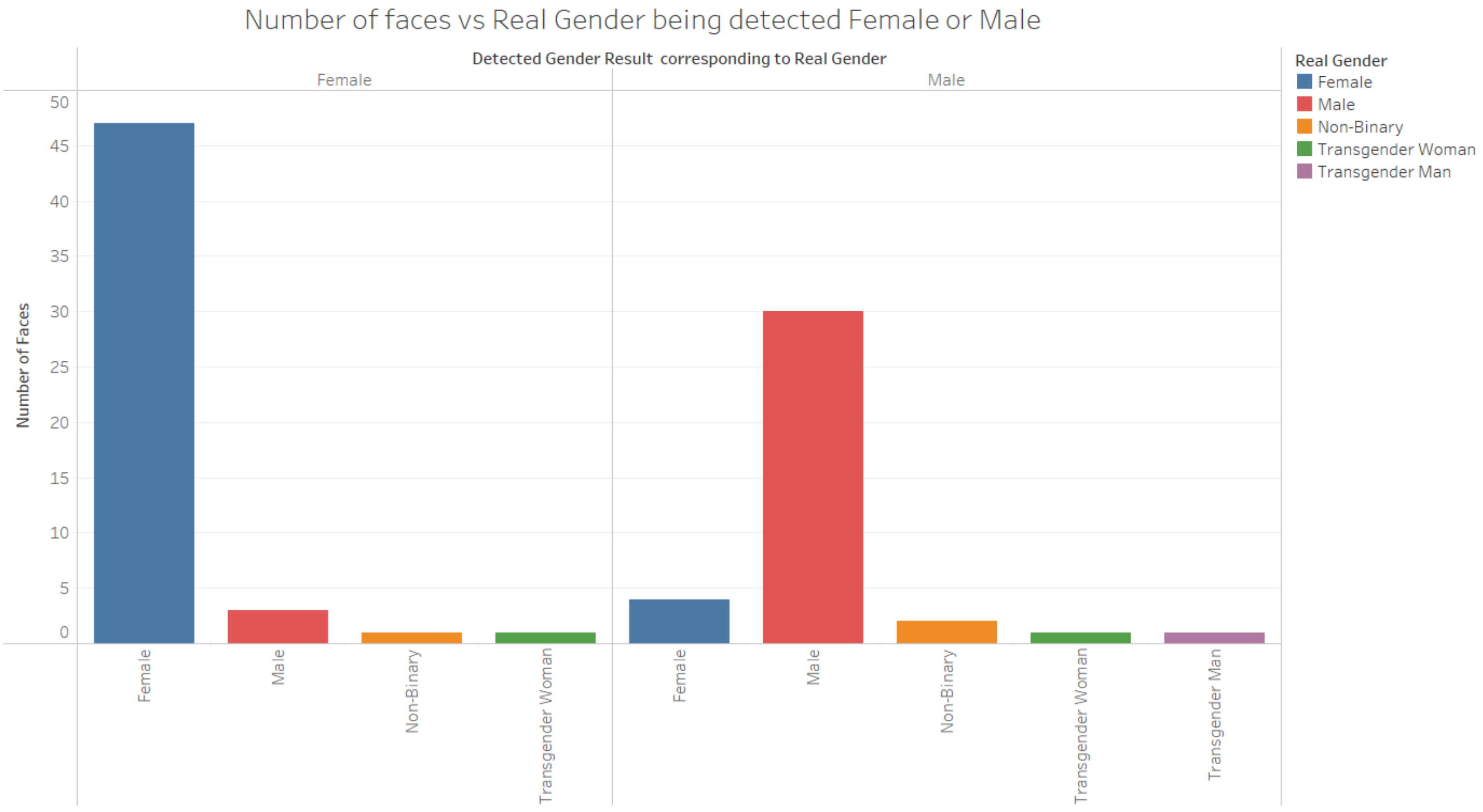
Analysis of differences across demographic groups and API performance

Data In Context

April 28th, 2024

Examining the data, there are noticeable differences observed in the average quality of results across demographic groups, with females having a higher detection quality (7.02), then males (6.84), and white having the highest average quality of score results (7.19) among races/ethnicities. The bar graph (plot 1 on next page) shows the number of faces detected as Male and Female against their real gender. It indicates a high accuracy in detecting females (47 out of 52 correct) and males (30 out of 38 correct), but also highlights significant errors particularly in non-binary and transgender categories.

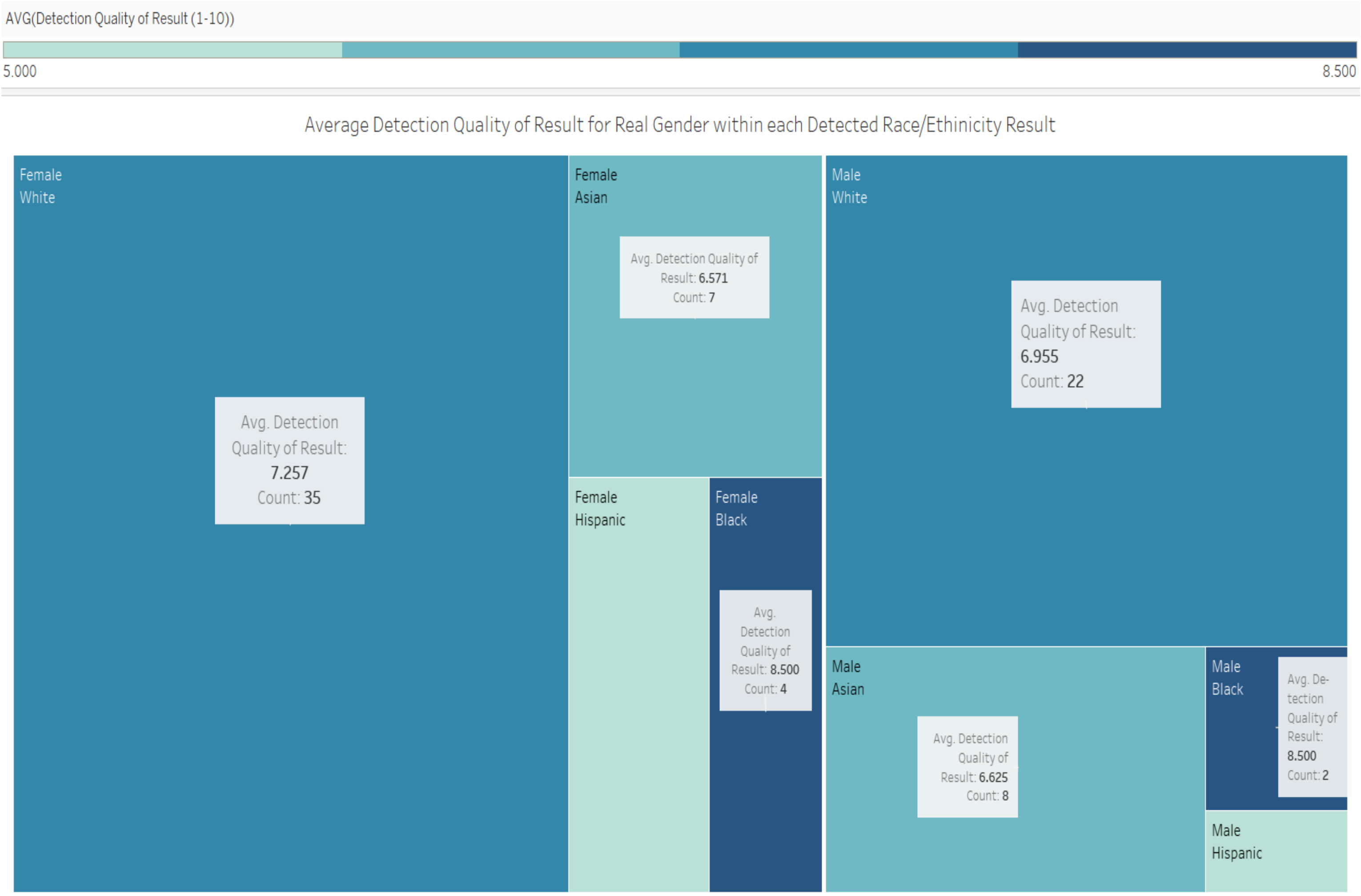
Plot 1:



Count of Sheet1 for each Real Gender broken down by Detected Gender Result . Color shows details about Real Gender.

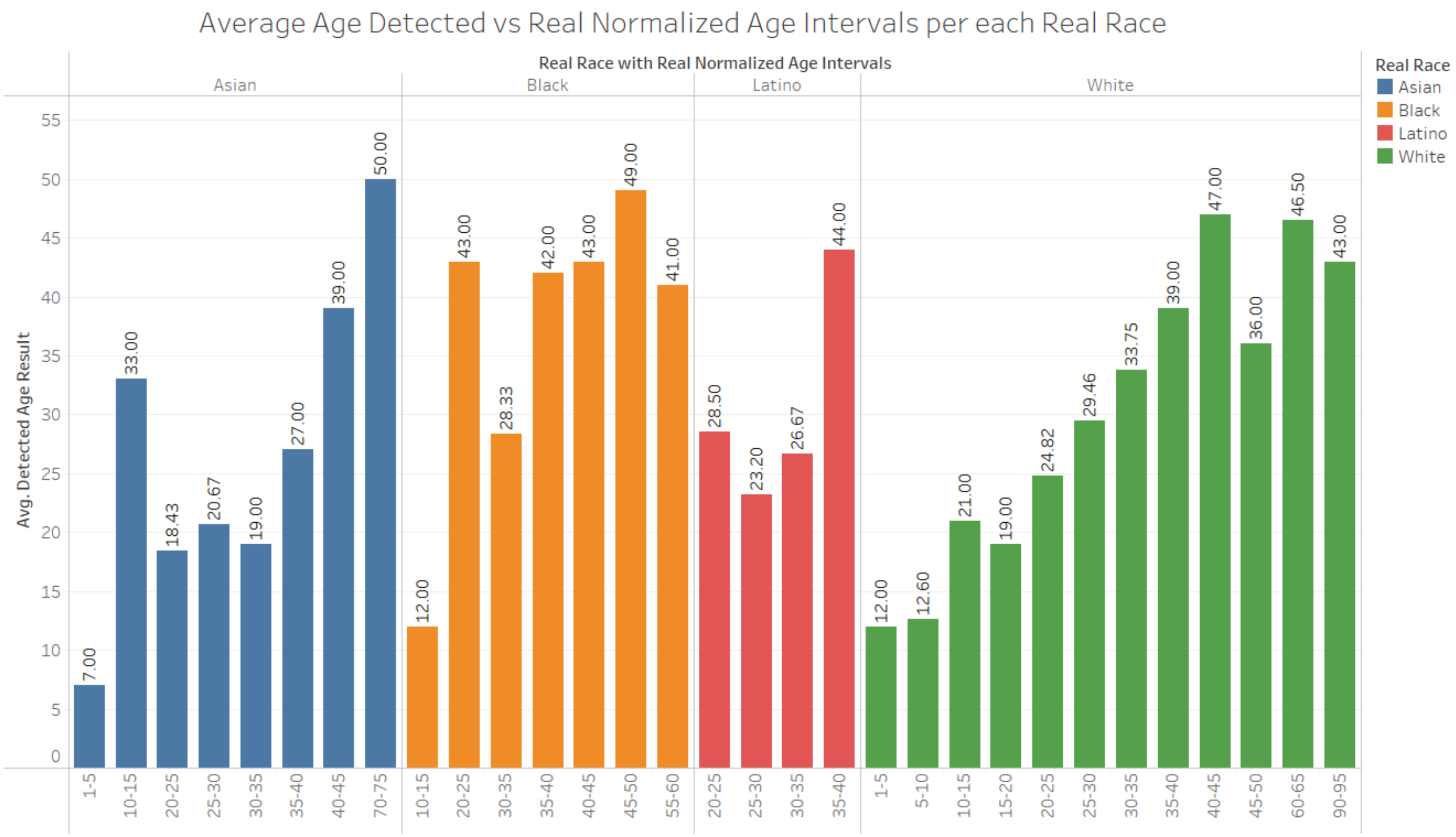
To check the detection accuracy of race within dominant genders seen from plot 1, Female and Male, I have used a tree map in plot 2 below displaying larger sizes for more number of faces and color gradients representing the quality of detection. Here, Black females and males show the highest detection quality scores, yet the number of faces/data for Black race is much lower compared to White counterparts.

Plot 2:



To examine the trend better, the bar plot below correlates real age groups with detected ages across various races, revealing significant discrepancies, particularly in the Asian demographic where younger ages are often overestimated, and older ages were often detected young. The least discrepancy can be seen among the White race age groups than other races.

Plot 3:



Average of Detected Age Result for each Real Normalized Age Intervals broken down by Real Race. Color shows details about Real Race.

Based on the visualizations and observations, one of the areas of good performance by API was a high gender detection accuracy for Female and Male genders. Both plots 1 and 2 suggest robust performance in accurately classifying traditional male and female genders. Females show a particularly high true positive rate, and males, while slightly lower, still exhibit good accuracy. Secondly, from plot 2, it can be observed that White individuals receive relatively high average detection quality scores (7.3 for females and 7 for males), indicating a well-trained model for these demographic categories. Black individuals, although represented less frequently, also receive high-quality scores, suggesting effective, albeit less faces/data, accuracy. There are clear areas of bad performances of the API seen through the detection results. In plot 1, it can be seen that the API struggles with gender diversity, as non-binary and transgender categories have frequent misclassifications to Female and Male genders, which indicates a lack of adequate training data representing these groups and lack of reliability. In addition, this shows lack of diversity in gender category, indicating the biased training dataset used not only in Betaface API but also the AI random face generator. Another factor showing bad performance is the inconsistent racial detection in minorities. Although some minority groups like Black individuals sometimes receive high detection scores (as seen in plot 2), the inconsistent detection quality of results scores across other minority groups such as Asians and Hispanics suggest an uneven performance indicating less external validity and reliability. Moreover, plot 3 highlights considerable inaccuracies in age detection, particularly among Asians, where ages are consistently underestimated. For 1 South Asian face observed with darker complexions, features like dark under eyes could be potential factors classifying them as higher age while the majority East Asian, missing features like South Asian, could be factor leading to 70-75 age group being detected as 50. The lack of intersectionality in the data representation and the apparent model bias towards "white-passing" features suggest that the API may not only be less accurate for certain demographic groups but also potentially perpetuate stereotypes. The lack of accurate recognition for non-binary and transgender individuals, combined with racial and age biases, points towards a skewed training dataset emphasizing the need for algorithmic auditing in Betaface as well as AI face generator too.

Recommendations for Mitigating Concerns:

With the noticable performance variations across different demographic groups in the Betaface API's detections, several recommendations can be made to address concerns about the reliability, validity, and fairness. Firstly, enhancing algorithmic fairness, implementing bias mitigation techniques such as re-weighting training data can help correct skewed outputs like seen in Asian age detection, while improving validity and reliability. Diversifying training datasets can promote data "pluralism[, improving external validity, and] rethink binaries and hierarchies", as suggested by D'Ignazio and Klein (Week 14 Slides), by integrating a broader spectrum of data, especially groups that the Betaface currently underrepresents like Hispanics, Non-binary, Transgender not included in training the current BetaFace API. Secondly, conducting regular algorithm audits, as highlighted by Milan & van der Velden (Week 14 Slides), can "proactively" identify biases and inaccuracies. In addition, maintaining transparency through publishing transparency reports is crucial to address power imbalances, a component lacking in both the Betaface API and AI random generator. Thirdly, as Buolamwini and Gebru mention, to develop "machine learning systems, we need to develop intersectional training data sets, intersectional benchmarks, and intersectional audits" (Costanza-Chock, 2020, p. 20) allowing to understand how overlapping identities such as race and gender affect facial recognition accuracy. Developing an "intersectional error analysis [targeting] gender classification performance on [poorly represented groups like] darker and lighter" (Buolamwin and Gebru, 2018, p. 11) among genders and Asian age group 50-75 (as in Plot 3) could be done to attain algorithmic fairness and avoid single-axis algorithmic bias audits. Adopting comparative and non-comparative justice measures is crucial by ensuring equitable error rates across all demographic groups prevents any single group from bearing the burden of inaccuracies disproportionately, in cases like the COMPAS criminal justice system algorithm with racial bias, leading to higher rates of false positives of Black defendants as in film Coded Bias (2020) (Week 14 Slides). Implementing proactive inclusion strategies by prioritizing enhancements for historically marginalized or underrepresented groups seen, Hispanic, Black and Asian races, helps correct systemic imbalances in datasets. Finally, enhancing validity and reliability through continuous performance monitoring and robust user feedback mechanisms will help gather insights on performance and areas of improvement. Cross-validation with external data by regularly testing the algorithm with datasets not used in the training phase can also show how well the algorithm can generalize to new unseen data. These measures align with the principles of data activism and justice, ensuring fairness, accuracy, and inclusivity in Betaface's facial recognition API.

Dataset limitations and challenges in analysis:

Looking back at the overall data analysis, it can be observed that my group faced considerable challenges in creating a dataset aimed at capturing a diverse representation of demographics. The AI face generator tool exhibited a lack of diversity with fewer dark skin tone faces and a combination of feminine and masculine facial features, posing classification challenges despite our efforts. Another significant challenge was the subjective estimation of individuals' ages, which was inherently guesswork and prone to error, potentially leading to misclassifications. Additionally, BetaFace API only provided binary gender options which restricted our ability to accurately represent non-binary and transgender individuals, despite our efforts to include these groups based on perceived gender cues from the AI-generated facial features. This issue was further compounded by Betaface's inability to provide detailed racial classifications, forcing us to simplify our categories into broad groups like White, Asian, Black, and Hispanic, which might have resulted in further misclassification—such as conflating Middle-Eastern individuals with other racial groups. The quality of result scoring was also subjective, relying heavily on individual interpretation within predefined guidelines, while aimed at consistency, still introduced a level of bias and variability into our analysis. These methodological constraints highlight the complexities and potential inaccuracies in constructing a dataset that truly reflects demographic diversity and indicate the need for more sophisticated analytical tools and more nuanced classification system in Betaface facial recognition API. Reflecting on the findings covered, I think it is crucial for future research to address these limitations by developing more inclusive data collection methods that can more accurately capture and represent the rich diversity of human identities.

Data In Context

April 28th, 2024

References

"Betaface | Advanced Face Recognition." *Betaface*, 2014, www.betaface.com/wpa/.

"Betaface Free Online Demo - Face Recognition, Face Search, Face Analysis." *Betaface*, 2014, www.betafaceapi.com/demo.html.

Buolamwini, Joy & Gebru, Timnit. "Gender Shades Intersectional Accuracy Disparities in Commercial Gender Classification (2018)." Canvas, uploaded by Jonathan Bullinger, 12 Jan. 2022, https://canvas.rutgers.edu/.

Costanza-Chock, Sasha. "Introduction to Design Justice (2020)." *Canvas*, uploaded by Jonathan Bullinger, 12 Jan. 2022, https://canvas.rutgers.edu/.

Week 14 slides. *Canvas*, uploaded by Jonathan Bullinger, 12 Jan. 2024, https://canvas.rutgers.edu/.