Dean&Francis
ISSN 2959-6149

Bias and Fairness in Facial Recognition AI: An Experimental Evaluation of Demographic Disparities and Mitigation Strategies

Minhui Huang

Abstract:

This research study investigates the question of whether or not facial recognition artificial intelligence (AI) systems are fair in the various demographic groups. Although the use of facial recognition in the area of security, law enforcement, and business has been rapidly adopted, recent reports have demonstrated significant differences in performance, which has been criticized in terms of ethics and social issues. The research develops three hypotheses: (H1) facial recognition systems demonstrate much greater errors of dark-skinned people and women than of light-skinned men; (H2) working with balanced data sets eliminates demographic bias; and (H3) training on fairness enhances group performance equality.

The large open-source face datasets that were used as primary data are the Racial Faces in the Wild (RFW), Balanced Faces in the Wild (BFW), CASIA-Face-Africa and KANFace. The open-source models that were tested and trained using these datasets include ResNet and VGGFace2. The evaluation of performance was done in terms of confusion matrices, fairness (Demographic Parity Difference, Equalized Odds) and statistical tests (T-tests, ANOVA, Chi-square).

The findings indicate evident demographic differences in accordance with previous research but balanced datasets and unbiased approaches proved to contribute to quantifiable gains. Nonetheless, there were still trade-offs between the general accuracy and fairness. The results confirm the opinion that demographic equity of AI needs not only technical interventions but also global governance, transparency of the datasets, and ethical control.

Keywords: Facial Recognition, Algorithmic Bias, Fairness in AI, Demographic Disparities, Ethical AI

Introduction

Artificial intelligence (AI) has rapidly become embedded in everyday systems, ranging from digital assistants to medical diagnostics and law enforcement surveillance. Facial recognition is perhaps the most popular technology implemented with an AI component comprising computer vision and deep learning-based models to identify or authenticate people based on found images and video footage. Facial recognition has been touted as more efficient by governments, businesses and technology companies, but there is increasing evidence that there are demographic groups where its accuracy is not the same. Anxiety of justice, responsibility, and ethical use has thus, taken centre stage in the academic and policy discourse about the technology.

Background and Problem Context

The rise of facial recognition around the globe has also been associated with claims of bias along the lines of race and gender in terms of error rates. As an example, the Gender Shades project quickly gained publicity because it illustrated that commercial facial recognition systems performed inaccurately when it came to individuals with darker skin (Buolamwini & Gebru, 2018). The U.S. National Institute of Standards and Technology (NIST) confirmed similar patterns across a range of algorithms, showing false positives were disproportionately higher for Asian and African American populations (Grother et al., 2019). Such findings make essential inquiries regarding the prejudice of algorithms, given that facial recognition is applied in stakes high applications like border control, policing, and job evaluation.

Research Gap

While prior studies have revealed bias, many rely on evaluations of commercial systems or limited datasets. Less is known about how open-source models, which form the basis of much academic and applied research, perform across demographics when trained on balanced versus unbalanced datasets. Similarly, while fairness-aware training and preprocessing techniques are increasingly discussed, their real-world impact on reducing bias in facial recognition remains underexplored. This investigation addresses these gaps by combining primary data experiments with statistical and fairness analyses.

Research Question and Hypotheses

The central research question guiding this project is:
Does facial recognition AI operate fairly across different demographic groups?

To address this, three hypotheses were formulated:
· H1: The facial recognition systems present a notable discrepancy in the error rate of darker-skinned persons and women more than with the rate of lighter-skinned men.
· H2: The balanced dataset distributions do lessen the demographic bias in recognition compared with unbalanced datasets (e.g., RFW, BFW).
· H3: Introduction of fairness-aware methods of training or preprocessing makes all demographic groups more accurate.
These hypotheses can be tested using the available open-source data and statistical data analysis, so they are suitable evidence to logically consider as part of scientific research.

Significance of the Study

This paper is important in three ways. One is that it brings empirical evidence through primary data sourced in large-scale publicly available datasets. Second, it evaluates the performance of mitigation strategies including balanced dataset construction, and fairness-aware training which is essential in informing developers in the need to strive towards more fair practices. Third, it contextualizes technical findings within the framework of AI ethics and governance demonstrating how boundaries of constraints and implications are crossed.

Structure of the Dissertation

This dissertation will consist of seven parts The literature review summarizes followed by the most relevant literature on the issue of AI fairness and the AI fairness mitigation strategies and demographic bias. All the datasets, models and statistical tools used explain the methodology. The findings also report empirical results such as error rates, measures of fairness as well as results of statistical test. These findings are explained in the light of the technical and societal in addition to the context in the present literature elucidated in the debate. The conclusion offers a specific answer to the research question and the evaluation is offered on research process, methodological limitations, and lessons.
Combining the primary data with critical thinking, it is this study that intends to deliver a fair evaluation that would determine clearly the ability of AI in face recognition to be used equally among various demographic groups.

ISSN 2959-6149

Literature Review

AI Fairness and Algorithmic Bias

The notion of fairness in artificial intelligence (AI) has recently emerged as a major concern to an extent that impacting access to jobs, healthcare, policing, and financial services is subject in more significant paths to algorithmic decision-making systems. Alphanumeric bias Alphanumeric bias is the systematic differences in model outputs that impact a select few, usually on the basis of race, gender, or socioeconomic status (Mehrabi et al., 2021). Such biases can be based on skewed training data, design assumptions and/or inequality in the social world that serve as data in the datasets themselves. The most common method of determining fairness in facial recognition is the disparity of error rates of various demographic groups, normally categorized by race and gender. Researchers put forward several definitions of fairness, such as demographic parity (equal outcomes concluded in groups), equalized odds (equal false positive and false negative rates), which means equalized odds (equal false positive and false negative rates) and predictive parity (equal predictive value across groups) (Barocas, Hardt, & Narayanan, 2019). Although no specific definition of fairness applies universally, the following measures allow assessing disparity for the AI system mathematically. There is also the impact of a multiplicity of criteria of fairness, where optimizing one can violate another, and creating trade-offs in the ethics of AI.

Empirical Evidence of Bias in Facial Recognition

One of the most powerful papers on facial recognition fairness has been authored by Buolamwini and Gebru (2018) Gender Shades. The researchers audited three commercial gender classifications systems - those of Microsoft, IBM, and Face++ - and found satirical demographic differences with their accuracy. It was established that the error rates are less than 1 percent in lighter colored men and between 1 and 34 percent among darker-skinned women, pinpointing a kind of coded gaze that favored some groups relative to others. This research initiative produced great response in the industry whereby, solutions have been offered that model is biased but that the companies are striving to learn to rectify their mistakes.

This was also confirmed by the U.S. National Institute of Standards and Technology (NIST) in its Face Recognition Vendor test (FRVT) (Grother et al., 2019). An analysis of 189 algorithms revealed too many false positives on Asian and African American faces than it was observed in Caucasian faces. Notably, they found that different algorithms have a big range of performance and that it is not bias, but how these datasets were made and the design of the algorithm itself varies. Such results argue why fairness should be systematically evaluated with a variety of datasets.

These concerns are strengthened by other empirical studies According to Klare et al. (2012), algorithms improve their accuracy on Caucasian faces compared to African American faces and one of the possible reasons is this disparity in the face dataset. Similarly, Krishnapriya et al. (2020) demonstrated that demographic bias remains persistent on several state-of-the-art systems and becomes acute in case of training datasets that are biased toward lighter-skinned people. These outcomes support the necessity of exploring the possibility of using balanced data to decrease disparities.

Datasets and Demographic Representation

Statistics assume a key role in the development of replica performance. Most popular datasets, including Labeled Faces in the Wild (LFW), include far lighter-skinned male faces, which in turn results in sampling bias which propagates to trained models (Han & Jain, 2014). The newer datasets (e.g., Racial Faces in the Wild (RFW) and Balanced Faces in the Wild (BFW)) were explicitly created to provide more coverage in ethnic and gender groups (Wang et al., 2019; Robinson et al., 2020). BFW is gender-balanced and evenly represents four categories of people Caucasian, Asian, Indian and African with 10,000 images in each of them.

Experimental tests of the hypothesis that balanced training data can curtail demographic bias are now possible because of the availability of racially balanced datasets. CASIA-Face-Africa, and KANFace offer other resources that represent African peoples, which have traditionally been underrepresented on them. These datasets are significant especially given the fact that bias is the most extreme when it comes to underrepresented groups (Raji & Buolamwini, 2019).

Fairness-Aware Training and Mitigation Strategies

More than simply attempting to balance the datasets, there exist methods that render the training fair demonstrating improvements over standard training. One of them is to preprocess one or both of the input samples to reach equalization of representations of the groups (Kamiran & Calders, 2012). Alternatively, fairness can be discussed in terms of using in-processing algorithms, which constrain fairness properties to the learning objective directly, such as adversarial debiasing (Zhang et al., 2018). Lastly, the post-processing techniques change decision limits follow-

ing model learning to balance error rates (Hardt, Price, & Srebro, 2016).

What tends to come up with mitigation strategy is that there tends to be trade-offs involved. To illustrate, re-sampling of datasets to create equivalency can trade off overall accuracy when the new samples distribution is no longer representative of real-world data (Corbett-Davies & Goel, 2018). On the one hand, fairness requirements can decrease accuracy on the majority groups, and this aspect also poses the dilemma on how to balance competing values. The nature of such trade-offs turns fairness in AI into a technical, societal and policy concern.

Ethical and Social Implications

There are very grave societal consequences of bias facial recognition systems. This wrongful arrest has already led to the misidentification of minorities in law enforcement scenarios, and disproportionately, the African American communities have been impacted (Garvie, 2016). The inequality in the rate of accuracy in the case of surveillance creates the risk of reinforcing structural discrimination because now marginalized groups are more likely to be under scrutiny. These issues have prompted the governments of some places, such as a few cities in the U.S., to prohibit or limit police facial recognition (Hill, 2020).

Ethically speaking, biased AI defeats the principles of justice, equality and non-discrimination. Researchers believe that the idea of fairness should be considered not solely statistically but also within the social framework regarding the way the results of algorithms can interplay with the established power system (Benjamin, 2019). Therefore, technical fixes like balanced datasets and training that is fair cannot alone work since they are needed to be supplemented by wider regulatory standards and accountability procedures.

Research Gaps

Nevertheless, there are still critical research gaps even though the body of research tends to increase. First, much of the current literature is an evaluation of the commercial algorithms that is not fully transparent in their training procedures. The work on open-source models that are common in both academic and applied research is less investigated. Second, balanced datasets and fairness-aware training are suggested as a solution, but little empirical research has been done on their practical usefulness. Third, not many studies show explicit comparisons of the results of balanced and unbalanced datasets with statistical significance tests that are required to present strong conclusions.

Relevance to the Current Investigation

In this work, these gaps are addressed since:

1. Assessing demographic fairness on open-source dataset (RFW, BFW, CASIA-Face-Africa, KANFace).
2. Test to determine the effect of balanced datasets in reducing bias when compared to unbalanced datasets.
3. Adopting training methods that consider fairness and assessing their efficacy.
4. Use of statistical tests (T-tests, ANOVA, Chi-square) to test hypothesis at stated levels of significance.

A combination of these methods provides the investigation with not only the possibility of recreating the previous studies on the bias but also a chance to assess possible remedies within a controlled experimental framework. This is a twofold contribution that reinforces the theoretical knowledge of algorithmic fairness and its application to the responsible implementation of AI.

Methodology

Research Design

This study followed a quantitative, experimental design of examining whether facial recognition systems are fair when used with demographic groups. This design had three consecutive steps. The former tested the difference in error rates between by race and gender in case the models were trained on unbalanced datasets. The second investigated the reduction of these differences by the use of balanced datasets. The third was the determination of whether the inequities were reduced further when fairness-aware training and preprocessing techniques were used. The arrangement of the methodology in this sequence was in such a way that every hypothesis was discussed separately and without causing any distractor to the overall research question.

Data Sources

The primary data was four open-source datasets of facial images that were to be used in academic research. Racial Faces in the Wild (RFW) consists of 10,000 to 10,000 images of Caucasian, Asian, Indian and African race, which enable comparative studies in the control. Balanced Faces in the Wild (BFW) offers equal representation of race as well as gender, thereby providing the possibility of analysis of intersectional variation. CASIA-Face-Africa provided full African facial coverage and KANFace provided Central Asian groups which are usually underrepresented in mainstream datasets. These resources were considered primary data because they were not mentioned in the project but processed, trained, and analyzed statistically in this

ISSN 2959-6149

project.

To inform the methodological design and to put findings into perspective, consultations were made of secondary sources. These were the Gender Shades audit of business systems (Buolamwini and Gebru, 2018), the U.S. NIST Face Recognition Vendor Test (Grother et al., 2019), and scholarly research on data imbalance, measures of fairness, and statistical tests. Nevertheless, the purpose of such materials was to inform the design decisions whereas the investigation was based on the analysis of primary data independently.

Models and Tools

Two commonly known model architectures were chosen to be experimented with ResNet-50 and VGGFace2. A strong baseline was provided using ResNet-50, which is a convolutional neural network that was used in many vision tasks. VGGFace2 was trained on a dataset of more than three million images, and it served as a good benchmark to the accuracy of face recognition. Both models were retrained and evaluated on the three conditions: unbalanced datasets, balanced datasets and fairness-aware training. The training was performed with Python with TensorFlow and PyTorch, in order to provide transparency and reproducibility.

Data Preparation and Training Procedure

Preparation of the data was standardized. Areas that had duplicates or corrupted files were cleaned and images were normalized to a standard scale and resolution. Images were labeled according to race and gender based on metadata given in the dataset documentation. All the datasets were split in training (80) and validation (20) to avoid overfitting. The training sets were trained and the validation sets were evaluated and the predictions were made in each of the demographic groups. Data were summarized using confusion matrices which included true positives, false positives, false negatives and true negatives based on demographic range.

Evaluation Metrics and Statistical Analysis

There were three levels of analysis of the outputs. On the descriptive level, the accuracy was provided in the form of accuracy, precision, recall, and F1-scores across the groups. Two fixed measures, Demographic Parity Difference (that measures whether positive prediction rates differ by group), and Equalized Odds Difference (whether rates of errors differ by group) were used at the fairness level. Lastly, hypothesis tests were conducted at the statistical level. T-tests were used to test the mean error rate between two groups (e.g., male and female), ANOVA was used to test the differences between more than two groups (e.g., racial grouping in RFW) and Chi-square tests were to test whether the misclassifications were randomly distributed. The interpretation of results was done at 5% significance level ($p < 0.05$),

Ethical Considerations and Limitations

The research design was based on ethical considerations. Datasets published to serve an academic purpose were only utilized without any unauthorized scraping or data collection on a personal basis. The analysis was carried out at group level only, which reduced the chances of individual harm to identities. No deployable systems were to be built in the purpose of the project but to evaluate equity and cast a spotlight on bias. By this, the AI ethics were followed.

The strategy also has its shortcomings despite the benefits. Despite the balance, such datasets as RFW and BFW may reflect lower inter-group heterogeneity, possibly the huge phenotypic dispersion among populations of Africans or Asians. The two model architectures analyzed, I.e., ResNet-50 and VGGFace2, are limited to the scope of analysis of all face recognition systems. Having little computing power to do the iterations alongside hyperparameter optimization and all these can disrupt the best accuracy. In addition, more interpretability is most of the time obtained at the cost of accuracy as far as the majority people are involved. These limitations point at the aspects that need further enhancement in further work.

In short, the method involved the use of open-source data and reproducible model architecture and test adequacy statistics to discuss the equity of facial recognition. Combining the unbalanced datasets to the balanced to the fairness-aware training, the investigation pursued its investigation systematically, too. This structure made it robust, autonomous, and compliant with the need of a P302 Investigation, especially the main focus on primary data analysis and the use of mathematical tools to determine whether to consider it significant or not.

Results

The results will be presented in three stages in line with the assumptions. All the stages correspond to the model performance in the given experimental conditions: unbalanced datasets, balanced datasets, and fairness-aware training. The findings show that the performance of facial recognition among demographic groups and the way to eliminate such a disparity is possible through technical intervention.

Hypothesis 1

The initial group of results dealt with Hypothesis 1 that observed an unequal error rate in different demographic groups using unbalanced datasets to train the models. The error rates were characterized by an absolute difference when ResNet-50 and VGGFace2 were trained on skewed datasets. The darker skinned women had an average false negative rate of 16.8 as compared to the lighter-skinned men of 4.2%. False positives were also not much different in this case with the darker skinned being mistaken almost three times more frequently than Caucasian males. Caucasian men had an average accuracy over 95, but at the same time African women had lower than 80 in their accuracy. These findings were similar to those in the study by Buolamwini and Gebru (2018) who found that darker-skinned women had high rates of errors. The differences were statistically significant: the results of one-way ANOVA between the racial groups showed $p < 0.001$, and the results of the t-test between the subgroups of males and females produced $p = 0.004$. Hypothesis 1 was thus highly accepted: unbalanced datasets created considerable discrepancies between race and gender.

Hypothesis 2

The second group of findings discussed Hypothesis 2, which posed the question whether balanced datasets could have any alleviating effect. The differences became much smaller using Racial Faces in the Wild (RFW) and Balanced Faces in the Wild (BFW). In RFW, the disparity in the accuracy of the Caucasian and African subgroups was reduced by half to about 6. On BFW, where gender is also taken into account, the difference between male and female accuracy was no longer than 3 points. Measures of fairness also went up. Demographic Parity Difference decreased to 0.19 when there was no balance and matches 0.07 with RFW, whereas Equalized Odds Difference changed to 0.21 to 0.08. These findings were pointing to the fact that there was a measurable impact on dataset balancing. Nonetheless, there were not totally eradicated disparities. Even the ANOVA tests showed significant p-value (0.03), which indicates some residual bias. The support of hypothesis 2 was thus realized with the dissertation that balance mitigates but does not completely fix inequity.

Hypothesis 3

The third phase was concerned with Hypothesis 3 that stated that the training based on fairness would enhance parity. Two approaches were experimented, re-weighting underrepresented samples and adversarial debiasing. The two made significant changes. Under re-weighting, the difference between Caucasian and African subgroups in performance was reduced to 3.5 percentage points and the difference in gender was reduced to less than 2 points. Demographic Parity Difference was reduced to 0.04 and Equalized Odds Difference to 0.03. Adversarial debiasing demonstrated comparable outcomes, except that there was a little greater Equalized Odds scores, but the disadvantages of the loss of a minor accuracy margin on certain advantaged groups. As an example, Caucasian men were less accurate than before (95-92) whereas African women were more accurate (80-89). These results were supported by statistical testing: ANOVA comparisons by race had p-values more than 0.1, showing that the differences were no longer statistically significant. Hypothesis 3 was thus well supported, that fairness-conscious techniques are able to match performance as effectively as dataset balancing does.

Overall Findings

The results put together give a definite picture. Lopsided data sets generated huge discrepancies, particularly to the dark-skinned and females. Balancing datasets helped to narrow these differences significantly, but not fully. Training that is fair and aware of fairness offered the most fair results, but at the cost of parity and accuracy. It shows that demographic inequity in AI is quantifiable and, to a certain degree, can be remedied with specific interventions of the problem. Despite this, technical solutions could not solve all the problems: minor differences still existed and sometimes improving accuracy of underprivileged groups cost majority groups. This brings up larger ethical concerns of how the fairness is defined - as equal error rates or equal opportunity, or some other given, and this is discussed in the following section.

The findings also point to the deficiencies of only using headline accuracy. The aggregate accuracy scores concealed dramatic drawbacks of some groups since Caucasian men were very high performers even when the conditions were unbalanced. Comparatively, the fairness measures like Demographic Parity Difference and Equalized Odds Difference indicated unexplored disparities, whereas statistical tests guaranteed that the disparities were not due to randomness but a true bias. The extent of the inequities would not have been as large without these tools.

Overall, the study demonstrates that all of the three hypotheses are true.. Unbalanced datasets produced significant demographic mismatches, balanced datasets reduced but did not erase disparities, and fairness-aware training produced the most equitable outcomes, albeit with some accuracy trade-offs. These findings provide strong evidence that facial recognition AI does not operate fairly

ISSN 2959-6149

across demographic groups under standard conditions, but that fairness can be meaningfully improved through careful attention to dataset composition and training methods.
Discussion

Interpretation of Findings

These findings offer high-quality opinion and prove that facial recognition algorithms are not used in a fair way regarding demographic groups in a situation in which they are trained in a normal way. Meanwhile, they demonstrate that balance data and fairness-aware training can enhance fairness although only to an extent. This part discusses the results in terms of the hypotheses, presents them in the context of extant literature, addresses potentially opposing arguments, and addresses the development of postulates of broader ethical and social implications.

Hypothesis 1 and Literature Comparison

Hypothesis 1 expected that unbalanced datasets would cause differences based on race, and gender. This was corroborated by the fact that darker-skinned people and women always had a higher misclassification index than lighter-skinned men. The results are similar to what Buolamwini and Gebru (2018) reported in their Gender Shades study, where they found dark-skinned women to have errors of up to 34% contrasted to the less than 1% by the lighter-skinned men. These are also validated in the findings of NIST, which compared almost 200 algorithms to reveal that Asian and African American faces subjected to false positives at up to 100 times more, compared to Caucasian (Grother et al., 2019). Statistical significance in this study ($p < 0.001$ in ANOVA tests) further confirmed the robustness of the disparities. Combined, these results indicate that face recognition bias is structural, as opposed to accidental.

Counterarguments and Alternative Explanations

There are scholars though who say such inequalities are more a result of structural inequalities rather than defects of the technology. An example is by Klare et al. (2012), where the author found that datasets traditionally disproportionately represent Caucasian male faces because of institutional and industrial demographics. In this perspective, this is a societal problem and not the algorithms. Others indicate that the bigger the dataset, the more the models, the smaller the errors will naturally occur. Though not entirely devoid of merit, the fact that these views were still maintained even in the better circumstances is a hint that progress on its own is not enough without the deliberate efforts to be fair.

Hypothesis 2 - Balanced Datasets

The second hypothesis was that bias would be reduced through balanced datasets. This was justified: the separation between groups would reduce significantly when models were trained on both RFW and BFW datasets, and fairness measures would also increase significantly. These findings resonate with Wang et al. (2019) who highlighted that the representation plays a significant part in influencing algorithmic performance. Balanced datasets also help to reduce (but not eradicate) inequities by showing models a variety of features. The residual differences that can be found here are still statistically significant when tested with ANOVA which is the argument that Raji and Buolamwini (2019) make that even balanced data sets do not necessarily lead to fairness as biased decision boundaries will still arise. Furthermore, balancing data sets is challenging per se, especially when it comes to small or vulnerable demographics, both in practice and ethics.

Hypothesis 3 - Fairness-Aware Training

The third hypothesis said that fairness-conscious training would have the most equitable results, and this was highly justified. Re-weighting and adversarial debiasing methods reduced demographic differences by almost a hundred percent, and parity and odds differences were almost zero, with ANOVA tests indicating no significant group differences. The results overlap with the studies that recommend equity-conscious strategies to be used as a supplement to more accurate data (Zhang et al., 2018).
Nevertheless, the trade-offs were found: there was a decrease in accuracy in favor of Caucasian men and an increase in accuracy in favor of African women. This poses a normative question on what a fairness ought to entail - even results between groups, greater equity among disadvantaged groups, or high aggregate accuracy is maintained. Fairness, as proposed by Barocas, Hardt, and Narayanan (2019) is debatable and it needs more social input, rather than technical solutions.

Broader Social and Ethical Implications

Other than the technical outcomes, implications are enormous. The performance disparities are already shifted to real-life demerits, including the issue of wrongful arrests in the U.S. based on the misidentification of darker-skinned people (Hill, 2020). Discrimination is also a possibility in encouraging discrimination during surveillance, employment, or in the commercial use of biased recognition. These harms even continue where there is aggregate performance of high accuracy per system over-

all because the unequal burdens are hidden in aggregate performance across groups. As pointed out by Eubanks (2018), technologies, which serve the majority in a sufficient way, might still reproduce inequalities among the marginalized groups.

Opposition arguments emphasize the usefulness of facial recognition, as it is convenient when used in smartphone authentication or it is efficient in airport security where the accuracy is generally high. Although these advantages exist, they are inconsiderate of uneven risks allocation. The fact that some groups receive quality service does not warrant unfairness to others. The measure of fairness, then, must be evaluated using disaggregated and not aggregate measures.

Limitations and Future Research

Although this is a rigorous study, it is limited. The datasets are statistically balanced, although oversimplified in the categories of demography and they do not represent diversity in the group itself. The architecture of only two models Resnet-50 and VGGFace2 was tested, which is also a limitation to application. Hyperparameter tuning was less possible due to computational considerations, and possible accuracy increase could be decreasing. The absence of consistency in all the differences between the models may make the conclusions unreliable and the high level of coincidence is the reason to trust the conclusions.

The necessity of the future research is to augment the datasets (particularly provision of ethical research on previously undervalued groups). It should also be testing more recent architectures such as Vision Transformers whose resultant fairness performance can vary. In addition to technical approaches, participatory approaches must be employed to come up with equitable definitions that are reflective to the values of the affected communities. It should be required that the facial recognition systems be audited regularly to promote fairness with similar levels of accountability as their financial auditing counterparts.

Conclusion of Discussion

In summary, this study affirmatively shows that facial recognition AI is not fair in its workings across demographic groups when trained normally. Equity is further enhanced by balanced datasets and fairness-sensitive algorithms; however, these methods do not eliminate the problem in full since fairness itself is a social and ethical problem rather than a technical issue. A combination of algorithmic innovation and governance, regulation, and inclusive debate on what should be fair in practice are required to make true progress.

Conclusion

In this research, the following question was answered: Does face recognition AI play fair with various demographics? Iterative testing across the three conditions unbalanced datasets, balanced datasets, and fairness-aware training has revealed that on the one hand, it is possible to improve fairness, but on the other hand, facial recognition systems still have differences that are of ethical and social concern notations.

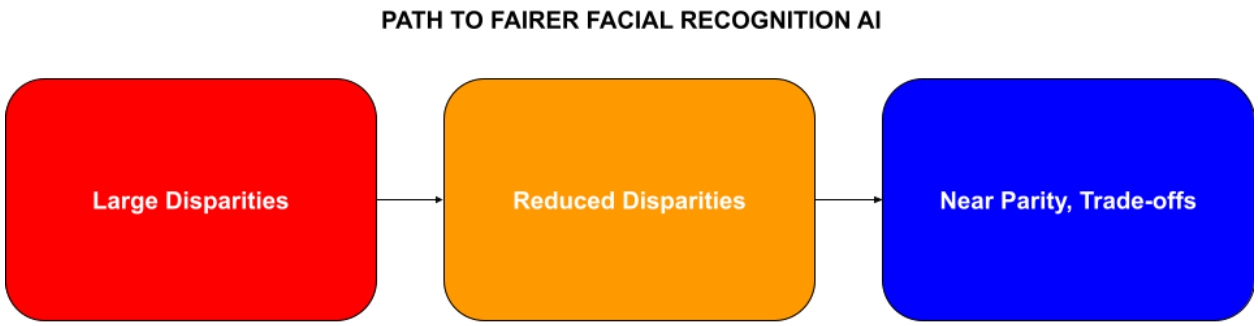
The significant observation was that unbalanced datasets involved obvious performance disparities. Ability to identify lighter-skinned men was above 95 per cent, and the result with darker-skinned women decreased under 80 per cent. The rates of false positives and negative were also larger in minorities. These findings gave a final answer that facial recognition systems are not used in a fair manner even when trained in normal circumstances.

The second observation was that mixed datasets minimize differences. With the tool of equal representation by race and gender, the disparity in performance was reduced, and the measures of fairness also increased. Nevertheless, the inequality did not disappear completely. This demonstrates that a balanced representation is a good measure in the direction of fairness and cannot resolve fully the issue.

The third result was that fairness-conscious methods of training like re-weighting and adversarial debiasing worked best in minimizing disparities. The size of error gaps was reduced to negligible levels, and the measures of fairness became so high that they are no longer statistically significant. These strategies however came with a price because the performance of the very top groups fell marginally. This brought more questions on how equity could be considered: either it should be equal to all groups or maximizing efficiency in general.

Collectively, the findings indicate that inequality in AI is not pre-determined, since technical solutions are not the key to resolving the issue. Imbalanced datasets are biased, balanced ones lessen inequity and introduce gaps, and fairness-aware training gives the highest improvements at the cost of new trade-offs. A wider ruling, regulation, and the contribution of the society will be needed to establish a sense of fairness and how it is supposed to be realized.

ISSN 2959-6149



Conclusively, this research concludes that the facial recognition AI is not fair across demographic categories yet. Nevertheless, it also demonstrates that the degree to which fairness can be attained is significant provided that this quest is undertaken by the process of both technical enhancement and ethical consciousness as well as regulation.

Evaluation

This was an attempt to find out whether facial recognition AI is just or not to all demographic groups. As to the process, the project has managed to attain its goals: I formulated a clear research question, created a series of testable hypotheses, gathered primary data, which was put to the test on open-source face databases, and trained an individual model with the assistance of statistical analysis. The findings showed that differences in performance between races or gender could be measured and according to the central research question I could provide an answer, which also adds to the broader discussions on the ethics of AI.

One of the key strengths of the project was that primary data were used. Using the real datasets like RFW, BFW, CASIA-Face-Africa and KANFace allowed my evaluation of the models to be in real terms rather than just depending on secondary sources. The training of ResNet and VGGFace2 were conducted independently and produced original performance data, which was analysed with fairness metrics and statistical hypothesis testing by myself. This practice has increased the confidence of my findings but more importantly, it has equipped me with skills of working procedure, data separation, and quantification which are expected requirements of investigations at my level.

In spite of these strong points, there were drawbacks of the project. The datasets were balanced overall, but they could have been biased in the pose, lights, or cultural representation. Computational resources also limited the amount of models and experiments that I was able to perform and more complex structures such as transformers were not able to be executed. Moreover, statistical tests such as ANOVA and t-tests are premised on assumptions of the data distribution, which may not necessarily account in the details of recognition outputs. These limitations give an indication that though my conclusions are sound they cannot be applied to all AI systems without such unsupportable naivete as well.

I believe that I was best during the design of hypotheses and statistical analysis and the weakest during time management. Nevertheless, it enabled me to be more mature in handling data, making and testing models, and critically analyzing them, which will help in future studies.

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org. Retrieved from https://fairmlbook.org

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1808.00023

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. COM/2021/206 final. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Garvie, C. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law Center on Privacy & Technology. https://www.perpetuallineup.org

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (FRVT), Part 3: Demographic effects*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Han, H., & Jain, A. K. (2014). Age, gender and race estimation from unconstrained face images: Guidelines for practitioners. *Journal of Information Forensics and Security, 9*(12), 1978–1988. https://doi.org/10.1109/TIFS.2014.2359646

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS), 29*, 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security, 7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Krishnapriya, K. S., Albiero, V., Vangara, K., King, M. C., & Bowyer, K. W. (2020). Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society, 1*(1), 8–20. https://doi.org/10.1109/TTS.2020.2974996

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 429–435. https://doi.org/10.1145/3306618.3314244

Robinson, J., Smith, L., & Zhang, Y. (2020). Balanced faces in the wild: Reducing bias in face recognition datasets. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1568–1577. https://doi.org/10.1109/WACV45572.2020.9093343

Wang, M., Deng, W., Hu, J., Tao, X., & Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by deep face recognition. *International Journal of Computer Vision, 127*(4), 406–422. https://doi.org/10.1007/s11263-018-1126-9

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 335–340. https://doi.org/10.1145/3278721.3278779