Everypixel API Audit (Mock Report)

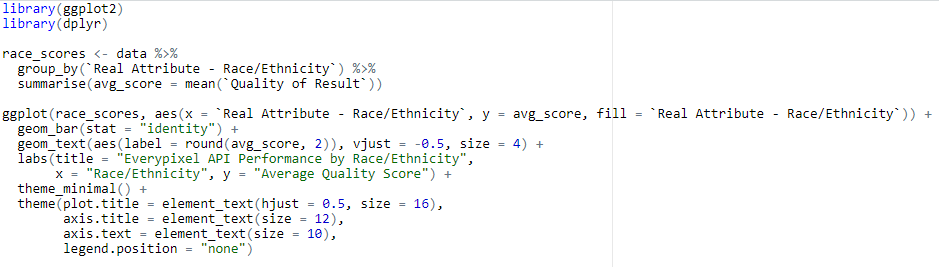
Over the last decade, facial recognition technologies have become increasingly used in a wide range of applications ranging from security systems to social media. These technologies rely on a large dataset of faces to train an artificial intelligence model using a process called machine learning. Next, complex mathematical algorithms are used to detect and analyze human faces in still images. The output of such technology can then be interpolated to assign demographic attributes such as age, gender, and race to specific faces. However, as the use of facial recognition technologies has grown, so have concerns regarding the accuracy and bias of these technologies across different demographic groups. Studies have shown that facial recognition algorithms can potentially have significant disparities in performance when analyzing different populations, with higher error rates being detected for people of color, women, and younger or older individuals when compared to middle-aged white men. These biases can have serious consequences, particularly in critical applications such as law enforcement and immigration. This report contains an audit that analyzes the performance of the Everypixel facial recognition API using a dataset of AI-generated faces from Generated Photos of Generated Media, Inc. The goal of the audit is to assess the accuracy and fairness of the algorithm across multiple different demographics. For this audit, the demographics chosen are age groups, genders, and races/ethnicities. The report will also include potential solutions to minimize the risks of low accuracies and high biases.

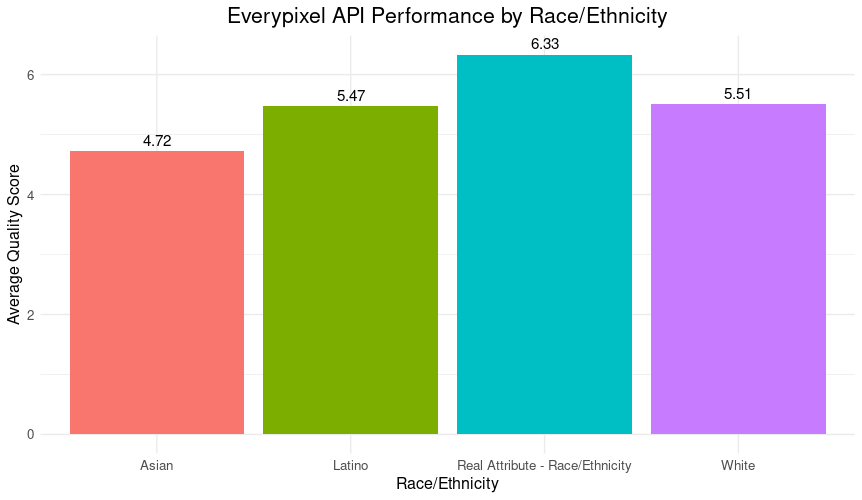
To evaluate the Everypixel API, a diverse dataset of human faces is first collected. The Generated Photos service was used to collect 120 different AI-generated images of faces, and each image had with it the demographic parameters used when prompting the AI to generate the image. The service currently only supports the generation of four major ethnic groups, these being White, Black, Latino, and Asian. To collect a batch of images, a script was written using JavaScript to scrape the images off of the website directly. However, even though the images presented were completely chosen at random, no images belonged to the Black ethnicity group.

With our dataset collected, the Everypixel API was then used to analyze the attributes of each face. Everypixel offers image recognition as a service through a cloud-based API. Images can be sent to the API to receive predictions about the image's content, including the presence and attributes of human faces. For each face in the dataset, the Everypixel API was used to predict the age, gender, and race/ethnicity. Based on the API’s response, our team generated personal confidence scores using nuanced judgment and bias, indicating the algorithm's certainty in each prediction.

To assess the accuracy of Everypixel API's predictions, the results were first compiled into a spreadsheet along with the real attributes of each face, as provided by Generated Photos. The quality of each prediction was then manually rated on a scale from 1 to 10, with 10 representing a perfect match between the predicted and real attributes. This quality score serves as a estimated measure of the API's performance per each face. By averaging the scores across different demographic groups, areas where the algorithm performs well and where it struggles can be identified.

The analysis conducted below revealed notable differences in Everypixel API's performance particularly across different ethnicities. To better understand the performance differences across this demographic metric, a bar chart was created showing the average quality score for each category. This chart was generated using R, a major programming language used for statistical analysis and data visualization.





As seen above, there are large disparities in the API's accuracy. The "Real Attribute" category represents faces that were accurately classified by race/ethnicity, and thus has the highest average quality score of 6.33. This suggests that when the API correctly identifies the race/ethnicity for a face, it also performs well when identifying all the other demographic metrics. Among the specified ethnic groups, the API performs best on White faces, with an average quality score of 5.51. Latino faces have the second-highest score of 5.47, indicating a relatively good performance for this group too. However, the API struggles most with Asian faces, which have a substantially lower average quality score of 4.72. This discrepancy raises concerns about the API's ability to accurately analyze and classify faces of Asian faces; the gaps in performance across these ethnic groups show potential biases and limitations in the API's training data and algorithms. The lower accuracy for Asian faces in particular suggests that the API may not have been sufficiently trained on a diverse range of Asian faces, leading to poorer performance for this demographic. This shows the high importance of checking facial recognition systems for fairness.

These disparities in performance also raise concerns about a concept known as algorithmic fairness. Algorithmic fairness refers to the principle that AI systems should treat all individuals equitably, regardless of their demographic characteristics. The Everypixel API's varying accuracy across ethnic groups suggests that the algorithm may be perpetuating biases and discriminating against certain populations. Furthermore, the disparities in performance for the API are detrimental to its quality as a service, and can have real-world consequences. For example, in a law enforcement context, a facial recognition system that is less accurate for certain demographics could lead to false identifications and eventually wrongful arrests. This occurred in the case of Robert Williams, an individual who was linked to a crime that they did not commit through the use of police facial recognition tools (Williams 1). Comparative justice focuses on the fairness of outcomes between different metrics; the lower accuracy for Asian faces compared to other groups suggests that the API may be disproportionately disadvantaging this population, leading to potentially unequal treatment. In any case, biased algorithms could also result in discriminatory practices such as targeted advertising or differential pricing.

It is also important to address the limitations of the approach of analysis conducted. As mentioned previously, one key limitation is the use of AI-generated faces from Generated Photos. While this data source allowed for easily obtaining a diverse set of faces, it did not perfectly capture the variations in appearance present in the real world. Additionally, Generated Photos did not provide options for all possible ethnicities or for non-binary gender identities. The image selection process, while randomized, could also introduce biases. The sample size of 120 faces, while sufficient for identifying broad trends, does not capture smaller differences in performance. A larger and more representative dataset would provide more definitive results. Finally, the quality scoring was based on subjective human judgments. While consistency and impartiality were the top priority, there is always the potential for individual biases to influence the scores. Additionally, it is important to consider the intersectionality of different demographic attributes. Intersectionality refers to the recognition that the experiences of individuals are shaped by the complex interplay of identities such as race, gender, and age. While this study focused primarily on ethnic disparities, it is important to remember that the API's performance may also vary at the intersection of these attributes. For example, the accuracy of the API may differ for Asian women compared to Asian men, or for younger Latino individuals compared to older Latino individuals. Analyzing the API's performance through an intersectional viewpoint would provide a more nuanced understanding of how different identities may face unique biases and challenges.

Based on the findings, several steps are recommended that Everypixel could take to improve the accuracy and fairness of their facial recognition API. Everypixel should carefully examine the data used to train their facial recognition algorithm. They should ensure that the training data is diverse and representative of different demographic groups, particularly those for which the API currently performs poorly. If the training data is found to be insufficient, Everypixel should invest in collecting more diverse data. This could involve partnering with organizations that have access to large datasets of faces or using artificial intelligence to generate more diverse training examples. Everypixel should also investigate the sources of bias in their current algorithm and develop techniques to avoid these biases. This could involve using alternate machine learning models. To provide more inclusive predictions, Everypixel should expand the demographic categories their API recognizes. This could include more specific ethnic groups as well as non-binary gender identities. To address any current ethical problems, Everypixel should be transparent with their customers and clients about the performance of their API across different demographic groups. They should report accuracy metrics and be clear about the limitations and potential biases of their service. They should also provide a clear process for users to report errors and provide feedback.

All things considered, facial recognition technology holds significant potential for society, but the risks associated with such technology raises significant concerns. The audit of the Everypixel API shows the importance of carefully evaluating these services and identifying places for service improvement. The performance differences seen across racial and ethnic groups show the need for more diverse and representative training data. As the use of facial recognition continues to grow, it is crucial that biases are mitigated and fairness is prioritized. The benefits of this technology should only be brought on in a way that respects the rights and privacy of all individuals.

Works Cited

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, pp. 77-91.

Everypixel API. n.d. https://www.everypixel.com/api.

Generated Photos. n.d. https://generated.photos/.

Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429-435.

Williams, Robert. "I Did Nothing Wrong. I Was Arrested Anyway." American Civil Liberties Union, https://www.aclu.org/news/privacy-technology/i-did-nothing-wrong-i-was-arrested-anyway.