Reducing bias through the careful curation of training data for developing face recognition algorithms

by Leslie G. Little

B.S. in Industrial Technology, December 2016, East Carolina University
M.S. in Technology Management, December 2018, North Carolina A & T State University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 10, 2025

Praxis directed by

Joseph P. Blackford
Associate Professor of Engineering and Applied Science

Enrique Campos-Náñez
Professorial Lecturer in Engineering and Applied Science

Michael C. Jones
Professorial Lecturer in Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Leslie G. Little has passed the Final Examination for the degree of Doctor of Engineering as of November 20, 2024. This is the final and approved form of the Praxis.

Reducing bias through the careful curation of training data for developing face recognition algorithms

Leslie G. Little

Praxis Research Committee:

Joseph P. Blackford, Associate Professor of Engineering and Applied Science, Chair

Enrique Campos-Náñez, Professorial Lecturer in Engineering and Applied Science, Advisor

Michael C. Jones, Professorial Lecturer in Engineering and Applied Science, Committee Member

© Copyright 2024 by Leslie G. Little
All rights reserved.

Dedication

To my husband, my backbone, my rock, and most often, my biggest headache, Billy Carson Little, Jr. (yeah..he forced me to write his entire name), thank you for the support, the laughter when I wanted to break down into tears, and for keeping the ship afloat while I spent countless hours behind the computer plugging away at this research. I don't say it often but, THANK YOU for all you do to keep the house running.

To my only son, Kaden Isaiah Little (hey, I'm on a roll with the entire government names, why stop now) you are truly the light that beams brightest. Don't ever let anyone or anything dim your light, son.

To my daughter, Imari Latrice… I'll be up your way soon to eat your homework! Make it fancy, Chef. You are a wonderful woman, and you continue to make me proud. Like Lauren said, "You are going to do big things".

To my celestial being, my oldest daughter Lauren Charis… there's not a moment that goes by, that I am not thinking of you. I miss you so much and am thankful for the 17.5 years I had with you. You always knew Mom could do it even when I didn't think I could.

And finally, this praxis is dedicated to my younger self, who at the most inopportune moment in life's history, decided to venture beyond conventional boundaries and confront this challenging experience.

Acknowledgments

Words cannot express my gratitude to my professors for their invaluable patience and feedback. I could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise.

I am also grateful to my cohort members for their editing help, late-night feedback sessions, and moral support. Thanks, should also go to my office staff and co-workers who impacted and inspired me with words of encouragement and after-hours emergency help amid my (*many*) emotional breakdowns.

Lastly, I would be remiss in not mentioning my mentors. Their belief in me has kept my spirits and motivation high during this process.

Abstract of Praxis

Reducing bias through the careful curation of training data for developing face recognition algorithms

While face recognition algorithms have their uses, they also carry the potential to be biased and do harm to some groups. In order to create better models, it is essential to comprehend how these algorithms are trained and what factors affect their accuracy and fairness. How the ethnic composition of training data affects the performance of facial recognition systems is the topic of this research. We show that gender mislabeling and racial bias are two of several biases that can have a major effect on accuracy when combined.

As part of our research, we compiled information from a wide variety of sources. We were able to significantly lower the likelihood of bias due to certain groups' underrepresentation, particularly African Americans, by including a wide range of perspectives and experiences in the data.

We find that changing a color image to a grayscale one yields different meanings, which allows us to assess how well our curated datasets analyze the performance of our FRT algorithms.

In order to implement our findings, we developed a Predictive Code Completion Model using Python code that made use of the Tkinter toolkit.

Table of Contents

Dedication ..... **iv**

Acknowledgments ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xii**

List of Equations ..... **xiii**

List of Symbols ..... **xiv**

List of Acronyms ..... **xv**

Glossary of Terms ..... **xvii**

Chapter 1—Introduction **1**

1.1 Background 1

1.2 Research Motivation 2

1.3 Problem Statement 3

1.4 Thesis Statement 4

1.5 Research Objectives 5

1.6 Research Questions and Hypotheses 6

1.7 Scope of Research 6

1.8 Research Limitations 7

1.9 Organization of Praxis 8

Chapter 2—Literature Review **11**

2.1 Introduction 11

2.2 AI in Law Enforcement 13
2.3 The Impact of Biased Training Data on African Americans 16
2.4 The Need for Curated Training Data 16
2.4.1 Challenges in Facial Recognition Data Curation 17
2.5 Trust in AI 18
2.6 Explainable AI on Algorithmic Bias Mitigation 20
2.6.1 Data Curation as a Mitigation Strategy 22
2.6.2 Algorithmic Auditing as a Mitigation Strategy 23
Through the implementation of a comprehensive algorithmic auditing procedure, researchers and engineers can attain a more profound comprehension of the origins of innate biases in FR systems and undertake specific measures to mitigate them. 24
2.7 Face Recognition Biases Beyond Demographics 24
2.7.1 “Face Recognition Biases: Too Bias, or Not Too Bias?” 26
2.8 Understanding Fairness of Gender Classification Algorithms 28
2.9 Summary and Conclusion 29
Chapter 3—Methodology 32
3.1 Introduction 32
3.2 Face Recognition & Bias (Data Collection) 33
3.2.1 Datasets suitable for doing experiments in deep face recognition. 34
3.2.2 Evaluation Datasets (Model Training) 35

3.2.2.1 MORPH-II Dataset 35

3.2.2.2 Labeled Faces in the Wild (LFW) dataset. 35

3.2.2.3 NIST FRVT Dataset. 36

3.2.3 Data Bias Mitigation 37

3.2.3.1 Model Data Quality 37

3.2.3.2 Data Augmentation 38

3.2.4 Organization of CNNs 39

3.2.4.1 Techniques for Training CNNs 40

3.3 Metrics 42

3.3.1 Datasets 45

3.3.2 Model Training 46

3.4 Performance Evaluation (Modeling Best Practices) 48

3.4.1 Analysis of FMR and FNMR 50

3.4.2 Processes for Ensuring Fairness in Model Training 52

Chapter 4—Results 53

4.1 Introduction 53

4.1.1 Obtaining NIST Dataset 54

4.1.2 Obtaining LFW Dataset 55

4.2 Executing Dataset Learning Environment 57

4.3 Details of Construction and Composition 59

4.4 Dataset Evaluation 61

4.4.1 Dataset Balancing 64

4.5 Bias Mitigation by Retraining FR Algorithms vs. Results from Related Studies 65

4.5.1 Summary of Key Studies from Literature Review 66

4.6 Research Findings and Hypotheses Validation 67

4.7 Ethics in FR 72

Chapter 5—Discussion and Conclusions 76

5.1 Discussion 76

5.2 Conclusions 77

5.3 Contributions to Body of Knowledge 79

5.4 Recommendations for Future Research 80

References 81

Appendix A 98

Appendix B 100

List of Figures

Figure 1: Diagram of Facial Recognition System..... 15
Figure 2: Blue face detection image 25
Figure 3: Racial Discrimination in FRT 27
Figure 4: Depiction of biometrics. 28
Figure 5: MORPH images 35
Figure 6: LFW sample images 36
Figure 7: How a CNN is put together 40
Figure 8: Confusion Matrix 45
Figure 9: Analysis of the progression of losses and validation accuracy 46
Figure 10: Flowchart of working Viola-Jones algorithm 49
Figure 11: ROC Curve 50
Figure 12: Types of errors..... 52
Figure 13: Biases in Facial Recognition Technology 69

List of Tables

Table 1: Number of images by Gender and Race 35
Table 2: Label distribution for LFW attributes 63
Table 3: Presents a summary of the number of samples from each cluster categorized by ethnicity and gender 64
Table 4: Comparative Results – Bias Mitigation vs. Results from related studies. 66

List of Equations

Equation 1: True Positive Rate 27
Equation 2: False Negative Rate 27
Equation 3: True Negative Rate 27
Equation 4: False Positive Rate 27
Equation 5: Binary Classification 42
Equation 6: Optimization Equation 42

List of Symbols

x	State of the system
y	Output of the system
z	Noise
n	Neural Network Layer
$\in$	Element of a set
$\mathbb{R}^m$	The set of real numbers; Vector spaces; Real value/M-dimensional
$\mathbb{R}^d$	The set of real numbers; Vector spaces; Real numbers/D-dimensional

List of Acronyms

AA-M	African American Male
AA-F	African American Female
AI	Artificial Intelligence
API	Application Programming Interface
BFW	Balanced Faces in the Wild
CNN	Convolutional Neural Network
DNA	Deoxyribonucleic Acid
FBI	Federal Bureau of Investigation
FC	fully connected (neural network layer)
FERET	**F**ac**E** **RE**ognition **T**echnology
FMR	False Match Rate
FNMR	False Non-Match Rate
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FR	Face/Facial Recognition
FRT	Face/Facial Recognition Technology
FRVT	Face/Facial Recognition Vendor Test (-ing)
GAN	Generative Adversarial Network (s)
LFW	Labeled (Labelled) Faces in the Wild

ML	Machine Learning
NIST	National Institute of Standards and Technology
NYPD	New York Police Department
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
ROC	Receiver Operating Characteristics
SDM	Signal Detection Model
SL	Supervised Learning
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
US	United States

Glossary of Terms

AdaBoost	A machine learning method that ensembles many weak classifiers together to form a strong classifier and trains the models sequentially, with each model addressing the weaknesses of its previous model.
API	Application Programming Interface – allows communication between software applications
ArcFace	A machine learning model that takes two photos of faces as input and calculates the distance between them to determine whether they are the same person.
BFW	Balanced Faces in the Wild is a dataset that serves as a proxy for assessing bias between racial and gender categories enabling characterization of the FR performance of each subgroup.
bias	A pattern mistake that happens when a model wrongly guesses about the data, which can lead to biased predictions and maybe even unfair results
classifier	An algorithm that divides data into groups or classifications. A classifier's objective is to learn from training data and then apply that knowledge to forecast fresh data.
confidence level	A metric indicating the confidence of a model in the accuracy of its prediction
confusion matrix	A table that compares the expected and actual values to demonstrate how well a machine learning (ML) system classifies data.
CONV layer	The convolutional layer serves as the fundamental component of a CNN.

data augmentation	A method for training machine learning models that generate new data points from existing data.
epoch (s)	Perform a single pass over the entire training set. The technique handles every data point, modifying the weights of the model to reduce mistakes and improve its forecasts.
FaceNet	A facial recognition system built by Google researchers Florian Schroff, Dmitry Kalenichenko, and James Philbina.
FERET	A project (database/evaluation/program) at the National Institute of Standards and Technology (NIST) that creates facial recognition algorithms, compiles a database of facial photos and evaluates the algorithms.
fine-tuning	Retraining a pre-trained model on a particular dataset or task is a machine-learning strategy that enhances performance.
fully connected	A network of interconnected neural nodes in which all inputs are linked to all outputs**.** Not every node in a convolutional layer is connected.
GhostFaceNets	GhostFaceNets is a collection of lightweight face recognition models built on top of GhostNetV1 and GhostNetV2, both of which are based on the Ghost module.
grayscale (-ing)	A method (technique) for training models with photos consisting solely of grayscale – removing color.
Jupyter Notebook	A free, open-source web tool that allows users to produce and distribute interactive documents that combine text, code, and other media. In addition to supporting more than forty programming languages, Jupyter Notebook also supports Python.
Kaggle	A website where experts in the fields of data science and machine learning can meet, network, and share knowledge

MAAD-Face	A free, publicly available **M**assively **A**nnotated **A**ttribute **D**ataset that consists of over 3M facial images.
marginalized group	A marginalized group consists of people who face discrimination and exclusion from mainstream society as a result of imbalanced power dynamics. Marginalization can stem from a variety of factors, including race, gender identity, sexual orientation, and age.
mitigation	The procedure of mitigating or eradicating bias in AI systems to enhance the fairness and accuracy of their prediction.
MORPH (II)	Collection (dataset) of mugshots of individuals ranging in age from 16-77
n-layer	Neural layer
POOL	CNN layer that preserves the most crucial information while reducing the input data's dimensions
Python	A machine programming language that can be used to teach computers to learn from data and make predictions.
ROC (curve)	A figure illustrating the performance of an ML model over all categorization thresholds.
Sensitive loss	Includes fairness as a learning goal during the algorithm's training phase. The technique enhances the performance and fairness of pre-trained representations without necessitating full retraining by acting as an add-on.
standard CNN	A deep learning model that processes and analyzes data in a grid pattern using a stack of layers.
TKinter	A Python library
Viola-Jones	In 2001, Paul Viola and Michael Jones created a machine-learning framework for object detection. Although it can be modified to detect other things, it was originally developed to detect faces.

Chapter 1—Introduction

1.1 Background

This study provides empirical evidence that bias, particularly against African Americans aged 18 to 65, can be mitigated by retraining facial recognition algorithms. This could help U.S. law enforcement accurately identify potential criminals.

Facial recognition technology (FRT) is utilized in several applications, such as enabling secure smartphone access and identifying criminal suspects from surveillance photographs, as a tool in the justice system (Wehrli et al., 2022). In collaboration with the academic community, various organizations advocating for citizens' rights and social justice have pinpointed negative societal outcomes resulting from the unquestioning application of FRT algorithms. These implications include wrongful arrests and an excessive level of government surveillance. In the United States, the negative outcomes of these consequences have had a greater impact on individuals belonging to racial and ethnic minority groups (Limante, n.d.). This issue is because algorithms are generally less precise when used on nonwhite people. Even more than this, FRT systems are being integrated into existing systems and institutions that have a history of unequal treatment and disparities (Terhorst et al., 2022). As research into racial biases in the operation of facial recognition technology continues to advance, enhancing facial recognition algorithms could become a comprehensive solution to mitigate these biases. Addressing population differences in training data is critical in preventing this phenomenon. This means making the labels on the data better and having outside auditors look over the datasets to make sure they are diverse and fair.

1.2 Research Motivation

Facial recognition technology is ubiquitous. Because this technology is almost unavoidable today, it is reasonable to expect it to be closely monitored and regulated. It isn't. The technology is fundamentally flawed and largely unregulated.

The need to address algorithmic bias, which occurs when an algorithmic decision unfairly benefits some groups at the expense of others, is what motivates this research. This matters because algorithms are the gatekeepers of civil liberties.

Imagine being handcuffed in front of your neighbors and family for stealing watches (Jefferson, 2020). After spending hours behind bars, you learn that the facial recognition software state police used on footage from the store identified you as the thief. But you didn’t steal anything; the software pointed cops to the wrong guy (Chicklas, 2021). Unfortunately, this is not hypothetical. This happened three years ago to Robert Williams, a Black father in suburban Detroit. Sadly, Williams’ story is not a one-off.

In a recent case of mistaken identity, facial recognition technology led to the wrongful arrest of a Black Georgian for purse thefts in Louisiana (Brodkin, 2023). Many times, these misidentifications cause irreparable damage.

Case in point, a Texas man is suing Macy's and Sunglass Hut's parent business for allegedly using error-prone facial recognition technology to wrongfully accuse him of armed robbery. Harvey Murphy Jr. was released after nearly two weeks in jail when prosecutors determined he was not physically present in the state when the crime occurred. During his imprisonment, Murphy was allegedly attacked and raped by three guys, resulting in lifelong damage (Fung, 2024).

Technology does not exist apart from the biases and bigotry that pervade our culture. According to studies, people of color, women, and nonbinary people have the lowest levels of facial recognition reliability (Garg et al., 2018). Conversely, criminal databases used by U.S. law enforcement disproportionately include individuals from minority groups, thereby exacerbating racial inequality. With law enforcement having access to this technology, the consequences can be fatal.

Racial disparities in the performance of FRT can be attributed to the underrepresentation of non-white faces in the sample datasets used to train the algorithms (Jefferson, 2020).

Most Americans support the use of facial recognition technology by law enforcement, but research shows that some law enforcement agencies rely too heavily on the system while human staff are still needed (Chong et al., 2022). Increased transparency and effective leadership may be necessary to resolve the issue.

1.3 Problem Statement

Using improperly trained AI facial recognition technology by law enforcement results in an error rate ranging from 0.1 to 9.3% accuracy (Grother et al., 2019) causing a disproportionate number of misidentification of African Americans.

The problem to be addressed through this study is to analyze the accuracy of facial recognition algorithms and suggest a strategy to mitigate their inaccuracies. Facial recognition is complex and relies on several factors; for instance, how up-to-date and accurate is the algorithm used by the justice system (Wu et al., 2023)? Which image feeds into the algorithm? To get witnesses to confirm or identify a list of suspects, what is the best way to display the list? People are quite concerned about how accurate the

program is (Albiero et al., 2020). However, the reliability of an algorithm is affected by several things. Among these elements are the decisions made by the developer, the quality of the dataset, the training process, and the structure of the neural network (Kabir et al., 2018).

There is a correlation between the input photo's quality and the output (Morales et al., 2021). The gender, age, and race of the searcher may also have an impact on the algorithm's performance. Up until now, one of the biggest worries has been the possibility that algorithms would unfairly hurt Black people due to the technology's disproportionately high mistake rates for people of color (B et al., 2024). When law enforcement uses facial recognition technology, there is significant concern about the potential for false arrests (Johnson et al., 2022).

1.4 Thesis Statement

To enhance outcomes and reduce biases in facial recognition, an effective training algorithm uses data augmentation with diverse and balanced datasets to reduce disparities in ethnicity, gender, or skin tone.

Federal, state, and local law enforcement agency's use of a fine-tuned algorithm with less accurate results increases the chances of bias during arrests because of facial recognition. One approach is to optimize the algorithm by removing color from the image, thus minimizing the likelihood of bias. Grayscale images eliminate color information, thereby reducing the impact of skin color changes on recognition accuracy (Bhatta et al., 2024). Training in grayscale may reduce bias; however, it is essential to maintain a diverse dataset that accurately represents various ethnicities and skin tones to optimize accuracy for all individuals. Facial recognition systems that use color images

may exhibit greater sensitivity to biases associated with skin color variation, particularly when accurately identifying individuals with darker skin tones.

Employing methodologies like the “blind taste test” can effectively reduce bias (Galhotra et al., 2017). This involves evaluating algorithms without regard to the demographic details of people in the images (Morales et al., 2021), focusing solely on performance and accuracy while reducing implicit biases.

Comprehensive data review and enhancement involves careful and systematic analysis of datasets and results, facilitating algorithmic improvements and mitigating bias in future software iterations (Terhorst et al., 2022). This requires enhanced data labeling and evaluation of external datasets to ensure diversity and representativeness.

1.5 Research Objectives

The primary goal of this research on FRT is to accurately identify and differentiate African Americans between the ages of 18 and 65, both individually and as a group. The number of incorrect identifications can fluctuate, contingent on the technological method employed for facial recognition (51Testing, 2021). The most accurate face identification algorithm has an error rate of 0.08% (admin, 2017). Facial recognition systems that incorporate liveness detection exhibit superior levels of accuracy.

This work will focus on implementing facial recognition with deep learning CNN architecture in Python to mitigate algorithmic biases.

This research will explore favorable strategies to reduce algorithmic bias by addressing the following research questions and hypotheses:

1.6 Research Questions and Hypotheses

The study questions and hypotheses to reach the thesis statement are listed below, considering the previously stated research objectives:

RQ1: Are facial recognition algorithms racially biased?

RQ2: Can a facial recognition algorithm model be built to improve accuracy and reduce built-in biases?

RQ3: Can the unjustness brought about by AI misidentification be mitigated by incorporating human oversight?

H1: Assuming the datasets used to train AI technology are biased, mostly using white male faces to train FR systems while ignoring faces of different races, the technology will unavoidably produce biased conclusions.

H2: With 90% accuracy, the facial recognition algorithm model can be retrained to correctly classify faces as Good or Bad Quality.

H3: Human involvement in data labeling, thorough data testing, and ensuring a diverse sample representative of the African American community effectively illustrate that basic best practice protocols in FRT mitigate biases.

1.7 Scope of Research

This study aims to investigate a significant issue where algorithms disproportionately affect African Americans, especially men and women between the ages of 18 and 65, by emphasizing a main concern: the reduced accuracy of FRT systems for African Americans, a phenomenon known as bias.

Analyzing a clear method to eliminate this prejudice—that is, retraining models for sensitive features like gender or color—helps one to appreciate why this approach is

virtually equal to reducing bias. If facial recognition models focus less on sensitive features, one would expect the differences in accuracy between groups to be minimized. This method seems to successfully help to eliminate bias in scientific writing. This process involves assessing model awareness of sensitive aspects and their impact on accuracy. In particular, blind pre-trained models are examined to reduce their sensitivity to significantly important features. Not surprisingly, no correlation was shown between awareness and accuracy, or between awareness and bias. In these experiments, blinding has somewhat compromised accuracy (Albiero et al., 2020). The seemingly simple concept of reducing bias in facial recognition rates by reducing awareness of sensitive features is ineffective; ignoring sensitive features is unrealistic for achieving less biased facial recognition.

The NIST-produced FRVT dataset provides a quantitative benefit to the research findings. Launched in 2000, the Face Recognition Vendor Testing Program (FRVT) at NIST was intended to test both prototype and commercially available facial recognition algorithms neutrally. The U.S. government is using the test results to help determine where and how to most judiciously apply facial recognition technology. To ensure more accurate reporting and complete this research, the NIST dataset was downloaded and extracted in March 2024.

1.8 Research Limitations

We recognize that our research study has certain constraints that may affect the practical implementation of our method in real-world scenarios. The following are limitations and challenges encountered within this research:

- Possessing a sufficient depiction of human faces, which differ widely in terms of size, shape, expression, and illumination. These characteristics may make it more difficult for the algorithm to generalize and adapt to different situations. A large number of the faces in the NIST FRVT dataset failed to provide a sufficiently representative sample of African American men and women in the 18-65 age range.
- The diversity and caliber of the datasets employed to train and assess the algorithms. Diverse datasets may display differing characteristics, such as resolution, format, annotation, distribution, and bias. This presents a significant issue in recognizing various skin tones, backgrounds, facial characteristics, and additional facial elements. The NIST dataset was partitioned into training and testing subsets and evaluated against the Labeled Faces in the Wild (LFW) dataset.
- The complex nature and constraints of the algorithms employed to execute the tasks. The algorithms must be precise and resilient to accommodate the unpredictability and diversity of faces and datasets.

  Recognizing these limitations allows for the refinement and expansion of understanding of CNNs used for facial recognition.

1.9 Organization of Praxis

This study examines the feasibility of mitigating bias through careful curation of the training data used in the development of algorithmic CNN models. This praxis paper presents our research findings organized into five chapters.

Chapter 1 – Introduction: The first chapter presents the introduction to the research topic and outlines our research objectives by offering a comprehensive overview of facial recognition technology. This chapter discusses the biases present in the flawed algorithms used in law enforcement's FRT and the effects these inaccuracies have on African Americans between the ages of 18 and 65. Our suggested solutions aim to mitigate these biases.

Chapter 2 – Literature Review: The second chapter presents the literature review of existing research on FRT. It provides a comparative analysis of previous studies related to this subject, highlighting both similarities and differences in findings. The literature review analyzes essential publications and sources to provide the basis for our research inquiry.

Chapter 3 – Methodology: The third chapter is titled Methodology, and it is where the CNN architecture and the methodology that was employed for the study are presented. Some important components of the research are discussed, including the origin of the data, the machine learning algorithms that were implemented in Python, and the assessment measures that were utilized throughout the research.

Chapter 4 – Results: The results of the retraining of the FRT algorithm are shown in the fourth chapter, which also includes a comparison of the method's performance to that of previous works that have utilized the NIST FRVT dataset that is used by law enforcement.

Chapter 5 – The fifth and final chapter of the research paper encompasses an overview of the research contributions to the existing body of knowledge, an examination of the implications and limitations of the findings, and recommendations for future

research in facial recognition technology and methods for mitigating bias in algorithms utilized by law enforcement. This chapter acts as the final summary of the research praxis.

This document is structured in the following manner: Chapter 2 examines the existing literature on FRT research, focusing on the biases present in facial recognition algorithms. It provides a comparative analysis of previous studies related to this subject, highlighting both similarities and differences in findings. Chapter 3 outlines the methodology employed for the FRT algorithms utilized in this research. This summarizes the strengths and weaknesses found. Chapter 4 presents the quantitative findings of this study focused on mitigating FRT biases through the careful curation of training data for the models. Chapter 5 presents the findings of the research, highlighting any constraints in the analysis and specifically examining the validation or refutation of the hypotheses.

Chapter 2—Literature Review

2.1 Introduction

Facial recognition technology has become increasingly prevalent in law enforcement and other societal domains, raising concerns about its potential impact on marginalized communities, particularly African Americans within the 18-65 age range. However, the accuracy and fairness of these systems have been a growing concern, as they have been shown to exhibit biases against this demographic group (Stark, 2019) (Guo et al., 2019) (Zeng et al., 2019). These biases can lead to disproportionate errors and discriminatory outcomes, undermining the technology's trust and effectiveness. Recognizing the critical importance of addressing these biases, researchers have explored a multitude of strategies with a particular emphasis on the pivotal role of carefully curating the training data used to develop these algorithms (Ferrara, 2023).

Facial recognition technology has been rapidly adopted by law enforcement agencies across the United States, with a growing number of police departments utilizing this technology for a variety of purposes, including surveillance, identification, and crime prevention (Radiya-Dixit & Neff, 2023). While the widespread use of this technology has the potential to aid law enforcement efforts, it has also raised significant concerns about its inherent biases and the infringement of civil liberties (Cagle, 2024), particularly for marginalized communities such as African Americans.

As stated in Chapter 1, identifying criminal suspects from surveillance photos is one way that facial recognition technology (FRT) is used in the legal system (Wehrli et al., 2022). False arrests are one of these repercussions. People who identify as members

of racial and ethnic minorities in the US suffer more detrimental effects as a result of these outcomes (Limante, n.d.). The issue arises from the fact that algorithms frequently perform less accurately when they are applied to non-white individuals. The FRT system is also being added to already-existing establishments and organizations, which exacerbates the problems with biases in facial recognition algorithms (Gentzel, 2021). To further elaborate (Gentzel, 2021), the deployment of FRT in practical environments, such as businesses or institutions, may intensify pre-existing biases in the algorithms, resulting in unjust discrimination against particular groups, notably individuals with darker skin tones, such as African Americans.

The study by (Robinson et al., 2020) shows that facial recognition technology (FRT) has the potential to exacerbate racial inequalities in policing. This study identified a significant bias in arrests made by law enforcement agencies, revealing a disproportionate representation of Black community members among those apprehended, as determined through an automated FRT. The identified biases are because of multiple factors, including inadequate representation of Black people in the training datasets utilized by the algorithms, the presumption of infallibility inherent in these systems, and the propensity for the personal biases of law enforcement officers to intensify these issues (Terhorst et al., 2022).

As facial recognition technology becomes more prevalent in law enforcement and other areas of society (Bump, 2018), concerns have arisen about the potential impact it could have on marginalized groups (Brey, 2004), particularly African Americans. It is essential to assess the sociotechnical ramifications and ethical concerns associated with the adoption of this technology by law enforcement agencies (Office, n.d.-a). This study

analyzes the existing state of facial recognition technology, any possible biases, and the wider ramifications for the African American community. Although eliminating racial profiling presents challenges (Roy, 2017) the significance of automating the tedious and repetitive facial recognition process is substantial.

2.2 AI in Law Enforcement

It is important to understand how law enforcement uses facial recognition as illustrated in *Figure 1*. Police departments across the United States have been quick to incorporate FRT into their investigative processes (Facial Recognition Gives Police a Powerful New Tracking Tool. It's Also Raising Alarms., 2018). To improve their ability to identify and monitor criminals, federal, state, and municipal agencies have partnered with FRT design firms such as Amazon and Google (The Facts on Facial Recognition with Artificial Intelligence, n.d.). Although law enforcement authorities now mostly use FRT to identify suspects from photographs taken by surveillance film or by a witness's camera, the applications of FRT will probably grow shortly (Wu et al., 2023). The government facial recognition market is expected to grow approximately threefold over the next decade. Currently, at least a quarter of state or local police agencies can use facial recognition systems for searches (Lewis & Crumpler, 2021).

The implementation of FRT by law enforcement is expected to transform its ability to identify suspects, much like the development of forensic DNA identification in the late 1980s (Tracy & Morgan, 2000). Similar to DNA evidence, FRT enables law enforcement officials to extract tiny biometric samples from a crime scene and then cross-reference them with department databases to generate a detailed profile of a suspect (Garvie & Frankle, 2016). In contrast to DNA analysis, FRT does not require the

presence of a suspect's bodily tissues, fluids, or other biological materials at the crime scene. Facial recognition technology simply requires an image of the suspect through closed-circuit television (CCTV) or other sources (Binns, 2018).

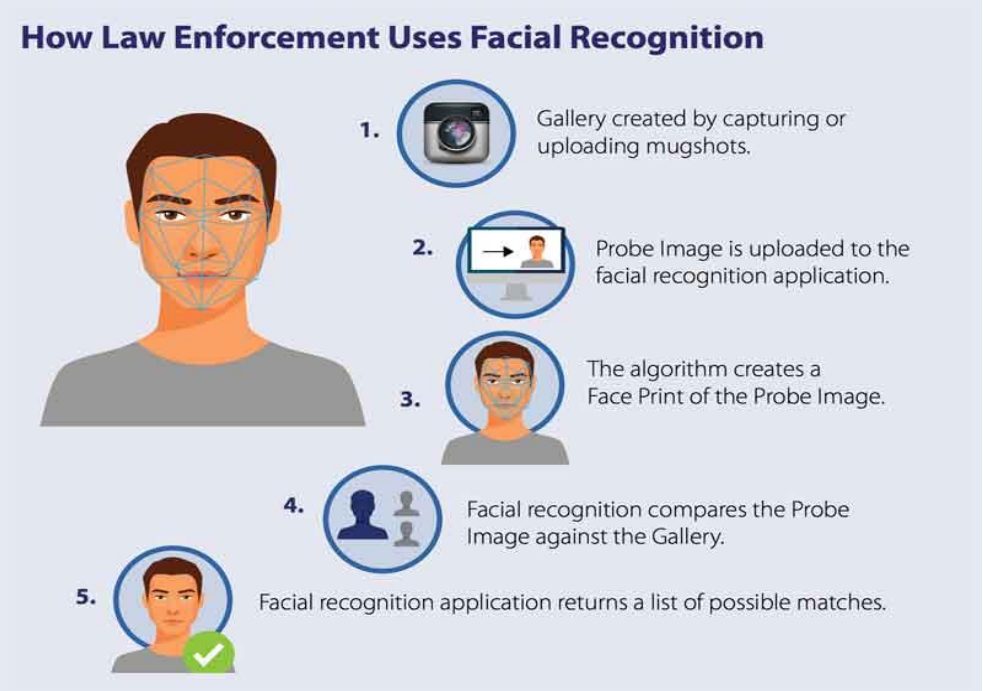
Studies on AI in law enforcement have produced inconsistent outcomes. Several scholars have expressed apprehensions over algorithmic prejudice (Stoica et al., 2018) asserting that it is both erroneous and exacerbates existing racial and gender biases in law enforcement, resulting in unequal treatment. Several police departments are employing AI to efficiently analyze large volumes of data to detect trends and patterns (Office, n.d.-a). Even more, than this, they utilize tools like ChatGPT (Bigman & Gray, 2018) to serve as investigation aides. In recent years, AI has experienced significant advancements and quick expansion. In addition, there is intense competition in the sector, as companies such as Clearview AI (Shaker, 2022), Amazon (Roach, 2018), and Microsoft (PatrickFarley, 2024) are launching new features and enhanced iterations of their FRT products. These businesses possess proficiency in developing tailored solutions for diverse situations. Notwithstanding considerable progress in deep-learning methodologies, federal tests (Office, n.d.-b) have indicated that most facial recognition algorithms demonstrate inadequate efficacy in accurately identifying individuals outside the white male demographic, as detailed by (Wu et al., 2023) in their investigation of algorithmic bias.

The majority of agencies use the same five-step procedure when identifying a suspect utilizing FRT. First, officers have a visual depiction of a criminal suspect. Second, police prepare the visual representation for inclusion in the FRT. Third, the FRT compares the visual representation against the system's database of faces, generally consisting of mugshots. The FRT generates a list of potential face matches for

an officer's examination; each match is usually paired with a corresponding confidence level, indicating the system's certainty over the identity of the suspect (Knox, 2018).

Figure 1: Diagram of Facial Recognition System

(Courtesy of Vigilant Solutions)



Finally, an officer looks over the list that the FRT generated and determines whether any of the potential matches demand further investigation (Lewis & Crumpler, 2021).

Numerous law enforcement agencies, such as the Federal Bureau of Investigation (FBI) and the New York Police Department (NYPD), see the outcomes of a Facial Recognition Technology (FRT) inquiry as merely an "investigative lead" (Nyoka et al., 2023). An "investigative lead" signifies that detectives must obtain further investigative evidence to establish the requisite probable cause for a valid arrest (Mcknight et al.,

2011). The FBI is contemplating the elimination of the "investigative lead only" protocol, permitting an FRT match to attain probable cause due to confidence in its program and anticipated expansion (Schuetz, 2021).

2.3 The Impact of Biased Training Data on African Americans

Several studies have underscored the substantial influence of biased training data on the effectiveness and equity of facial recognition algorithms (Ferrara, 2023). These biases may originate from multiple sources, including the underrepresentation of specific demographics, stereotypic labeling, or the intrinsic biases in the training datasets. Research indicates that facial recognition systems frequently demonstrate elevated error rates when analyzing images of individuals from underrepresented racial and gender groups, resulting in disproportionate inaccuracies and potential discrimination. The African American community, particularly those between the ages of 18 and 65, has experienced significant effects from the use of FRT by law enforcement. Current studies show that FR algorithms frequently encounter difficulties in accurately recognizing individuals with darker skin pigmentation resulting in elevated rates of false positives and false negatives for African Americans (Hong et al., 2022) (Radiya-Dixit & Neff, 2023) (Dooley et al., 2023). This bias may lead to the disproportionate representation of African Americans in the criminal justice system (Brodkin, 2023), along with a heightened likelihood of erroneous arrests and detentions (Allyn, 2020).

2.4 The Need for Curated Training Data

One of the primary factors contributing to bias in facial recognition algorithms is the composition of the training data used to develop these models (Dooley et al., 2023) (Merler et al., 2019). Datasets used to train facial recognition algorithms often lack

diversity, with a predominance of images representing certain demographic groups, such as Caucasian males. Lack of inclusion in the initial dataset may cause the algorithm to acquire biased patterns, leading to increased error rates for underrepresented groups.

To address this issue, researchers have highlighted the importance of carefully curating the training data to ensure a balanced representation across different demographic characteristics, such as gender, race, age, and ethnicity (Leavy et al., 2020) (Merler et al., 2019) (Kuhlman et al., 2020). By incorporating a diverse set of facial images in the training data, the algorithms can learn to recognize a wider range of facial features and characteristics, reducing the risk of biased performance and thus enhancing trust in the system. Researchers ((Ahn et al., 2021) (Kuhlman et al., 2020) (Leavy et al., 2020) (Merler et al., 2019)) have proposed several strategies to mitigate bias in facial recognition technology through the careful curation of training data, including actively augmenting existing datasets, establishing standards and best practices for data collection and curation, and fostering interdisciplinary collaboration with diverse stakeholders.

The extensive implementation of FRT in law enforcement, lacking curated training datasets, has sparked significant concerns regarding potential misuse and violations of civil liberties. These surveillance practices may make African Americans feel even more marginalized and targeted (Albiero et al., 2020) resulting in a breakdown of trust between communities and law enforcement (Kolla & Savadamuthu, 2023).

2.4.1 Challenges in Facial Recognition Data Curation

The effectiveness of FR systems is inherently dependent on the degree of accuracy and variation in training information used in their development. Current training datasets frequently lack essential diversity and representation, resulting in

biased algorithms that underperform for specific demographic groups (Fu et al., 2020). This issue is especially pronounced among underrepresented groups, including ethnic minorities and women, where the available data is frequently inadequate or biased (Parra et al., 2022).

A study analyzing the Diversity in Faces dataset (Merler et al., 2019) revealed substantial disparities in the representation of various facial characteristics, with certain features being overrepresented while others are underrepresented. The absence of diversity may result in algorithms that exhibit reduced accuracy and increased susceptibility to errors for individuals with underrepresented facial characteristics.

Moreover, studies (Chi et al., 2021) have underscored the need for greater diversity in the formulation of FR technology. The underrepresentation of marginalized groups in the technology sector can exacerbate biases in algorithms (Beauxis-Aussalet et al., 2021), as the design and decision-making processes may insufficiently address the needs and concerns of these populations.

2.5 Trust in AI

It is imperative that (Duenser & Douglas, 2023) expand on our understanding of how AI ethics principles, which have the potential to guide the creation of AI systems, are linked to people's perceptions of their trustworthiness and the circumstances in which that trustworthiness matters to them. Trust is goal-oriented and context-dependent. In this situation, a few of these elements might bolster an individual's judgments of trustworthiness and actions that are trustworthy when interacting with an AI (Ahn et al., 2021). Consequently, we argue that these issues require more in-depth discussion and

empirical research in order to prevent trust and credibility in AI systems from devolving into a vague catch-all for the "good" (Lora, n.d.). AI should be viewed as a socio-technical system in which the relationships among various stakeholders may affect perceptions of reliability, and how these relationships may strengthen or weaken trust, as well as affect how widely accepted and used AI becomes (Radiya-Dixit & Neff, 2023). Our attention to how the interactions and dynamics between several stakeholders, their trustworthiness, and whether they adhere (or are seen to adhere) to AI ethics principles will determine how trust and confidence perceptions of AI systems, developers, providers, and users, and how this influences the acceptance or adoption of AI technologies (Beauxis-Aussalet et al., 2021).

Face recognition systems serve two principal functions: one for law enforcement and the other for commercial applications. Face recognition technology is primarily employed in law enforcement, specifically in video surveillance through real-time matching of video sequences and in mugshot databases via static matching (Chellappa et al., 1995). AI ethics principles such as explainability and transparency are frequently believed to enhance user trust in the system (Reinhardt, 2023). Nevertheless, there is a lack of clear data demonstrating how these features influence users' perceptions of the system's trustworthiness. Biases can be generated and reinforced in the machine-learning process via model predictions obtained from various sources (Jiang & Nachum, 2019). Inadequate data, erratic data collection, and substandard data practices throughout various phases of the model's development can lead to bias in the model's assessments (Leavy et al., 2020). If the machine learning pipeline you are employing contains intrinsic biases,

the model will not only adopt them but also exacerbate and potentially amplify them (Ateniese et al., 2015).

AI systems ought to be regarded as sociotechnical systems, indicating that the people who designed, constructed, deployed, and utilized the system are equally significant as the system itself in assessing its trustworthiness (Radiya-Dixit & Neff, 2023). If we neglect these details, the phrases "trust in AI" and "trustworthy AI" could start to mean any good thing about AI systems (Reinhardt, 2023). Increased awareness of the adverse consequences associated with the deployment of AI systems in real-world scenarios has resulted in the development of numerous strategies to mitigate these negative effects (Meenakshi, 2013).

2.6 Explainable AI on Algorithmic Bias Mitigation

It is essential to identify, evaluate, and mitigate any potential biases that could affect the outcome when developing and maintaining a new machine learning model. In the next study, the authors (B et al., 2024) emphasized the objectivity of automated decision-making by classifying FRT as AI. This gives the misleading appearance that the outcomes generated by these technologies are devoid of errors that human eyes or thoughts frequently commit, as well as free from the biases and preconceived notions that we struggle to overcome (Raposo, 2024). Nevertheless, AI is neither a remote technology nor an entirely impartial algorithm; rather, it consists of codes authored by humans and so adheres to the guidelines established by humans (Arneson, 1999). These regulations frequently exhibit either direct or indirect prejudice or contribute to the perpetuation of inequities that our communities still maintain (Pflanzer et al., 2023).

To mitigate the risks associated with the use of facial recognition technology, it is essential to set up rigorous safeguards and oversight protocols. This may include the formulation of transparent and accountable policies (Young et al., 2019), the creation of independent oversight boards (Ughade, 2023), and the execution of stringent bias testing and audit procedures (51Testing, 2021) (Radiya-Dixit & Neff, 2023). Enhanced public awareness and community involvement can empower African Americans to assert their rights (Fjeld et al., 2020) and demand accountability from decision-makers (Bai et al., 2022).

To safeguard the rights and welfare of all people, regardless of race or ethnicity, policymakers, technology firms, and law enforcement agencies must collaborate to create and execute facial recognition systems that are fair and impartial and uphold civil liberties. Inequalities in facial recognition software can have significant consequences, particularly for underrepresented groups. Studies reveal that these systems often perform less accurately for women and individuals with darker skin tones, leading to higher rates of false positives and false negatives for these groups. Serious repercussions have followed from this, such as false arrests (Allyn, 2020) or denials of service access (Chicklas, 2021).

Mitigating bias in facial recognition technology is essential for guaranteeing that these systems are just, precise, and fair for all people, regardless of their demographic attributes (Chouldechova, 2017). To accomplish this, researchers and developers (Crawford & Calo, 2016) must collaborate to meticulously curate training data, conduct thorough algorithmic audits, and promote interdisciplinary cooperation among various

stakeholders and experts, all aimed at creating more inclusive and equitable facial recognition technology.

Moreover, the utilization of FRT relies on human judgment since it was implemented and operated by humans (Jacovi et al., 2021), thus presenting additional possibilities for human prejudice. This study (Bach et al., 2024) examines the issues arising from the potential biases and prejudices associated with the use of FRT by law enforcement authorities. (Agarwal, 2020) also explores how FRT deployment can lead to bias and discrimination. The article (Dhar et al., 2021) explores the introduction of bias into the FRT software and its potential manifestation during use, leading to adverse side effects that disproportionately impact specific population groups. The research by (Silva et al., 2016) examines the feasibility of overcoming these problems, with a specific emphasis on data and social viewpoints.

2.6.1 Data Curation as a Mitigation Strategy

One promising approach to mitigating bias in FRT involves the meticulous curation of the training data utilized for algorithm development (Kolla & Savadamuthu, 2023). By ensuring that the training data accurately reflects the diverse populations the technology is intended to serve, the algorithm can be trained to recognize a wider range of facial features and characteristics (Ferrara, 2023).

Researchers have developed a variety of strategies to improve the fairness and accuracy of FR systems through data management, as argued by (de Laat, 2018). These include:

- Actively seeking out and incorporating data from underrepresented groups to improve demographic diversity in the training dataset (Chi et al., 2021).
- Using synthetic data generation techniques to augment existing datasets and create more balanced representations (Leavy et al., 2020).
- Establishing standards and best practices for data collection and curation with a focus on ethical and responsible handling of sensitive personal data (Yarger et al., 2019).
- Fostering interdisciplinary collaboration between experts in machine learning, computer vision, and social sciences to ensure that the development of FRT is informed by a diverse range of perspectives and experiences (Shi, 2023).

By adopting these strategies, researchers and developers of FRT algorithms can work to mitigate the risks of bias, thereby ensuring that these systems are more inclusive and fairer to all people.

2.6.2 Algorithmic Auditing as a Mitigation Strategy

In addition to curating training data, an important way to mitigate bias in FRT is through a comprehensive algorithm audit (Raji et al., 2020). Systematic evaluation of algorithm performance on various datasets is crucial for algorithm auditing. This involves a thorough review of the prevalence of errors among various demographic categories and formulating strategies to address any identified biases.

Researchers have suggested various methodologies for performing algorithmic auditing, including:

- Testing the algorithms on benchmark datasets that are designed to evaluate fairness and accuracy across different demographic groups (Galhotra et al., 2017).
- Conducting sensitivity analysis to understand how the algorithms' performance is impacted by different facial features or characteristics (Roomi et al., 2011).
- Evaluating the algorithms' fairness metrics, such as demographic parity, equal opportunity, and equalized odds, to identify and address any disparities in their performance (Hardt et al., 2016).
- Implementing feedback loops and iterative refinement processes to continuously monitor and improve the algorithms' fairness and accuracy over time (Joachims et al., 2016).

Through the implementation of a comprehensive algorithmic auditing procedure, researchers and engineers can attain a more profound comprehension of the origins of innate biases in FR systems and undertake specific measures to mitigate them.

2.7 Face Recognition Biases Beyond Demographics

This article (Gong et al., 2020) evaluated two popular face recognition databases beyond demographics. The findings indicate that besides population trends, other variables substantially influence the detection effectiveness of the "FaceNet" and

"ArcFace" models. While the observed differences brought about by various features continue to be fairly consistent for FaceNet (FaceNet - Using Facial Recognition System, 2020), ArcFace's (InsightFace, n.d.) decision threshold has a significant impact on these differences. FR tools are changing how important decisions are made more and more. New research has shown that FR solutions act very differently depending on the user's background (Tolba et al., 2005). But for FR technology to be reliable, it's important to know how a lot of different facial features affect FR, not just populations. Because of this, the researchers looked at FR bias across a lot of different traits. The effect of 47 characteristics (Parker, 2023) of the verification performance of two well-known FR models is looked into. The tests were done on the MAAD-Face attribute database (Terhörst et al., 2021), which is open to the public and has over 120M high-quality attribute comments. They added control group-based validity values to figure out if the changes in performance were due to unbalanced test data so that false claims about biased performances don't get made (Hupont Torres & Fernández, 2019).

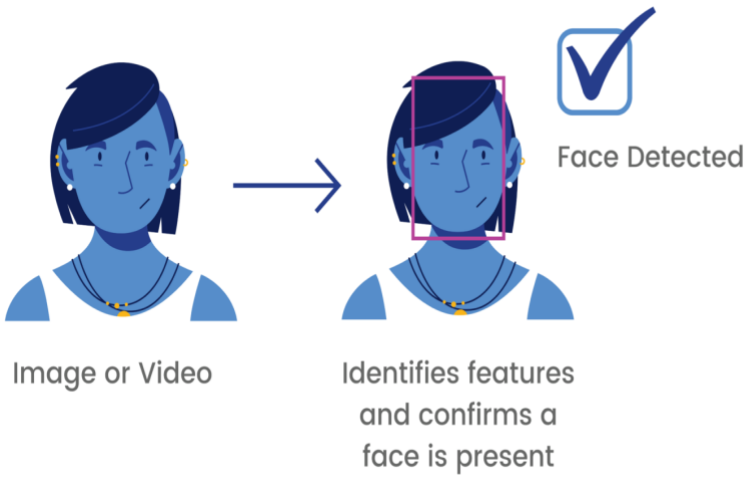


Figure 2: Blue face detection image

The findings indicate that a wide range of factors other than age or gender significantly impact an individual's recognition performance. These consist of ornaments, facial shapes, abnormalities of the face, and hairstyles and colors. Based on what was found in this study, (The Misgendering Machines, 2022) more work needs to be done to make the FR system more reliable, fair, and easy to understand. The researchers explain the large differences in performance observed. However, some observations are not clear enough and need to be further explored in future studies. The results of this study suggest that further improvements need to be made to the resilience, understandability, and fairness of FR systems (Terhorst et al., 2022).

2.7.1 "Face Recognition Biases: Too Bias, or Not Too Bias?"

The article by (Robinson et al., 2020) studied the imbalanced training sets of CNNs trained on faces recognized by face detection systems and the biases involved. This study utilized the Balanced Faces in the Wild (BFW) dataset. It consists of eight subgroups that are fairly divided by gender and race. By illustrating the challenges associated with categorizing subjects into a singular subset, they demonstrated the utility of creating subgroups, as the FR algorithm infrequently errs in these subdivisions. They purchased a Sphereface (Liu et al., 2018) and believed that this "state-of-the-art convolutional neural network (CNN)" was biased due to an inequitable training set. Upon establishing that the data was biased, it became evident that applying the same threshold across various racial and gender groups resulted in false positive rate discrepancies of up to twice the standard level. Moreover, distinct percentage disparities in the scores were evident between the subgroups.

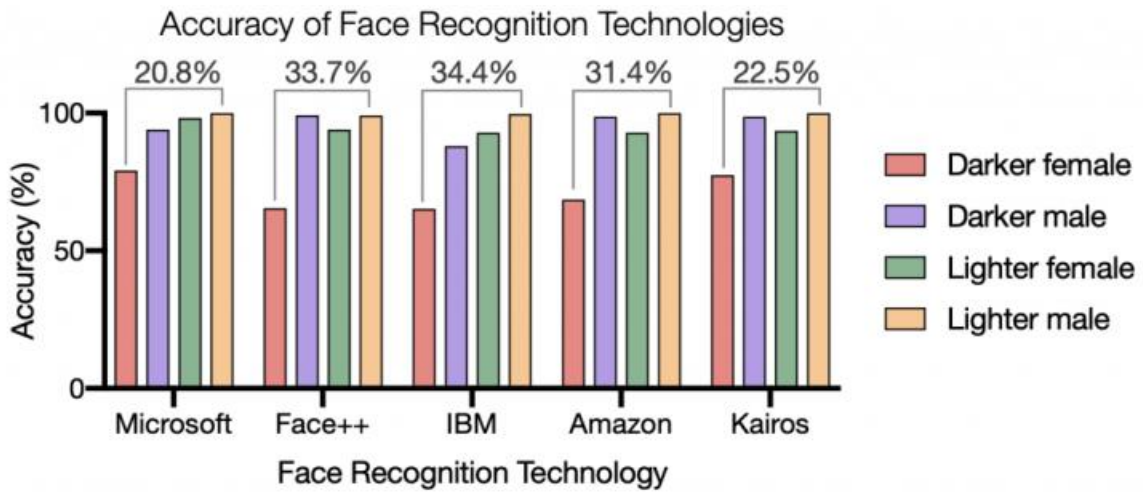


Figure 3: Racial Discrimination in FRT

(Courtesy Science in the News)

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

Equation 1: True Positive Rate

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

Equation 2: False Negative Rate

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

Equation 3: True Negative Rate

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Equation 4: False Positive Rate

Furthermore, they standardized the false positive rate (FPR) by establishing a cutoff for each subgroup, which minimized the disparities and resulted in an elevated true positive rate (TPR) (Klare et al., 2012). They believed that the majority of people were raised among diverse groups, enabling them to learn from inherently biased datasets (Wu & Bowyer, 2023). Ultimately, manual evaluations showed that individuals exhibited bias,

as most people believed their group was the most representative (Ståhl et al., 2020). The study, data, and resources may be utilized in various capacities. This is merely a minor aspect of the broader issue of bias in machine learning, as the article demonstrated its significance through data (Darabant et al., 2021).

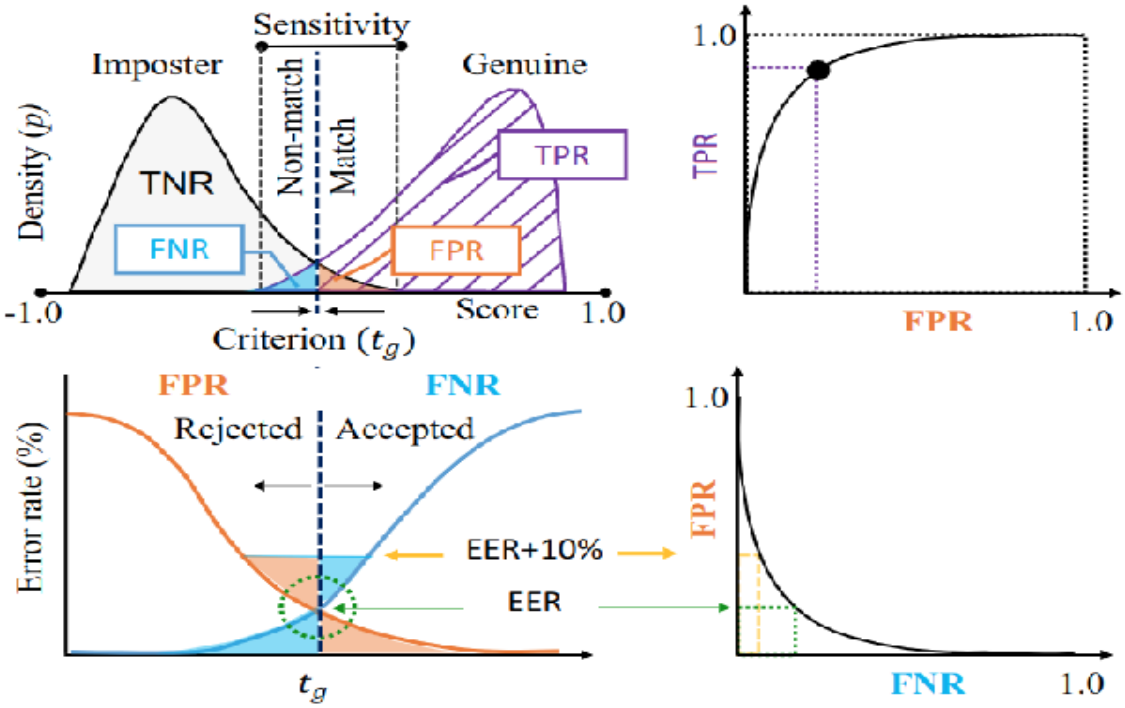


Figure 4: Depiction of biometrics.

*The signal detection model (SDM) (top-left) shows the sensitivity related to a single threshold t_g. Translated to error rate (bottom left), a direct trade-off between FNP and FPR. In practice, this is entirely application-dependent. Specifically, the specific choice in desired FPR *%) (bottom-right). (Courtesy (Robinson et al., 2020)*

2.8 Understanding Fairness of Gender Classification Algorithms

This paper investigates how different gender-race groups perform differently while using gender classification algorithms. Within this study, researchers (Buolamwini & Gebru, 2018) did experiments that showed algorithms with different architectures might not all work the same, even if they were trained on a set of people of

equal race and gender. Therefore, a specific algorithm is not to blame for the bias in the method for classifying people by gender. In all of the tests that were done, Black Race and Black women in particular had the lowest success rates. The highest accuracy rates were seen in Middle Eastern men and (Wu et al., 2023) Latina women, which was also seen in. The reason could be the way skin tones reflect (Roach, 2018) light in different conditions and the shape of the face. Bias in the training set may lead to larger differences in accuracy. Furthermore, the study found that the high mistake rates might be because black females' faces are different from those of the other females and look a lot alike with black male faces. The bias in the method for classifying people by gender is not caused by a specific algorithm, which proves my thesis (Krishnan et al., 2020).

2.9 Summary and Conclusion

Strong protections and monitoring systems must be put in place to deal with the possible dangers of FRT. Some possible steps in this direction include creating open and responsible policies, setting up impartial oversight committees, and instituting thorough procedures for testing and auditing bias. Researchers and developers can make these systems more accessible, accurate, and equitable for people of all demographics by applying these strategies.

Facial recognition algorithms could be made much more accurate and fairer by implementing the suggested strategies for minimizing bias in training data and conducting thorough algorithmic audits. Researchers and developers can strive to build more diverse and representative training datasets by proactively seeking out and integrating data from underrepresented groups, employing synthetic data generation techniques, and establishing standards for responsible data handling. Thus, this can

ensure that the algorithms can correctly identify a wider range of facial features and traits, which in turn can help reduce the risk of demographic biases.

Researchers and developers can further understand the sources of bias in their algorithms and take targeted actions to address them by implementing robust algorithmic auditing processes. These processes include testing on benchmark datasets, conducting sensitivity analysis, and evaluating fairness metrics.

Together, data curation and algorithmic auditing, along with multidisciplinary cooperation and a dedication to responsible and ethical development, have the potential to greatly enhance the accuracy and fairness of face recognition technology, to the advantage of everyone in society. A multi-pronged strategy is needed to tackle the pressing problem of bias in FRT. Researchers and developers can make these systems more inclusive and equitable for everyone by implementing rigorous algorithmic auditing procedures and meticulously curating the training data used to build these algorithms. This will help mitigate the risks of demographic biases.

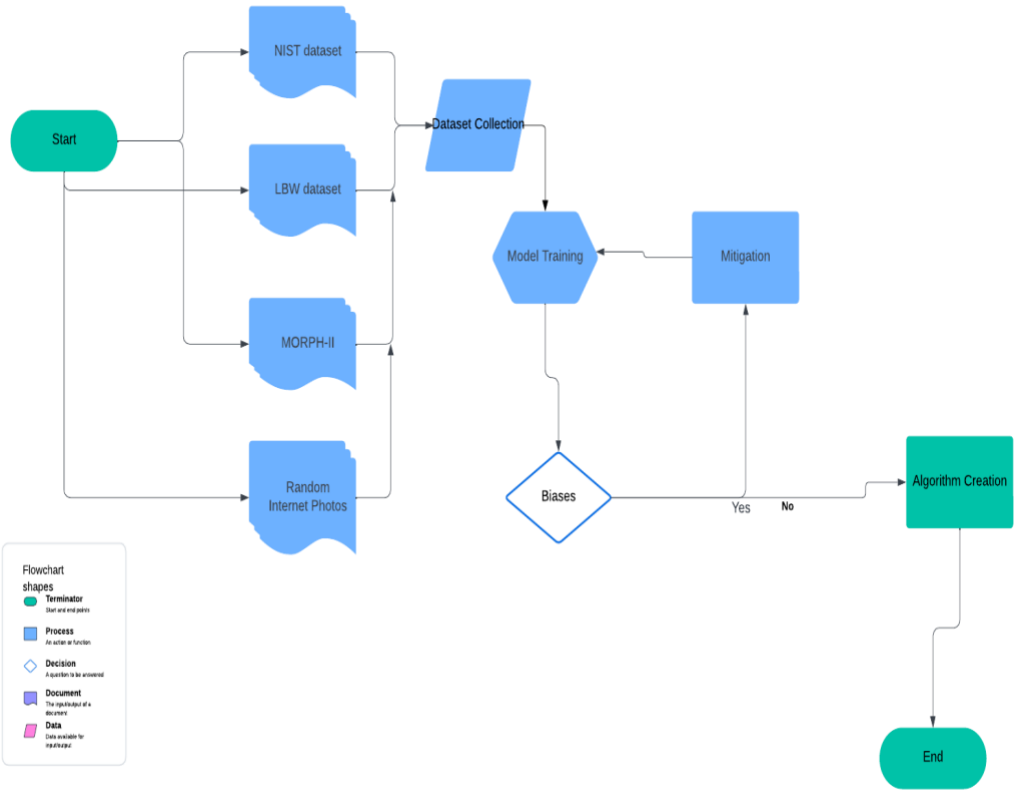
Law enforcement's extensive use of facial recognition technology has serious consequences for African Americans, especially those between the ages of 18-65. As a result of this prejudice, African Americans may be more likely to be wrongfully arrested and held in prison, and they may already be overrepresented in the criminal justice system. To address these issues, strong oversight mechanisms must be in place, equipped with the necessary tools to uphold their rights and hold those in power accountable. Strong protections and monitoring systems must be put in place to lessen the dangers that come with using face recognition technology. Some possible steps in this direction include creating open and responsible policies, setting up impartial oversight committees,

and instituting thorough procedures for testing and auditing bias. A combination of meticulous data curation for training, rigorous algorithmic auditing, and multi-disciplinary collaboration is necessary to mitigate bias in face recognition technology.

Chapter 3—Methodology

3.1 Introduction

This chapter delineates the approaches implemented to guarantee effective curation of data related to FRT algorithms used by law enforcement, specifically concerning African Americans aged 18 to 65. Below are the steps we followed to reduce algorithmic biases in the FR datasets that could be used by law enforcement.



Step 1: Data Collection. We performed our tests on the NIST FRVT dataset, Labeled Faces in the Wild (LFW) dataset, MORPH-II, and random photos from the internet. Thanks to carefully constructed deep learning models and ever-increasing information processing capabilities, modern FRTs can now learn from image datasets containing millions of faces. Our deep learning models relied on image datasets, which

served as the development and test set benchmarks. However, we found the majority of facial recognition datasets rely on crowdsourcing, which further introduces bias into the data.

Step 2: Model Training. We took the time to thoroughly investigate and analyze the dataset that we had chosen to use to test and train the models before we started developing them. We had to ensure these groups were protected and address any data bias or quality issues that may affect them.

Step 3: Implement Responsible Data Modeling Best Practices. After carefully curating our data and putting our model into production, we had to make sure that bias would not be introduced, and if it was introduced by accident, it would be quickly identified and mitigated. This was a constant process**.**

Furthermore, we will outline the assessment criteria and validation protocols employed to gauge the efficacy of our proposed approach. We outline the criteria employed to assess model performance, validate the model, and juxtapose our findings with prior research methodologies. The principal objective of our praxis study is to assess the precision of algorithmic models applicable to law enforcement while tackling concerns associated with intrinsic data biases in existing methodologies.

3.2 Face Recognition & Bias (Data Collection)

In this section, we describe how to significantly reduce algorithmic bias during data collection. Reducing bias in AI systems depends largely on the utilization of diverse, representative data. To ensure that the data fairly reflects the diversity of the target population, we collected information from many sources for our research.

Incorporating a diverse array of experiences and viewpoints in the data helped us demonstrate how to greatly reduce the dangers of bias caused by the underrepresentation of certain groups, particularly African Americans. Our work helps reduce (but not eliminate) current biases and limitations of existing datasets by collecting more diverse facial images of non-white racial groups. We empirically show that when the majority ethnic mix is not Caucasian, this results in a noticeable improvement in generalization performance to new image datasets.

3.2.1 Datasets suitable for doing experiments in deep face recognition.

There are multiple facial datasets available for training and testing deep facial recognition algorithms. Below, is specific information regarding a few of them. The MORPH-II dataset provided the study data. Current publicly available facial datasets are significantly biased towards Caucasian faces due to the huge amount of available data, while other ethnic groups (such as African Americans) are significantly underrepresented. Through the use of public sources, the MORPH-II dataset was developed in 2006 to facilitate research on face-aging. This dataset is frequently and extensively utilized by the scientific community. This research is particularly pertinent because it contains a large collection of photographs that feature African American subjects. It was discovered that the MORPH-II meta-data was extraordinarily accurate, which led to alterations in the curation process not being needed for more than five percent of the original photos. Both mugshots with a resolution of 200x240 and 400x480 are included in this collection.

Figure 5: MORPH images

Facial photographs from MORPH. The numbers next to each facial photo indicate age, ranging from 15 to 77 years, consistent with the target population of our current study. The average number of available facial photos per person was approximately four.

3.2.2 Evaluation Datasets (Model Training)

3.2.2.1 MORPH-II Dataset

The MORPH-II Dataset is a facial age estimation dataset containing 55,134 photos of 13,617 participants ranging in age from 16 to 77 years old. This dataset was used in comparative analysis to LFW and NIST FRVT.

Table 1: Number of images by Gender and Race

Table 1: Number of Images by Gender and Race (n=55,134)

Gender	African American (AA)	Caucasian (C)	Oriental (O)	Latinx (L)	Other	Total
Male	36,832	7,961	141	1,667	44	46,645
Female	5,757	2,598	13	102	19	8,489
Total	42,589	10,559	154	1,769	63	55,134

3.2.2.2 Labeled Faces in the Wild (LFW) dataset.

The LFW dataset is a substantial collection of facial images. It was specifically created to evaluate the performance of face recognition

algorithms in realistic, uncontrolled environments. The photographs have been sourced from the Internet and encompass a wide range of variances in expression, stance, age, illumination, and resolution. The LFW database has photographs of 5749 individuals, with a cumulative count of around 13000 photos. This collection contains a variety of photos with cluttered backgrounds.

Figure 6: LFW sample images

(Courtesy Labelled Faces in the Wild)

3.2.2.3 NIST FRVT Dataset.

The NIST FRVT test is critical for evaluating the effectiveness of face recognition algorithms. In our praxis evaluation, we compare LFW and MORPH-II with NIST FRVT and note that their algorithms can benefit significantly from modifications to the deployed test database and probe image format.

3.2.3 Data Bias Mitigation

All too often, public datasets contain many pairs of redundant photos that are only distinguished by different cropping or compression methods. Consequently, when evaluating image models, it is common for a single image to appear in the two collections of images gallery (training) and the test set, so skewing the accuracy metrics in the machine learning assessment. Typical datasets include low-quality images and some inaccurate labels, and the prevalence of these features varies across different demographic groups. After addressing numerous imbalances and mislabels in the NIST FRVT and LFW photos, a dataset was constructed through rigorous manual curation. This dataset outperformed earlier facial recognition bias audit datasets.

3.2.3.1 Model Data Quality

This process of interracial transformation is often used to produce or equilibrate datasets that were previously imbalanced in terms of race, thereby achieving racial parity. Altering the racial composition of the individuals seen in the images can be performed using a variety of methods, including the utilization of generative adversarial networks (GANs) or AdaBoost. As an additional recommendation, we propose that the quality of the model photos be improved in addition to these. As a consequence of this, a great deal of network enhancements were put into place, which led to a reduction in the amount of disparity in performance that existed between races. Some examples of these enhancements include loss functions, modifications to the architectures, modifications to the information, and the utilization of new data. Other examples include the

exploitation of additional data. Our goal was to demonstrate the present state of the field in terms of reducing racial bias in facial recognition, so we collected all of these contributions and studied them. Below is what we have included:

(1) When looking at both racially balanced and unbalanced datasets, it is clear many of the FR datasets, that were extensively used, were considerably lacking in terms of equality across different races.

(2) A discussion of the meanings of the procedures that were used to create new data to balance datasets, as well as an examination of the strategies themselves.

(3) An outline of the many network enhancements that have the potential to effectively reduce the error gap that exists between different ethnic groups when it comes to activities that require FR. Consequently, the tight connection that exists between racial bias and gender bias, we also discuss gender bias in the context of these three specific contributions.

3.2.3.2 Data Augmentation

One novel approach to training generative models is the use of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). This task is viewed as a supervised learning problem with two sub-models: the discriminator's job is to distinguish between actual (domain) and false (generated) instances. New instances can be generated by training the generator. Within a mutually exclusive adversarial framework, the two models were concurrently trained so the model using discrimination was deceived approximately fifty percent of the time, indicating that the model used for

generation was producing credible examples. An ever-evolving area of study, Generative Adversarial Networks (GANs) create realistic examples for a variety of problem areas, fulfilling the potential of generative models. For example, converting summer photos to winter or daylight to midnight photos is an example of image translation. Gray-scaling the images while retraining the dataset is the GANS method we used in this study.

3.2.4 Organization of CNNs

A Convolutional Neural Network (CNN) is essentially a network of interconnected artificial neurons called perceptrons, which operate together to provide parallel signal processing across a large network topology. Adjustable weights are employed to regulate the interaction among the perceptron components. An essential element of a Convolutional Neural Network (CNN) is the hidden layers, which are strategically positioned between the input and output layers of the network. The hidden layers provide the allocation of nonlinear weights to the input and guide them towards the output. That is, they assist in the application of nonlinear mathematical functions to specific components of the network to generate a desired result. For instance, in the context of a face recognition job, a single hidden layer can be optimized to detect the colors present in the input image, while another hidden layer can be optimized to recognize the specific physical characteristics of the face. Collectively, the network can identify and categorize a face inside the input image. A standard CNN consists of several layers, which can be classified into three main kinds. The three types of layers used in this context are convolution layers (CONV), subsampling layers (POOL), and fully connected

layers (FC). Typically, these layers are organized in a precise method to convert the network's input into a meaningful representation that produces an output prediction, as depicted in *Figure 7*.

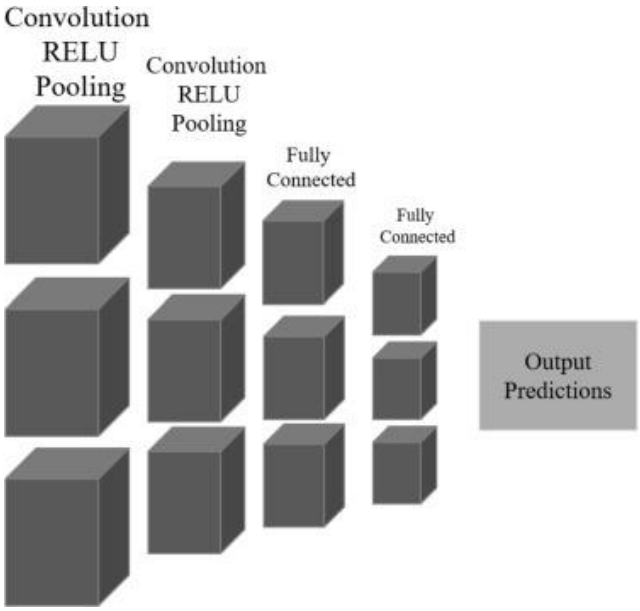


Figure 7: How a CNN is put together

It has several layers, such as fully connected layers, convolution layers, and pooled layers.

3.2.4.1 Techniques for Training CNNs

In general, there are three methods for implementing CNNs - training a network from the beginning, fine-tuning an existing model, or utilizing pre-existing CNN features. Transfer learning is the term used to describe the latter two processes. Training CNNs from scratch using the backpropagation algorithm necessitates the automated learning of a vast number of parameters, making it imperative to have a substantial amount of data. Furthermore, CNNs require significant computer resources due to their high demand for data processing. Moreover, the process entails the

fine-tuning of multiple hyperparameters. Therefore, it is uncommon to train a complete network from the beginning. Fine-tuning entails transferring the learned weights of the initial ***n layers*** from a base network to a target network, followed by the continuation of backpropagation using a new dataset. Therefore, the target network is trained using the new dataset for a distinct goal, typically unrelated to that of the base network. Fine-tuning is typically employed when the new dataset is of modest size (ranging from tens to hundreds of thousands) and exhibits significant dissimilarity from the dataset used to train the base network. Utilizing the weights from the previous network for initialization aids the backpropagation mechanism, resulting in expedited automatic learning of more specialized characteristics. When dealing with a relatively small dataset, typically consisting of a few hundred data points, even adjusting the weights through fine-tuning can lead to overfitting. Nevertheless, due to the efficient learning capabilities of CNNs in acquiring generic picture characteristics, it becomes feasible to employ a pre-trained network solely as a static feature extractor. Therefore, the process involves extracting characteristics from incoming data by mapping them onto the activations of a particular layer in the pre-trained network. Subsequently, the acquired representations are inputted into uncomplicated classifiers to address the current problem. Researchers have utilized the off-the-shelf feature extraction method to obtain encouraging outcomes (Weiss & Khoshkoftaar, 2018), (Day & Khoshgoftaar, 2017). Likewise, data

augmentation is the most straightforward approach to address the issue of overfitting faced by deep neural networks. This technique operates by artificially augmenting the dataset's size using diverse ways, such as altering the image orientation through flipping (which generates a mirrored image) and rotating the original photographs, so producing a new image. This guarantees that the learning system deduces characteristics from input with varying orientations.

3.3 Metrics

Let's examine the issue of binary classification, where we have a collection of paired training data sample:

Equation 5: Binary Classification

$\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{n}$. These samples consist of features $\boldsymbol{x} \in \mathbb{R}^m$ and labels $\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{d}}$. The objective is to discover a practical relationship $f: X \rightarrow Y$, which is determined by the parameter θ, that minimizes a specific loss function $\mathcal{L}(\boldsymbol{\theta})$across the entirety of our training dataset. Put simply, our objective is to find a solution to the optimization problem stated as follows:

Equation 6: Optimization Equation

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\theta)$$

When presented with a new test sample, $(\boldsymbol{x}, \boldsymbol{y})$, the classifier should ideally produce an output $\hat{\boldsymbol{y}} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ that is similar to a $\hat{\boldsymbol{y}}$ is close to $\boldsymbol{y}$, with the concept of similarity being determined by the original loss function. Let's suppose that each data

point is additionally accompanied by a continuous latent vector $\boldsymbol{z} \in \mathbb{R}^{\boldsymbol{k}}$ represent the concealed and sensitive characteristics of the sample.

The concept of a biased classifier can be formalized in the following manner:

- A classifier $f_{\boldsymbol{\theta}}(x)$ is considered biased if its choice changes after being exposed to more inputs of sensitive features. To clarify, a classifier is considered fair about a specific collection of underlying characteristics, $\boldsymbol{z}$, if: $f_\theta(x) = f_\theta(x, z)$. When determining whether a picture includes a face, factors such as skin color, gender, and age are all hidden variables that can influence the outcome. However, modifying the values of these variables should not impact the classifier's final judgment.

- A dataset with almost homogeneous samples across the latent space should ensure a classifier's fairness across multiple latent variables. Stated differently, the training distribution itself shouldn't be biased to underrepresent some groups while overrepresenting others. Specifically, we are arguing that within a single class, the unobserved latent variables should also be balanced, not that our dataset should be balanced concerning the classes—that is, include about the same number of faces as non-faces in the dataset. This would support the idea that all cases of a single class would be handled fairly by the classifier, such that even if a latent variable was altered to the opposite extreme (e.g., skin tone from light to dark), the accuracy of the classifier would not be altered.

In addition, if we have a test set that is labeled and covers a range of sensitive latent variables, $\boldsymbol{z}$, we may assess the bias of the classifier by calculating its accuracy for each of the sensitive categories, such as skin tone. The classifier's overall accuracy refers

to the average accuracy across all sensitive categories. On the other hand, bias is the measure of variation in accuracies observed across different realizations of these categories, such as distinguishing between light and dark faces. For instance, if a classifier achieves the same level of performance regardless of the precise occurrence of a certain hidden factor (such as skin tone), it will exhibit no variation in accuracy, and thus, be referred to as unbiased about that factor.

Conversely, if some manifestations of the underlying variable result in improved or diminished performance of the classifier, the range of accuracies will expand, thus leading to an amplification of the classifier's bias. Utilizing a collection of human-defined sensitive variables to guarantee fair representation during training necessitates laborious manual annotation of each variable throughout the whole dataset. Moreover, this approach is susceptible to possible human prejudice in determining which variables are considered sensitive or not. In this study, we tackle this issue by acquiring knowledge about the hidden factors of the category through a completely unsupervised approach. We then utilize these acquired factors to dynamically adjust the dataset during the training process. In the subsequent part, we shall delineate the architecture employed for acquiring knowledge about the latent variables.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	True Positive	False Positive
$\hat{Y} = 0$	False Negative	True Negative

The matrix that reports the number of instances in each category is commonly known as a *confusion matrix*, and many binary prediction performance evaluation metrics are derived from a confusion matrix:

1. True positive rate (sensitivity, recall, hit rate) = true positive / positive = true positive / (true positive + false negative). It measures the fraction of correctly identified positive instances among all positive instances.
2. True negative rate (specificity, selectivity) = true negative / negative = true negative / (true negative + false positive). It measures the fraction of correctly identified negative instances among all negative instances.
3. Precision (positive predictive value) = true positive / (true positive + false positive). It measures the fraction of correctly identified positive instances among all instances identified as positive.

Figure 8: Confusion Matrix

3.3.1 Datasets

Our classifiers are trained using a dataset of images, $n = 4 \times 10^5$ which includes positive examples (images of faces) and negative instances 2×10^5 (images of nonfaces). The dataset is divided into training and validation sets (80% and 20%). The NIST dataset provided positive samples, which were then cropped to a square shape according to the annotated face-bounding box. The negative examples were sourced from the LFW dataset, encompassing a diverse range of non-human categories. All photos were resized to 64×64.

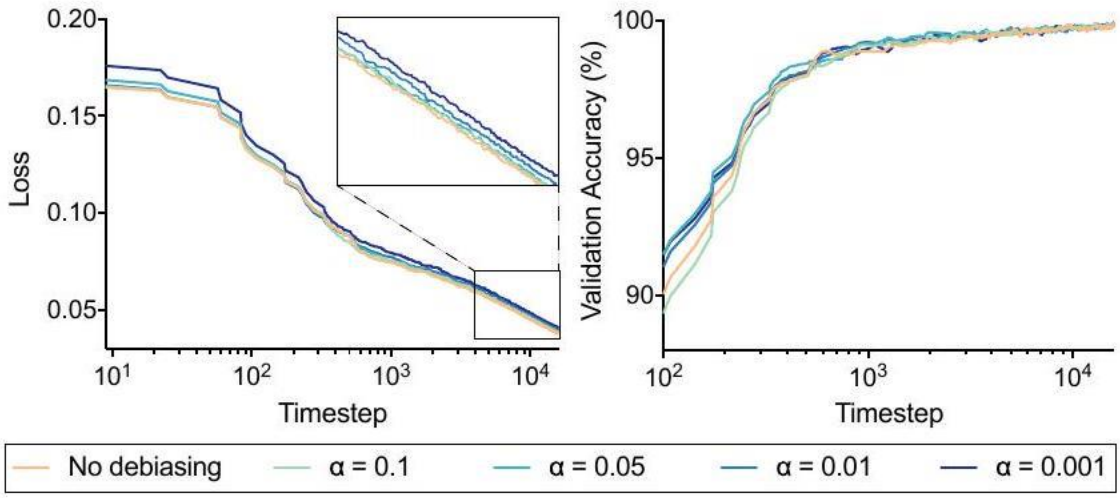


Figure 9: Analysis of the progression of losses and validation accuracy

The convergence of the overall loss on the training set (left) and the accuracy of classification on the validation set (right) for models with different levels of debiasing.

3.3.2 Model Training

To perform the traditional facial detection task, we utilize a convolutional neural network that consists of four consecutive convolutional layers. These layers use (5×5 filters with 2×2 strides) to extract features. The final classification is performed using two more fully linked layers, each containing 1000 and 1 hidden neuron, respectively. ReLU activation and batch normalization are applied to all layers in the network. To reduce bias in our CNNs, we have incorporated the technique of "grayscaling." This involves converting the training and classification input photos from color to grayscale, which in turn reduces variations in skin color. The main factor causing misclassification appears to be the difference between the images classified by the convolutional neural network (CNN) and the dataset used to train the CNN.

The presence of inequality in image processing algorithms can have different consequences. The current data standards do not meet the necessary criteria for equality. We further substantiated previous studies that have indicated that commercial face categorization software was inadequate performance when it comes to accurately identifying individuals from minority groups, especially those who are darker-skinned females. The accuracy of classification algorithms can be improved by providing additional training using only photos from the group that the system struggles to classify effectively. In the study by (Buolamwini & Gebru, 2018) they recommended that the error gap between males and females, as well as lighter and darker classifications in artificial intelligence, be closed. Since default camera settings are often designed to better expose lighter skin than darker skin tones, underexposed and overexposed photos lose critical information, leading to incorrect classification measures within the AI system. The failure of benchmark datasets to capture particular demographics may be a contributing factor in the frequent targeting and mistrust of minority groups. Inaccurate facial recognition systems often misidentify people of color, women, and young people, leading to dangerous and life-threatening situations. The authors emphasize the important need to ensure the biological and demographic accuracy of these systems to protect the public and ensure technologies remain accountable and transparent. This gave us the idea that the demographics of a CNN training dataset can change how biased a CNN facial classification algorithm is. We hypothesize that by assuring equal representation of each race in our training data, we may reduce bias in the CNNs, and vice versa. To evaluate our hypothesis, we divided our data

into three distinct racial categories based on the available labels: Oriental, African American, Latinx, and Caucasian.

Next, we validate the classifier by performing a contest elimination procedure. It will be more difficult to accurately classify a race if that race is not included in the training data. To demonstrate the impact of missing race data on classification accuracy, we exclusively train our Convolutional Neural Network (CNN) using data from only two of the available races. (For example, train using photographs of African American and Caucasian individuals, and then validate with images of Oriental, Latinx, and other individuals.) A higher representation of a certain race in the training data will result in increased accuracy when categorizing people of that race. To demonstrate the correlation between the frequency of a particular race in a dataset and the accuracy of its categorization, we exclusively train our Convolutional Neural Network (CNN) using data from that specific race. (For example, train the model using only photographs of African American people and assess its performance on images of Oriental, Latinx, and Caucasian images.)

3.4 Performance Evaluation (Modeling Best Practices)

To assess the efficacy of our curated datasets in examining the correctness of our FRT algorithms, we determine that changing a picture from color to grayscale yields varied significances.

Facial recognition is the accurate identification of faces in photos or videos based on the pre-existing database of faces. This approach consists of two stages: face detection and face recognition. The detection step is identifying the face inside an image or video

frame, typically accomplished using techniques such as the Viola-Jones algorithm and deep learning approaches, among others. The recognition step entails the specific task of recognizing or verifying the discovered face. This is often achieved by including more sophisticated machine learning and deep learning methods such as CNNs.

CNNs, such as the Viola-Jones method, break down the two-dimensional matrix of pixel values in a digital image into smaller constituents. However, rather than moving rectangular shapes over the image to find matches, they focus on specific sections of the image by employing tiny grids of pixel values known as kernels. The kernels generate feature maps by systematically traversing the entire image, attempting to identify matches for the specific feature that each of them has been trained to detect in the neural network.

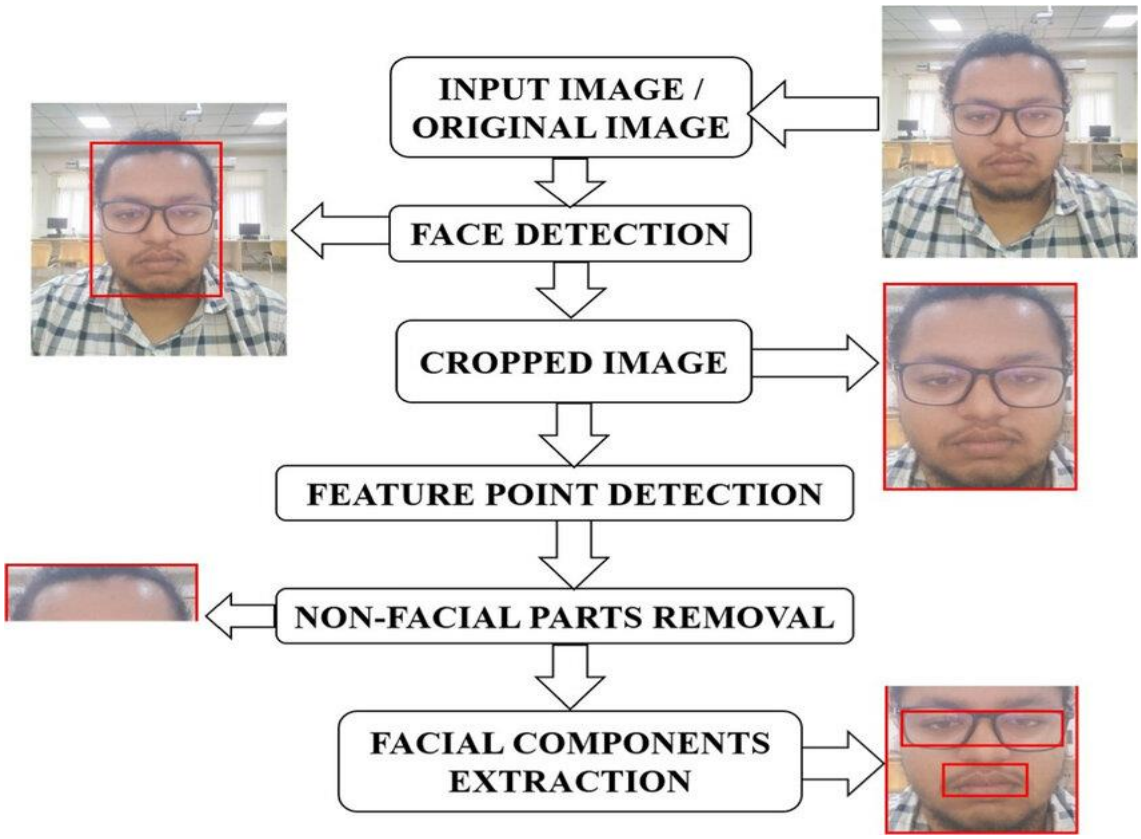


Figure 10: Flowchart of working Viola-Jones algorithm

3.4.1 Analysis of FMR and FNMR

The FMR is a crucial parameter for assessing the precision of a biometric system. Once the right FMR range has been chosen for the job, representatives from different datasets can be used to compare possible matches and find the one with the lowest correlated FNMR (or highest validation rate). The ROC curve is an often-utilized instrument tool in the comparative evaluation of competing algorithms on the same dataset, as it intuitively represents the disparity among false negative rate (FNMR) and false positive rate (FMR). Given the set false match rate, it is easy to determine which matcher exhibits the highest true match rate.

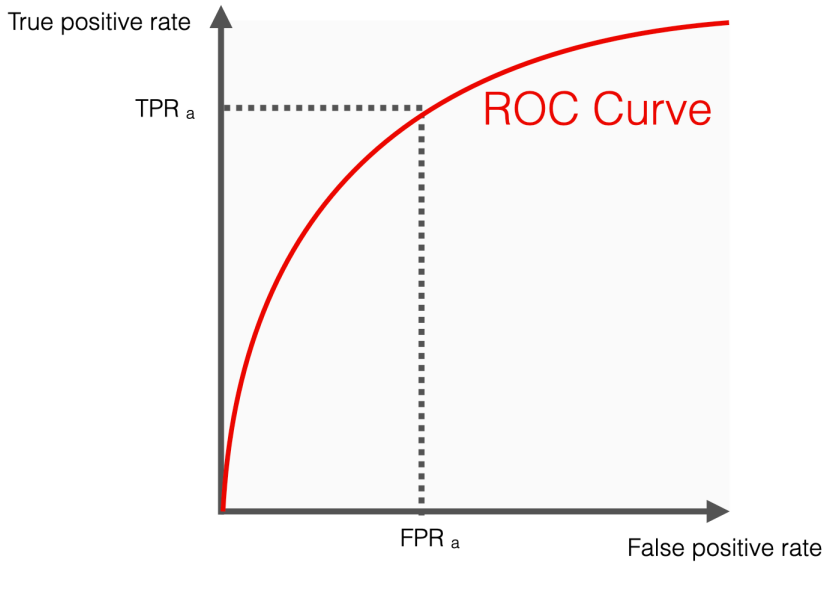


Figure 11: ROC Curve

(Courtesy Evidently AI)

We are currently improving the accuracy metrics of facial recognition systems, regardless of the latest technology and computational environments. At present, virtualization and artificial intelligence are employed to enhance the efficiency of facial recognition and decrease time complexity. In the proposed framework, we tackle these difficulties with meticulous attention. The classifier performance evaluator does multiple performance assessments and classifies the FR accuracy into true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). Accuracy is a particularly compelling yet intricate metric, appropriate for comprehensive comparisons between individual classifiers and the proposed system. There are a total of four possible outcomes, given the two test predictions («accept» or «reject») and two true outcomes («accept» or «reject»):

Figure 12: Types of errors.

Blue means "accept" and yellow means "reject". The background color shows the real values, and the image box color shows the decision made by the algorithm. - (Courtesy NTechLab)

The other two results go by various names, but two of them are true. Sometimes called false accept, false positive, or false match, Type I errors are those that don't match the data. Type II errors are those that don't match the data.

3.4.2 Processes for Ensuring Fairness in Model Training

Throughout the training and testing phases, we scrutinized fairness measures with the same precision as we assessed model correctness. We assessed whether the model exhibited bias against any demographic, particularly focusing on our target protected group of African Americans aged 18 to 65. Statistically, facial recognition accuracy ought to exhibit minimal variance among the groups. This evaluation can be validated using a z-test due to the sample size exceeding 50. We determined that all classifiers exhibit superior performance on male faces compared to female faces, with a difference in error rate ranging from 8.1% to 20.6%. All classifiers exhibit superior performance on lighter faces compared to

darker faces, with a discrepancy in error rate ranging from 11.8% to 19.2%. All classifiers exhibit inferior performance on darker female faces, with an error rate ranging from 20.8% to 34.7%.

Chapter 4—Results

4.1 Introduction

This investigation aims to evaluate the effectiveness of curating facial recognition algorithms informed by empirical evidence of bias, specifically targeting African Americans aged 18-65. The retraining of these algorithms may enhance the ability of US law enforcement to accurately identify potential criminals. We used two widely recognized face recognition models in our experiments: 1) NIST FRVT and 2) LFW. We generated facial encoding for specific face images in the FRVT dataset, the image underwent alignment, scaling, and cropping. The images received preprocessing, followed by labeling and classification according to their positions, specifically front, left, right, etc. The preparation process of the LFW dataset was the same as that of the FRVT dataset. Both models employed a CNN architecture and were trained via Adaboost within the Viola-Jones algorithm for identity verification. We developed Python code with the Tkinter library to create a Predictive Code Completion Model for executing our environment. In this chapter, we will present the results of our models' assessment and how well they perform under data augmentation.

Furthermore, a key element of our study was to compare the results of our curation approach with those of previous studies using similar techniques. This comparison seeks to understand the potential advantages or drawbacks of similar data augmentation in relation to its centralized counterpart in the context of mitigating

algorithmic biases. We hope our findings will open doors for future research efforts and contribute efforts in further mitigating FR algorithmic biases.

4.1.1 Obtaining NIST Dataset

NIST tested nearly 200 facial recognition algorithms from nearly 100 developers to assess demographic differences, using four sets of images that included more than 18 million images of more than 8 million people. The supported research aims to build face recognition systems. The FERET database was established to facilitate sponsored research and FERET evaluations. Before the FERET program, it was impossible to accurately test or compare facial recognition algorithms. The majority of US law enforcement agencies use this dataset.

Instructions on getting the FERET database:

> **To obtain the color FERET database**, request an account by following the steps:
>
> **1.** Read the Release Agreement
>
> **2.** Email the form below in plain text:
>
> *To: colorferet@nist.gov*
>
> *Subject: color FERET download request*
>
> *Please create an account that will allow me to download the color FERET database. I will abide by the Release Agreement version 1.*
>
> and we will email back a reply detailing how to use the account.
>
> Once approved, if you cannot access the download:
>
> You can install **wget** by downloading its (free) source code from:

- http://ftp.gnu.org/pub/gnu/wget/ and compiling it. Or you can use a search engine to locate a pre-compiled version of **wget** for your platform.

Or you can download the 6 files (5 data files and 1 file containing their MD5 checksum) manually using any web browser.

4.1.2 Obtaining LFW Dataset

Labeled Faces in the Wild (LFW) is a collection of pictures of faces that were made to help researchers study the problem of recognizing faces without any restrictions. Researchers at the University of Massachusetts, Amherst, make and keep up with this database. The Viola-Jones face analyzer examined 13,233 photos of 5749 people online and placed them in the center.

This dataset is made up of 11 files. The images are in the file: 1fw-deepfunneled.zip. The other 10 files are information that is useful and could help you make training and testing sets for your model.

How to Download Kaggle Dataset

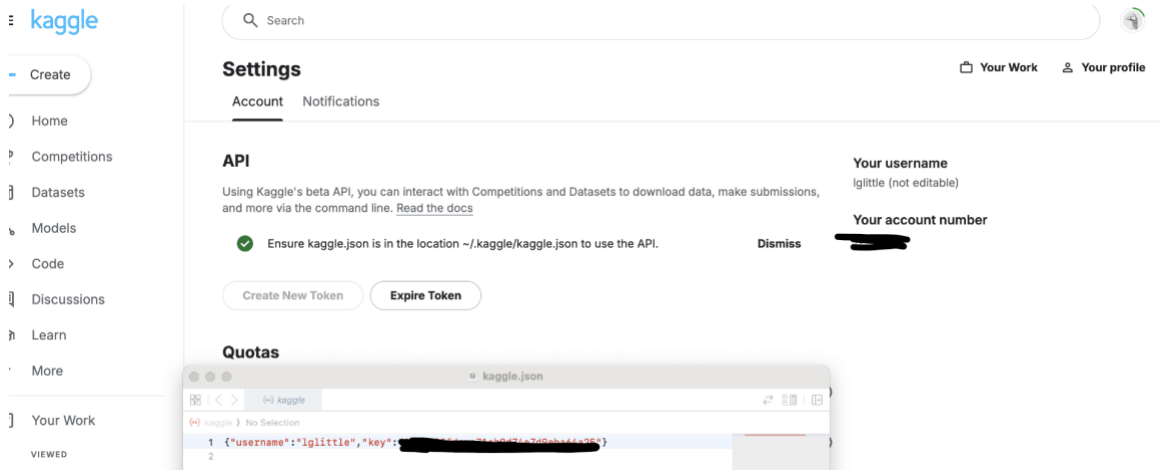
Method 1: Jupyter Notebook

Step 1: Download and Install the required packages.

```
!pip install opendatasets

!pip install pandas
```

Step 2: Visit the Kaggle website (www.kaggle.com). Go to your profile by clicking the picture, then settings, then API. Finally, click Create New Token.



Step 3: Open your Jupyter Notebook, import the opendatasets library and download your Kaggle LFW dataset by pasting the link.

```
import opendatasets as od

import pandas

od.download("https://www.kaggle.com/datasets/jessicali9530/lfw-dataset)
```

On the next field: Provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds

Your Kaggle username: lglittle

Your Kaggle Key: [*redacted*]

```
import opendatasets as od

import pandas

od.download("https://www.kaggle.com/datasets/jessicali9530/lfw-dataset)
```

Step 4: Now we are ready to use the dataset.

Method 2:

Manually downloading the Kaggle dataset

Step 1: Visit http://www.kaggle.com and select the Dataset tab

Step 2: Search for the LFW dataset and Click Download

Step 3: The downloaded file will be in ZIP form, unzip it

Step 4: Upload to your Jupyter notebook

4.2 Executing Dataset Learning Environment

We utilized the Python Tkinter software library to generate a total of 93 Convolutional Neural Networks (CNNs). Each of our Convolutional Neural Networks (CNNs) comprised three sets of convolutional layers, which were subsequently followed by batch normalization and max pooling layers. The three groups adopted the Rectified Linear Unit (ReLU) activation function due to its good empirical performance and computationally simple derivatives. After the three groups, there was a subsequent group consisting of dropout layers.

```
import tkinter as tk
from tkinter import ttk
import tensorflow as tf
from tensorflow.keras import layers, models

def generate_cnn():
    """Generates a CNN model with random hyperparameters."""
```

The dropout layers aim to mitigate overfitting by nullifying a portion of the connections between layers during each iteration. The last group consists of a dense layer or a fully connected layer to transform the features into categories.

We utilized images sourced from the NIST Database to train our Convolutional Neural Networks (CNNs). The dataset comprises 8 million high-resolution images of people in different poses, occasionally facing the camera directly against a plain white background. The images depict people wearing various accessories such as hats, scarves, and eyewear. The subject's race, gender, and emotion were each assigned labels to the corresponding image.

A total of 26 Convolutional Neural Networks (CNNs) were trained and used to classify color images, while another set of 26 CNNs was trained and used to classify grayscale images. In each group, four Convolutional Neural Networks (CNNs) were trained exclusively on a single race for a duration of 15 epochs. Additionally, four CNNs were trained exclusively on a single race for a duration of 20 epochs, while another four CNNs were trained on all races except one for a duration of 20 epochs. The 'baseline' algorithm consisted of 26 Convolutional Neural Networks (CNNs) in each group.

We trained these CNNs on each of the four races using a five-fold cross-validation method, for a total of 40 epochs. This algorithm was designed to achieve

optimal performance and minimize bias.

```
from keras.models import Sequential
from keras.layers import Dense, Flatten, Conv2D, MaxPooling2D

model = Sequential()
model.add(Conv2D(32, (3, 3), activation='relu', input_shape=
(face_images.shape[1:])))
model.add(MaxPooling2D(2, 2))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(2, 2))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(2, 2))
model.add(Flatten())
model.add(Dense(1024, activation='relu'))
model.add(Dense(class_count, activation='softmax'))
model.compile(optimizer='adam', loss='categorical_crossentropy',
metrics=['accuracy'])

model.fit(x_train, y_train, validation_data=(x_test, y_test),
epochs=20, batch_size=10)
```

Subsequently, we employed our Convolutional Neural Networks (CNNs) to categorize images from the LFW dataset based on gender. The irregularity of these images was significantly greater compared to those from the NIST Database. The images showcase various backgrounds, multiple colors, faces shot at angles, full body frames, and lower image resolution. We employed the NIST dataset for testing to enhance the simulation of performance in contexts that closely resemble real-world scenarios.

4.3 Details of Construction and Composition

The database construction process can be delineated into the following steps:

1) Collecting raw images; For our research, not only did we gather images from both NIST FRVT and LFW databases, but we also augmented the datasets with overfitting images from the internet.

2) Executing a face detection algorithm and manually removing false positives.

```
class SecondWindow(Toplevel):
    def __init__(self, parent, selected_file, recognizer):
        super().__init__(parent)
        self.title("File Selected")
        self.geometry("800x600")
        self.selected_file = selected_file

        # Recognize Selected Face
        id, confidence = recognize_face(recognizer, selected_file)
        filepath = ""

        text = "Face not Recognized"
```

3) Discarding duplicate images.
4) Annotating the identified individuals.

```
        if confidence > 90:
            text = f"Recognized Face --- ID: {id}"
            filepath = os.path.join("lookup_data", str(id))
```

5) Cropping and rescaling the detected faces, and

```
import numpy as np
import pandas as pd
from sklearn.datasets import fetch_lfw_people

faces = fetch_lfw_people(min_faces_per_person=100, resize=1.0,
slice_=(slice(60, 188), slice(60, 188)), color=True)
class_count = len(faces.target_names)

print(faces.target_names)
print(faces.images.shape)
```

6) Creating training and testing pairs for the different iterations run on the datasets.

```
hist = model.fit(x_train, y_train, validation_data=(x_test, y_test),
batch_size=10, epochs=10)

acc = hist.history['accuracy']
val_acc = hist.history['val_accuracy']
epochs = range(1, len(acc) + 1)
plt.plot(epochs, acc, '-', label='Training Accuracy')
plt.plot(epochs, val_acc, ':', label='Validation Accuracy')
plt.title('Training and Validation Accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend(loc='lower right')
plt.plot()
```

4.4 Dataset Evaluation

To obtain dependable assertions on the influence of various features on FR, it is essential to have a database that 1) offers a substantial quantity of facial photos, 2) includes numerous attribute annotations, and 3) maintains good quality. We selected the

publicly accessible Labeled Faces in the Wild database, utilizing photos from VGG-Face, as it meets our experimental criteria. The LFW dataset comprises more than 13,233 facial pictures sourced from the internet. It comprises 5,749 IDs, including 1,680 people with two or more photographs. A comprehensive list of the LFW annotation qualities is included in Table 2, detailing the number of photos that possess (positive) and lack (negative) these attributes.

Facial Recognition Models: We use two widely recognized face recognition models for the experiments: 1) GhostFaceNetV2-1 (LFW) and 2) NIST FRVT.

Attribute	Negative	Positive	Attribute2	Negative3	Positive4
Male	7484	5749	Round Face	50	6956
Young (18-30)	3289	9024	Double Chin	7720	82
Middle-aged (31-49)	1190	4412	Obese	0	0
Senior (50-65)	8789	32	Dark Eyes	9076	552
Asian	10321	129	Light Eyes	105	7035
Caucasian	621	11196	Bushy Eyebrows	12107	5
African American	9521	1532	Arched Eyebrows	33	7277
Bald	7691	59	Mouth Open	6333	874
Hair	583	8320	Mouth Closed	1036	6333
Glasses	6291	30	Smiling	1009	262
No Glasses	30	6291	Frowning	1036	6333
Hat	7277	3	Full Lips	93	3782
No Hat	0	7277	Wide Nose	2	48
Beard	3302	15	Pointy Nose	86	90
No Beard	3302	0	Earrings	2811	4289

Oval Face	10234	10	No Earrings	209	5297
Square Face	8922	48	Necktie	8402	4
Any Facial Hair	2798	26	No Necktie	4	8402

Table 2: Label distribution for LFW attributes

In the following investigation, we examine the essential attributes of ethnicity and gender. The Labelled Faces in-the-Wild (LFW) dataset was designed to examine racial bias in facial recognition (FR) systems. Each photograph is categorized into one of the four classifications: Caucasian, Oriental, African American, and other, denoting four separate ethnic groupings. Each subgroup has around 10,000 photographs of 3,000 people for facial verification. Furthermore, we assigned a gender identity to each photograph alongside the race designation. Our reports suggest that the model attains an accuracy of about 99.8%. The gender prediction is assessed using a continuous integer score ranging from 0 to 1, where lower values indicate men and higher values indicate women. The gender score was divided into fixed clusters of samples based on a threshold of $\boldsymbol{s_{gender}} < 0.5$ for males and $\boldsymbol{s_{gender}} > 0.5$ for females. Using a dataset consisting of at least 4 face photos per person**,** each individual was classified as male or female based on their average score. The representation of Black women is significantly low, accounting for fewer than 1% of all samples.

Table 3: Presents a summary of the number of samples from each cluster categorized by ethnicity and gender

	Caucasian	**other**	**Oriental**	**African American**	**Total**
Male	9621	8312	7641	23,198	48,772
Female	6872	2965	6110	745	16,692
Total	16,493	11,277	13,751	23,943	65,464

4.4.1 Dataset Balancing

Succeeding the assessment of each dataset, before the training phase, we conducted a comprehensive evaluation of every set. The initially downloaded dataset was extracted and systematically divided into more than 93 CNNs. The labeling and training of these sets were conducted exclusively using the information obtained from the NIST and LFW coding. Our findings suggest a significant absence of representation for African Americans. The dataset's accuracy was evaluated, yielding an error rate of 42.3% for African American women and 34.7% for African American men. The dataset was rebalanced by increasing the representation of African Americans overall, followed by retraining. The implementation of data augmentation led to a significant improvement in dataset performance, achieving a confidence level of 90%. We determined a confidence level exceeding 90%, which allowed our model to identify faces with complete accuracy, achieving a 100% success rate.

```
class SecondWindow(Toplevel):
    def __init__(self, parent, selected_file, recognizer):
        super().__init__(parent)
        self.title("File Selected")
        self.geometry("800x600")
        self.selected_file = selected_file

        # Recognize Selected Face
        id, confidence = recognize_face(recognizer, selected_file)
        filepath = ""

        text = "Face not Recognized"

        if confidence > 90:
            text = f"Recognized Face --- ID: {id}"
            filepath = os.path.join("lookup_data", str(id))
```

Our objective in balancing this dataset is to enhance the model's performance by removing any bias towards the class with a greater number of samples.

See **Appendix B** for Minitab results: Analysis of Variance, Goodness-of-Fits Tests, Regression Equations, and Confusion Matrix results.

4.5 Bias Mitigation by Retraining FR Algorithms vs. Results from Related Studies

Papers	Citation	Model Used	Dataset	Accuracy
Balancing Biases and Preserving Privacy on Balanced Faces in the Wild	(Robinson, et.al., 2020)	CNN	BFW	AA-M=0.971 AA-F=0.955
Face Recognition Accuracy Across Demographics: Shining a Light into the Problem	(Wu, et.al., 2023)	CNN (FR Model)	PIE & Multi-PIE; LFW	AA-M=0.24 AA-F=0.28
Fly in the Face of Bias: Algorithmic Bias in Law	(Schuetz, 2021)	CNN	LFW	AA=0.97

Enforcement's Facial Recognition Technology and the Need for an Adaptive Legal Framework				
Proposed Curation Model		CNN	NIST FRVT, LFW, MORPH-11	AA-M=0.999 AA-F=0.998

Table 4: Comparative Results – Bias Mitigation vs. Results from related studies.

Within the segment that came before this one, in the literature review, we reviewed previous studies and discussed them in depth. We found some works to be particularly influential, including Face Recognition Accuracy Across Demographics: Shining a Light into the Problem by Wu, et al.; Fly in the Face of Bias: Algorithmic Bias in Law Enforcement's Facial Recognition Technology and the Need for an Adaptive Legal Framework by Schuetz; and Balancing Biases and Preserving Privacy on Balanced Faces in the Wild by Robinson, et al. To lay the framework for our discussion on the comparative strengths and limitations of our proposed training data curation methodologies, let us examine these studies.

4.5.1 Summary of Key Studies from Literature Review

-Study by Robinson, et al. (2020): This research study was focused on showing that FRT has the potential to exacerbate racial inequalities in policing. This study identified a significant bias in arrests made by law enforcement agencies, revealing a disproportionate representation of Black community members among those apprehended, as determined through an FRT. The main findings of the experimental results on the

NIST FRVT, MORPH-11, and LFW datasets proved if our methodology was used in curating the datasets within those algorithms, the accuracy of FRT correctly identifying African Americans would be near 100% (0.99). Our model curation method outperformed the one used to train the datasets used in this evaluation.

-Study by Wu, et al. (2023): This research examined how variations in facial illumination may contribute to discrepancies in face recognition accuracy among different demographic groups. This study revealed minimal differences between men and women, as both sets of imitation photo pairs had faces that were either too dark or excessively bright, resulting in elevated false match rates (FMR). They suggested that illumination modifications for each individual might decrease those FMR. As this study was a first of its kind, being unprecedented, we did not validate or refute its findings. This may warrant additional investigation in future studies.

-Study by Schuetz (2021): This study advocates for legislative bodies to establish legal frameworks that address the troubling implications of FRT. This study helped me stay "grounded" and concentrated on the consequences of false-positive results as well as the drive to highlight necessary changes.

Our proposed training data management strategy improves accuracy on some of the top FR datasets by achieving near-perfect accuracy for African Americans aged 18-65.

4.6 Research Findings and Hypotheses Validation

This study aims to systematically explore the subsequent research questions and hypotheses:

RQ1: Are facial recognition algorithms racially biased?

RQ2: Can a facial recognition algorithm model be built to improve accuracy and reduce built-in biases?

RQ3: Can the unjustness brought about by AI misidentification be mitigated by incorporating human oversight?

HQ1: Assuming the datasets used to train AI technology are biased, mostly using white male faces to train FR systems while ignoring faces of different races, the technology will unavoidably produce biased conclusions

HQ2: With 90% accuracy, the facial recognition algorithm model can be retrained to correctly classify faces as Good or Bad Quality.

HQ3: Human involvement in data labeling, thorough data testing, and ensuring a diverse sample representative of the African American community effectively illustrate that basic best practice protocols in FRT mitigate biases.

Research Question One:

For the dataset containing only race, we utilized recorded statistics. We randomly selected 3,000 samples from each dataset and labeled them with ethnic identifiers. As expected, the majority of current face attribute datasets, particularly those focused on African Americans, exhibit bias. Our investigation shows that facial recognition algorithms are racially biased. *Figure 14,* from the NIST Assessment of Face Recognition Technology, highlights the persistent disparities in how face detection and recognition algorithms are applied to African Americans. The algorithms showed considerable

disparities in accuracy between genders, with the most notable error rate reaching 30 percent, particularly for the faces of women of color

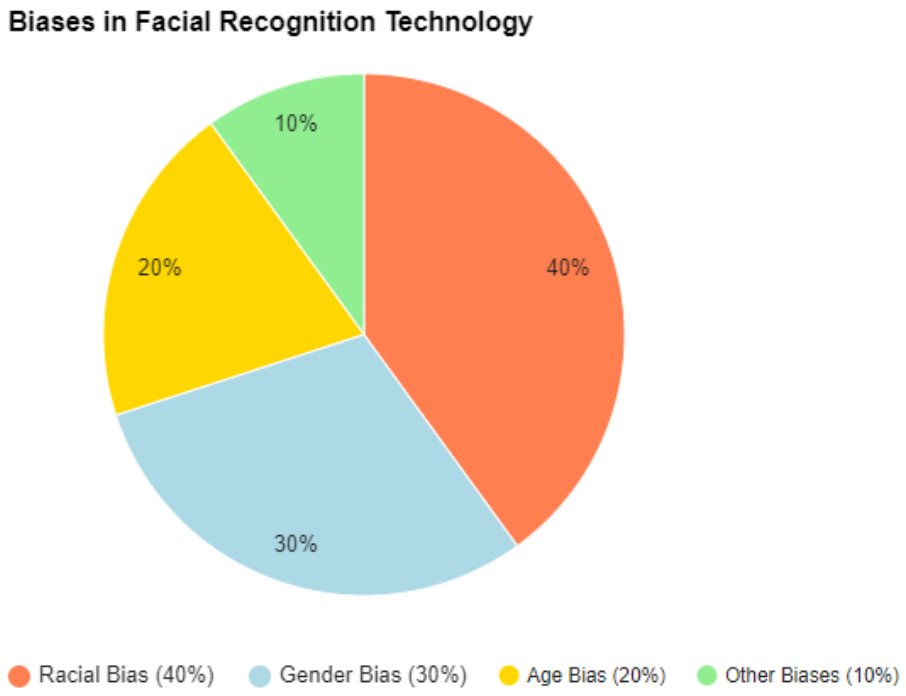


Figure 13: Biases in Facial Recognition Technology

(Courtesy NIST FRT)

Research Question Two:

To make a facial recognition algorithm model that is more accurate and less biased, it is important to carefully choose training data that is diverse and representative, use fair learning methods like "sensitive loss," and check for biases regularly while the model is being built.

Essential considerations for mitigating bias in FR algorithms:

- Diverse training data: The paramount component in mitigating bias is the use of a training dataset that encompasses a diverse array of faces. We found this to be the case within our testing and immediately augmented with the representative group of African Americans between the ages of 18 and 65.
- Data augmentation: Creating synthetic faces with different features helps to supplement the diversity of the training dataset.

Research Question Three:

Human oversight can alleviate bias in facial recognition technology by scrutinizing the algorithm's outputs, ensuring decisions are not exclusively reliant on the automated system, and by vigilantly observing for bias patterns in the data and results, facilitating necessary system modifications; this encompasses training personnel to identify and correct potential biases in facial recognition matches, especially in sensitive contexts such as law enforcement.

We meticulously examined each image to accurately categorize our models according to their positioning, such as facing front, left, right, and whether the glasses were on or off. We accurately identified African Americans based on morphological characteristics, including hair texture, skin color, and nose and lip size. By including human assistance throughout the data labeling phase of our model configuration, we achieved a substantial enhancement in accuracy, resulting in a model performance with a confidence level of 90 percent or higher. Although this step is time-consuming and often overlooked, it ensures high-quality labeled data, which is critical to training our effective facial recognition algorithms.

Hypothesis Validation:

Hypothesis One:

It is a well-documented fact that African Americans face a higher likelihood of misidentification or lack of identification by facial recognition systems, as evidenced by numerous studies, including my research. I have demonstrated the approach to mitigate this bias through the careful curation of data, involving significant human intervention during the labeling process.

Initially, the software developer can address the numerous reasons why this theory is valid. The data-driven variant addresses the commonly held problem that image training datasets are inherently biased. Images must be utilized to accurately represent the group. The method employed in training images needs to ensure a uniform standard of quality throughout various competitions. It is essential to ensure sufficient representation of all racial groups, particularly African Americans aged 18-65, in the algorithms used for facial image computation.

Hypothesis Two:

A facial recognition algorithm model exhibiting 90% accuracy can be reconfigured to categorize faces as Good or Bad quality, thereby determining the suitability of a facial image for recognition based on criteria such as clarity, lighting, and angle, potentially achieving a similar accuracy level of about 90%, contingent on “the training data and the quality of the new classification system.”

This research was conducted by revising the training dataset. We trained the model on a fresh dataset explicitly categorized as Good or Bad quality faces, enabling it to discern the characteristics that distinguish high-quality photos from low-quality ones. We also modified the algorithm to extract pertinent features such as sharpness, contrast,

lighting conditions, and facial occlusion to assess the quality of a facial image. We also analyzed the FR systems that utilize color images, as they were shown to exhibit greater sensitivity to biases associated with skin color variations, particularly when accurately identifying people with darker skin tones.

Grayscale images eliminate color information, thereby reducing the impact of skin color changes on recognition accuracy. Training in grayscale may mitigate bias; however, it is crucial to uphold a diverse dataset that appropriately reflects all ethnicities and skin tones to enhance accuracy for everybody, particularly for African Americans.

Hypothesis Three:

Human intervention in data labeling, comprehensive testing of data, and ensuring a diverse sample representative of the African American community effectively demonstrated that simple best practice protocols can significantly improve the algorithmic accuracy of facial recognition technology.

Mitigating bias in facial recognition algorithms is critical to achieving fair and equal outcomes. Developers and organizations can take bias mitigation measures by training algorithms on different datasets covering various demographics. Continuously auditing and evaluating the accuracy of these systems across different populations can reduce the risk of misidentification and ensure that the technology is fair and reliable for everyone.

4.7 Ethics in FR

There are two important ethical elements to this work. A clear issue emerges as to the potential for racial bias in facial classification systems. Underneath the surface exists a discussion regarding gender and its changing importance in modern society. Biased

training data can penalize underrepresented groups. Unfortunately, as shown by (Buolamwini & Gebru, 2018) (Esteva et al., 2017), this is generally the characteristic of the data available to engineers and scientists across various fields. Racial minorities and other structurally disadvantaged groups seem to have a lower representation in prestigious academic institutions, research laboratories, and various data-collecting organizations. The data generated by these organizations indicates their lack of presence in that context. It is essential for engineers and scientists across various fields to recognize this trend and actively strive to reduce its influence on their work. An instance illustrating the consequences of image processing occurred in 2015 with Google, whose image classification algorithm erroneously categorized photographs of Black people as "Gorillas" (Vincent, 2018). Historically, the notion of racial superiority has frequently been linked to a communicative approach that seeks to dehumanize people of color or categorize them as outsiders (Kuhlman et al., 2020). When erroneous data is unchecked and used to develop flawed algorithms, it destructively perpetuates a harmful narrative.

The introduction of machine learning techniques in the criminal justice system has intensified the consequences of biased data. Judges have begun utilizing criminal risk assessment algorithms to aid in their sentencing decisions (Roy, 2017). The algorithms produce a recidivism score through the examination of a convicted offender's profile, providing a numerical estimate of the likelihood that the person will engage in criminal behavior again. The effectiveness of these algorithms in addressing biases in the criminal justice system remains a subject of contention. Although algorithms can potentially render decisions free from the biases that human judges and juries may possess, this does not ensure that their outcomes are unbiased. Although they can improve our judicial

process, they still exhibit considerable shortcomings. The real danger lies in our inclination to assume that these algorithms possess objectivity solely because of their mathematical or computational characteristics, although they do not.

The project has prompted significant discourse regarding gender recognition systems, which is equally as critical as the topic of race. We have consistently utilized the term gender, as it is the official designation in both the Labeled Faces in the Wild dataset and the NIST databases. The datasets classify people according to their biological sex (male/female) yet label this classification as gender. This prompts the question of the feasibility of determining an individual's gender or biological sex based solely on a photograph of their facial features. Ultimately, we concluded that it is not feasible to definitively establish an individual's gender; rather, the system is identifying the individual's gender expression. This underscores the issue surrounding the socially constructed nature of gender and how labeling an image as either man or woman reinforces the binary framework of gender, which is increasingly being challenged. This presents a potential danger to people who do not adhere to this binary structure. (Hamidi et al., 2018) contends that the binary perspective on gender overlooks non-binary people who do not fit within the current classification system. Furthermore, relying solely on visual indicators to determine gender suggests a limited perspective that ignores the existence of transgender people. As a result, the probability of misidentifying the gender of a transgender person is markedly increased, potentially leading to discriminatory treatment (Goel et al., 2018). (Hamidi et al., 2018) examines strategies for developing gender-inclusive systems. The interviewees suggest methods that allow users to enter or confirm their identity. Whether a human or a computer conducts the assessment, (Hamidi

et al., 2018) agrees with (Wei et al., 2020) and (*The Misgendering Machines*, 2022) that determining gender through physical characteristics is impossible. Consequently, the obligation of self-identification should be assigned to the user. The utilization of a gender classifier in this study functions as a method to quantitatively evaluate the probability of racial bias within the system. Nonetheless, this classifier, due to its existing constraints, does not account for the non-binary aspects of gender.

Chapter 5—Discussion and Conclusions

5.1 Discussion

This study presents evidence that biases, specifically toward African Americans aged 18 to 65, can be alleviated through the retraining of facial recognition algorithms. Assume that fundamental models (with minimal parameters) may precisely deduce sensitive characteristics from the learning process of a face recognition neural network. Consequently, these features will probably be understood. The ability of a linear model to accurately predict a test subject's gender from their image utilizing face embeddings and gender labels assesses the model's gender awareness. Moreover, our results demonstrate that the algorithms display differing inaccuracies when forecasting individuals from various racial backgrounds. Among White people, the models demonstrate a greater tendency to differentiate between two photos of the same person. The opposite was true for appearances classified as black. Again, blinding the model had minimal effect on this. Improving the quality of the training data seems to be the more effective method for mitigating bias in the frequency range.

The significant consequences of different error types become evident when examining a common application of facial recognition technology: law enforcement. False negatives in law enforcement denote the inability to recognize a suspect due to a disparity between the camera footage and the person's image in the database. False positives denote occurrences where camera footage of an innocent person is erroneously recognized as representing the subject of interest. The facial recognition algorithm overestimates the probability of identifying white people as potential suspects due to a greater incidence of false negatives than false positives. Conversely, when faces are

classified as black, the algorithm exhibits a greater propensity to misidentify an innocent person as a suspect rather than mistakenly overlooking an actual suspect. The disparity between the privileges afforded to white people and the discrimination faced by members of racial minority groups renders this problem particularly complex.

When people are obliged to engage in face matching, the presence of biases resulting in a decreased capacity to identify persons belonging to particular races may not be attributed to prejudices but rather to restricted exposure to these faces. Our tests reveal that the FR models most likely follow the same rule: awareness of the racial or gender composition of the groups does not affect the variation in their performance levels. Rather, it results from their past limited exposure to persons from that specific demographic group. Therefore, we urge machine learning experts to rely on their sense of fairness to FR instead of trying to solve the issue of bias in fairness reporting. Ignoring subtle features does not seem in itself to be a sufficient medium for objective and fair description. FR developers who find biases in the performance rates of their models should improve the training data of their model as advised in past research instead of trying to blind their model.

5.2 Conclusions

Our research is coupled with specific constraints. The primary constraint is the data utilized for our investigations. We endeavored to use a dataset that was well-balanced about racial and gender demographics. We selected the NIST FRVT dataset and the LFW dataset, as the NIST FRVT dataset is predominantly utilized by law enforcement in facial recognition technology. Compared to many other publicly evaluated datasets, the LFW dataset contains a large number of populations. Yet, the

dataset lacks labels for additional variables, including gender. Consequently, we had to label the dataset independently, utilizing “a pre-trained classifier. We discovered that faces identified as female, particularly those also categorized as African American or black, are underrepresented in the LFW dataset. By curating the training data for models that law enforcement will use in face recognition technology against African Americans between the ages of 18 and 65, we were able to successfully demonstrate the value of minimizing algorithmic biases in this work. If face recognition algorithms are not taught with a diverse set of examples, our findings suggest that they will struggle to correctly classify images belonging to underrepresented groups. Grayscale images can help a little with this issue, but it won't solve the problem entirely.

The results show that there are risks that could be involved with developing any kind of face classification system. Though gender was used as a performance metric in this study, it doesn't mean we can build a system that can correctly identify a person's gender from just an input picture. Facial classification is an inadequate method for determining an individual's gender identity, as visual assessment alone cannot reliably convey this information. We painstakingly reviewed every picture to correctly classify our models based on their orientation, including whether they were looking forward, to the left, or right, and whether or not they were wearing eyeglasses. The modest measure of integrating human monitoring, although time-consuming and sometimes missed, proved helpful in reducing much of the bias in our facial recognition algorithms.

5.3 Contributions to Body of Knowledge

This study demonstrates the impact of curating algorithmic training data and the outcomes of employing a predictive model rather than depending exclusively on human analysis. The primary contributions to this research are listed below:

1. **Limited or Obsolete Technology and Algorithms.** We demonstrate that taking a more modern approach can solve problems caused by limited or outdated algorithms and techniques by building predictive models for retraining data algorithms. The complexity and diversity of human facial features are often too much for those old or simplistic algorithms to manage. The fact that this limitation often leads to errors, particularly among African American populations, highlights the need for constant updates and improvements in algorithmic design.
2. **Analytics and Design Subjectivity.** Through screening, we were able to demonstrate that our system achieved true accuracy after retraining using a larger number of diverse faces representative of the African American population. This corrects for bias in facial recognition systems, which may unconsciously absorb the biases of those who create and direct them.
3. **Reinforce equality through better data labeling.** We incorporated diverse perspectives when labeling the data. We were acutely aware of any potential bias and specifically sought to populate our sample with the largest, most diverse sample possible. This further mitigates any bias in our algorithmic model.

5.4 Recommendations for Future Research

This research observation highlights significant issues in current FR solutions, particularly regarding labeling, robustness, fairness, and explainability. Here are some potential topics worth exploring for future research:

-Improving the data utilized for training the FR models appears to be the most straightforward and least problematic approach to enhancing error rates. Within this research, this was the focus, curating the training data algorithm. Future research could deep dive into “training the trainer”. And what method to employ to regulate data augmentation.

-Future research should examine the contexts in which FRT may be inappropriate and explore strategies to prevent its misuse, such as implementing relevant policies. Advocating for reduced bias in FR will benefit those who are excluded only after these issues have been resolved.

-Engaging a third party to review your data labeling policies and confirm regulatory compliance offers a chance for an objective examination ML operations. This methodology was used within this study to ensure we mitigated bias in the algorithm.

-External controls can promote accountability, openness, and fairness by carefully examining current labeling procedures and seeking improvements.

-Regulating FRT. Future research could substantiate the need for governmental oversight based on the number of false arrests against African Americans by law enforcement.

References

51Testing. (2021, March 16). A Complete Guide of Facial Recognition Test. *Medium*. https://51testing.medium.com/a-complete-guide-of-facial-recognition-test-45f21ce4b582

admin. (2017, May 30). Facial recognition algorithms | Principle of work and analysis | NtechLab. *NtechLab*. https://ntechlab.com/blog/2017/05/30/notes-on-evaluating-face-recognition-software/

Agarwal, P. (2020). *Sway: Unravelling unconscious bias*. Bloomsbury Publishing.

Ahn, D., Almaatouq, A., Gulabani, M., & Hosanagar, K. (2021). *Will We Trust What We Don't Understand? Impact of Model Interpretability and Outcome Feedback on Trust in AI* (arXiv:2111.08222). arXiv. https://doi.org/10.48550/arXiv.2111.08222

Albiero, V., Zhang, K., & Bowyer, K. W. (2020). How Does Gender Balance In Training Data Affect Face Recognition Accuracy? *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. https://doi.org/10.1109/IJCB48548.2020.9304924

Allyn, B. (2020, June 24). "The Computer Got It Wrong": How Facial Recognition Led To False Arrest Of Black Man. *NPR*. https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig

Arneson, R. J. (1999). *What, If Anything, Renders All Humans Morally Equal?* (R. J. Arneson, Ed.; pp. 103–128). Blackwell. https://philarchive.org/rec/ARNWIA

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., & Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, *10*(3), 137. https://doi.org/10.1504/IJSN.2015.071829

B, A., H, C., P, L., & O, L. (2024). Impact of Explainable AI on Reduction of Algorithm Bias in Facial Recognition Technologies. *2024 Systems and Information Engineering Design Symposium (SIEDS)*, 85–89. https://doi.org/10.1109/SIEDS61124.2024.10534745

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback* (arXiv:2212.08073). arXiv. http://arxiv.org/abs/2212.08073

Beauxis-Aussalet, E., Behrisch, M., Borgo, R., Chau, D. H., Collins, C., Ebert, D., El-Assady, M., Endert, A., Keim, D. A., Kohlhammer, J., Oelke, D., Peltonen, J., Riveiro, M., Schreck, T., Strobelt, H., van Wijk, J. J., & Rhyne, T.-M. (2021). The Role of Interactive Visualization in Fostering Trust in AI. *IEEE Computer Graphics and Applications*, *41*(6), 7–12. https://doi.org/10.1109/MCG.2021.3107875

Bhatta, A., Mery, D., Wu, H., Annan, J., King, M. C., & Bowyer, K. W. (2024). *What's color got to do with it? Face recognition in grayscale* (arXiv:2309.05180). arXiv. https://doi.org/10.48550/arXiv.2309.05180

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Brey, P. (2004). Ethical aspects of facial recognition systems in public places. *Journal of Information, Communication and Ethics in Society*, *2*(2), 97–109. https://doi.org/10.1108/14779960480000246

Brodkin, J. (2023, January 4). *Black man wrongfully jailed for a week after face recognition error, report says | Ars Technica*. ARS Technica. https://arstechnica.com/tech-policy/2023/01/facial-recognition-error-led-to-wrongful-arrest-of-black-man-report-says/

Bump, P. (2018, November 29). *Facial Recognition for Law Enforcement—6 Current Applications*. Emerj Artificial Intelligence Research. https://emerj.com/ai-sector-overviews/facial-recognition-in-law-enforcement/

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Cagle, M. G., Matt. (2024, February 7). When it Comes to Facial Recognition, There is No Such Thing as a Magic Number | ACLU. *American Civil Liberties Union*.

https://www.aclu.org/news/privacy-technology/when-it-comes-to-facial-recognition-there-is-no-such-thing-as-a-magic-number

Chellappa, R., Sirohey, S., & Wilson, C. L. (1995). Human and Machine Recognition of Faces: A Survey. *IEEE*, *83*(5), 705–740. https://doi.org/DOI: 10.1109/5.381842

Chi, N., Lurie, E., & Mulligan, D. K. (2021). Reconfiguring Diversity and Inclusion for AI Ethics. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 447–457. https://doi.org/10.1145/3461702.3462622

Chicklas, D. (2021, April 13). *Farmington Hills Father Sues Detroit Police Department for Wrongful Arrest Based on Faulty Facial Recognition Technology | ACLU of Michigan*. ACLU Michigan. https://www.aclumich.org/en/press-releases/farmington-hills-father-sues-detroit-police-department-wrongful-arrest-based-faulty

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*. https://doi.org/10.1016/j.chb.2021.107018

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313. https://doi.org/10.1038/538311a

Darabant, A. S., Borza, D., & Danescu, R. (2021). Recognizing Human Races through Machine Learning—A Multi-Network, Multi-Features Study. *Mathematics*, *9*(2), Article 2. https://doi.org/10.3390/math9020195

Day, O., & Khoshgoftaar, T. M. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*, *4*(1), 29. https://doi.org/10.1186/s40537-017-0089-0

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., & Chellappa, R. (2021). PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15067–15076. https://doi.org/10.1109/ICCV48922.2021.01481

Dooley, S., Sukthanker, R. S., Dickerson, J. P., White, C., Hutter, F., & Goldblum, M. (2023). *Rethinking Bias Mitigation: Fairer Architectures Make for Fairer Face Recognition* (arXiv:2210.09943). arXiv. https://doi.org/10.48550/arXiv.2210.09943

Duenser, A., & Douglas, D. M. (2023). Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intelligent Systems*, *38*(6), 19–26. https://doi.org/10.1109/MIS.2023.3322586

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist–level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

FaceNet—Using Facial Recognition System. (2020, May 2). GeeksforGeeks. https://www.geeksforgeeks.org/facenet-using-facial-recognition-system/

Facial recognition gives police a powerful new tracking tool. It's also raising alarms. (2018, July 30). NBC News. https://www.nbcnews.com/news/us-news/facial-recognition-gives-police-powerful-new-tracking-tool-it-s-n894936

Ferrara, E. (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Fu, R., Huang, Y., & Singh, P. V. (2020). *AI and Algorithmic Bias: Source, Detection, Mitigation and Implications* (SSRN Scholarly Paper 3681517). https://doi.org/10.2139/ssrn.3681517

Fung, B. (2024, January 23). *Lawsuit: Facial recognition software leads to wrongful arrest of Texas man; he was in Sacramento at time of robbery - CBS Sacramento*. https://www.cbsnews.com/sacramento/news/texas-macys-sunglass-hut-facial-recognition-software-wrongful-arrest-sacramento-alibi/

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness Testing: Testing Software for Discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510. https://doi.org/10.1145/3106237.3106277

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Garvie, C., & Frankle, J. (2016, April 7). Facial-Recognition Software Might Have a Racial Bias Problem. *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/

Gentzel, M. (2021). Biased Face Recognition Technology Used by Government: A Problem for Liberal Democracy. *Philosophy & Technology*, *34*(4), 1639–1663. https://doi.org/10.1007/s13347-021-00478-z

Goel, N., Yaghini, M., & Faltings, B. (2018). Non-Discriminatory Machine Learning Through Convex Fairness Criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.11662

Gong, S., Liu, X., & Jain, A. K. (2020). *Jointly De-biasing Face Recognition and Demographic Attribute Estimation* (arXiv:1911.08080). arXiv. https://doi.org/10.48550/arXiv.1911.08080

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Grother, P. J., Ngan, M. L., & Hanaoka, K. K. (2019). Face Recognition Vendor Test Part 3: Demographic Effects. *NIST*. https://www.nist.gov/publications/face-recognition-vendor-test-part-3-demographic-effects

Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., & Morris, M. R. (2019). *Toward Fairness in AI for People with Disabilities: A Research Roadmap* (arXiv:1907.02227). arXiv. https://doi.org/10.48550/arXiv.1907.02227

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173582

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning* (arXiv:1610.02413). arXiv. https://doi.org/10.48550/arXiv.1610.02413

Hong, J.-C., Li, Y., Kuo, S.-Y., & An, X. (2022). Supporting schools to use face recognition systems: A continuance intention perspective of elementary school parents in China. *Education and Information Technologies*, *27*(9), 12645–12665. https://doi.org/10.1007/s10639-022-11084-7

Hupont Torres, I., & Fernández, C. (2019). *DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition*. 1–7. https://doi.org/10.1109/FG.2019.8756625

InsightFace. (n.d.). Retrieved October 13, 2024, from https://insightface.ai/arcface

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Jefferson, B. (2020). *Digitize and Punish: Racial Criminalization in the Digital Age*. University of Minnesota Press. https://doi.org/10.5749/j.ctvz0h9s7

Jiang, H., & Nachum, O. (2019). *Identifying and Correcting Label Bias in Machine Learning* (arXiv:1901.04966). arXiv. https://doi.org/10.48550/arXiv.1901.04966

Joachims, T., Swaminathan, A., & Schnabel, T. (2016). *Unbiased Learning-to-Rank with Biased Feedback* (arXiv:1608.04468). arXiv. https://doi.org/10.48550/arXiv.1608.04468

Johnson, T. L., Johnson, N. N., McCurdy, D., & Olajide, M. S. (2022). Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly*, *39*(4), 101753. https://doi.org/10.1016/j.giq.2022.101753

Kabir, H. M. D., Khosravi, A., Hosen, M. A., & Nahavandi, S. (2018). Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications. *IEEE Access*, *PP*. https://doi.org/10.1109/access.2018.2836917

Klare, B., Burge, M., Klontz, J., Vorder Bruegge, R., & Jain, A. (2012). Face Recognition Performance: Role of Demographic Information. *Information Forensics and Security, IEEE Transactions On*, *7*, 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Knox, E. (2018, August 8). Why confidence matters in facial recognition systems. *NYU Center for Cyber Security*. https://cyber.nyu.edu/2018/08/08/confidence-matters-facial-recognition-systems/

Kolla, M., & Savadamuthu, A. (2023). The Impact of Racial Distribution in Training Data on Face Recognition Bias: A Closer Look. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 313–322. https://doi.org/10.1109/WACVW58289.2023.00035

Krishnan, A., Almadan, A., & Rattani, A. (2020). Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1028–1035. https://doi.org/10.1109/ICMLA51294.2020.00167

Kuhlman, C., Jackson, L., & Chunara, R. (2020). *No computation without representation: Avoiding data and algorithm biases through diversity* (arXiv:2002.11836). arXiv. https://doi.org/10.48550/arXiv.2002.11836

Leavy, S., O'Sullivan, B., & Siapera, E. (2020). *Data, Power and Bias in Artificial Intelligence* (arXiv:2008.07341). arXiv. https://doi.org/10.48550/arXiv.2008.07341

Lewis, J. A., & Crumpler, W. (2021). *Questions about Facial Recognition*. https://www.csis.org/analysis/questions-about-facial-recognition

Limante, A. (n.d.). *Bias in Facial Recognition Technologies Used by Law Enforcement: Understanding the Causes and Searching for a Way Out*. https://doi.org/10.1080/18918131.2023.2277581

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2018). *SphereFace: Deep Hypersphere Embedding for Face Recognition* (arXiv:1704.08063). arXiv. https://doi.org/10.48550/arXiv.1704.08063

Lora, M. (n.d.). *A university professor wants to expose the hidden bias in AI, and then use it for good*. Retrieved March 6, 2024, from https://techxplore.com/news/2022-09-university-professor-expose-hidden-bias.html

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions*

on Management Information Systems, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Meenakshi, M. (2013). Real-Time Facial Recognition System—Design, Implementation and Validation. *Journal of Signal Processing Theory and Applications*. https://doi.org/10.7726/jspta.2013.1001

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). *Diversity in Faces* (arXiv:1901.10436). arXiv. https://doi.org/10.48550/arXiv.1901.10436

Morales, A., Fierrez, J., Vera-Rodriguez, R., & Tolosana, R. (2021). SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(6), 2158–2164. https://doi.org/10.1109/TPAMI.2020.3015420

Nyoka, A. F., Sauli Godfrey, K., Mwakilembe, H. G., Jonas Mugendi, L., Mbita, O. D., Assoumana, A., & Msongaleli, D. L. (2023). Reliable Face Identification System for Criminal Investigation. *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, 1–5. https://doi.org/10.1109/ISDFS58141.2023.10131738

Office, U. S. G. A. (n.d.-a). *Facial Recognition Technology: Federal Law Enforcement Agency Efforts Related to Civil Rights and Training | U.S. GAO*. Retrieved July 29, 2024, from https://www.gao.gov/products/gao-24-107372

Office, U. S. G. A. (n.d.-b). *Facial Recognition Technology: Federal Law Enforcement Agency Efforts Related to Civil Rights and Training | U.S. GAO*. Retrieved October 13, 2024, from https://www.gao.gov/products/gao-24-107372

Parker, D. (2023, December 12). UMES granted $327K to study facial recognition bias. *47abc*. https://www.wmdt.com/2023/12/umes-granted-327k-to-study-facial-recognition-bias/

Parra, C. M., Gupta, M., & Dennehy, D. (2022). Likelihood of Questioning AI-Based Recommendations Due to Perceived Racial/Gender Bias. *IEEE Transactions on Technology and Society*, *3*(1), 41–45. https://doi.org/10.1109/TTS.2021.3120303

PatrickFarley. (2024, August 28). *What is the Azure AI Face service? - Azure AI services*. https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-identity

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., & Nam, C. S. (2023). Ethics in human–AI teaming: Principles and perspectives. *AI and Ethics*, *3*(3), 917–935. https://doi.org/10.1007/s43681-022-00214-z

Radiya-Dixit, E., & Neff, G. (2023). A Sociotechnical Audit: Assessing Police Use of Facial Recognition. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346. https://doi.org/10.1145/3593013.3594084

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, January 3). *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing*. arXiv.Org. https://arxiv.org/abs/2001.00964v1

Raposo, V. L. (2024). When facial recognition does not 'recognise': Erroneous identifications and resulting liabilities. *AI & SOCIETY*, *39*(4), 1857–1869. https://doi.org/10.1007/s00146-023-01634-z

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, *3*(3), 735–744. https://doi.org/10.1007/s43681-022-00200-5

Roach, J. (2018, June 26). *Microsoft improves facial recognition to perform well across all skin tones*. The AI Blog. https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., & Timoner, S. (2020). Face Recognition: Too Bias, or Not Too Bias? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1–10. https://doi.org/10.1109/CVPRW50498.2020.00008

Roomi, S. Md. M., Virasundarii, S. L., Selvamegala, S., Jeevanandham, S., & Hariharasudhan, D. (2011). Race Classification Based on Facial Features. *2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 54–57. https://doi.org/10.1109/NCVPRIPG.2011.19

Roy, M. (2017). Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy . New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811). *College & Research Libraries*, *78*, 403–404. https://doi.org/10.5860/crl.78.3.403

Schuetz, P. (2021). Fly in the Face of Bias: Algorithmic Bias in Law Enforcement's Facial Recognition Technology and the Need for an Adaptive Legal Framework. *Minnesota Journal of Law & Inequality*, *39*(1). https://doi.org/10.24926/25730037.391

Shaker, Z. (2022, November 14). *How Facial Recognition Investigative Training Can Help Avoid Wrongful Arrests*. Clearview AI. https://www.clearview.ai/post/how-facial-recognition-investigative-training-can-help-avoid-wrongful-arrests

Shi, G. (2023). Technical Analysis of Machine Learning Algorithms in Artificial Intelligence Image Recognition. *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, 1–5. https://doi.org/10.1109/ASIANCON58793.2023.10270097

Silva, Y., Rich, C., & Hall, D. (2016). BullyBlocker: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 1377–1379. https://doi.org/10.1109/ASONAM.2016.7752420

Ståhl, N., Falkman, G., Karlsson, A., & Mathiason, G. (2020). Evaluation of Uncertainty Quantification in Deep Learning. *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, *1237*, 556–568. https://doi.org/10.1007/978-3-030-50146-4_41

Stark, L. (2019). Facial recognition is the plutonium of AI. *XRDS*, *25*(3), 50–55. https://doi.org/10.1145/3313129

Stoica, A.-A., Riederer, C., & Chaintreau, A. (2018). Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. *Proceedings of the 2018 World Wide Web Conference*, 923–932. https://doi.org/10.1145/3178876.3186140

Terhörst, P., Fährmann, D., Kolf, J. N., Damer, N., Kirchbuchner, F., & Kuijper, A. (2021). *MAAD-Face: A Massively Annotated Attribute Dataset for Face Images* (arXiv:2012.01030). arXiv. https://doi.org/10.48550/arXiv.2012.01030

Terhorst, P., Kolf, J. N., Huber, M., Kirchbuchner, F., Damer, N., Moreno, A. M., Fierrez, J., & Kuijper, A. (2022). A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Transactions on Technology and Society*, *3*(1), 16–30. https://doi.org/10.1109/TTS.2021.3111823

The Facts on Facial Recognition with Artificial Intelligence. (n.d.). Amazon Web Services, Inc. Retrieved October 13, 2024, from https://aws.amazon.com/rekognition/the-facts-on-facial-recognition-with-artificial-intelligence/

The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition | Trans Reads. (2022, June 7). https://transreads.org/the-misgendering-machines-trans-hci-implications-of-automatic-gender-recognition/

Tolba, A., El-Baz, A., & El-Harby, A. (2005). Face Recognition: A Literature Review. *International Journal of Signal Processing*, *2*, 88–103.

Tracy, P., & Morgan, V. (2000). Big Brother and His Science Kit: DNA Databases for 21st Century Crime Control. *Journal of Criminal Law and Criminology*, *90*(2), 635.

Ughade, N. (2023, December 15). *How To Mitigate Facial Recognition Bias in Identity Verification*. Hyperverge.Co. https://hyperverge.co/blog/mitigating-facial-recognition-bias/

Vincent, J. (2018, January 12). *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*. The Verge. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, *2*(3), 509–522. https://doi.org/10.1007/s43681-021-00108-6

Wei, D., Ramamurthy, K. N., & Calmon, F. (2020). Optimized Score Transformation for Fair Classification. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 1673–1683. https://proceedings.mlr.press/v108/wei20a.html

Weiss, K. R., & Khoshkoftaar, T. M. (2018). A Study of the Impact of Base Traditional Learners on Transfer Learning Algorithms. *International Journal on Artificial Intelligence Tools*, *27*(06), 1850022. https://doi.org/10.1142/S0218213018500227

Wu, H., Albiero, V., Krishnapriya, K. S., King, M. C., & Bowyer, K. W. (2023). Face Recognition Accuracy Across Demographics: Shining a Light Into the Problem. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1041–1050. https://doi.org/10.1109/CVPRW59228.2023.00111

Wu, H., & Bowyer, K. (2023, April 17). *What Should Be Balanced in a "Balanced" Face Recognition Dataset?* British Machine Vision Conference. https://www.semanticscholar.org/paper/What-Should-Be-Balanced-in-a-

%22Balanced%22-Face-Wu-Bowyer/e03191bce4de9f0df453862b7178cd499ff0954f

Yarger, L., Cobb Payton, F., & Neupane, B. (2019). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review*, *44*(2), 383–395. https://doi.org/10.1108/OIR-10-2018-0334

Young, M., Magassa, L., & Friedman, B. (2019). Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, *21*. https://doi.org/10.1007/s10676-019-09497-z

Zeng, Y., Lu, E., Sun, Y., & Tian, R. (2019). *Responsible Facial Recognition and Beyond* (arXiv:1909.12935). arXiv. https://doi.org/10.48550/arXiv.1909.12935

Appendix A

Figure A-1: ML Notations

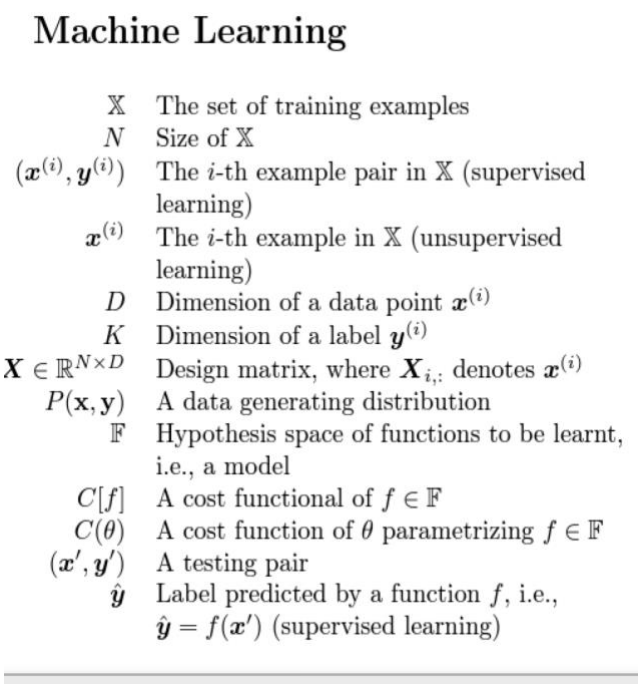
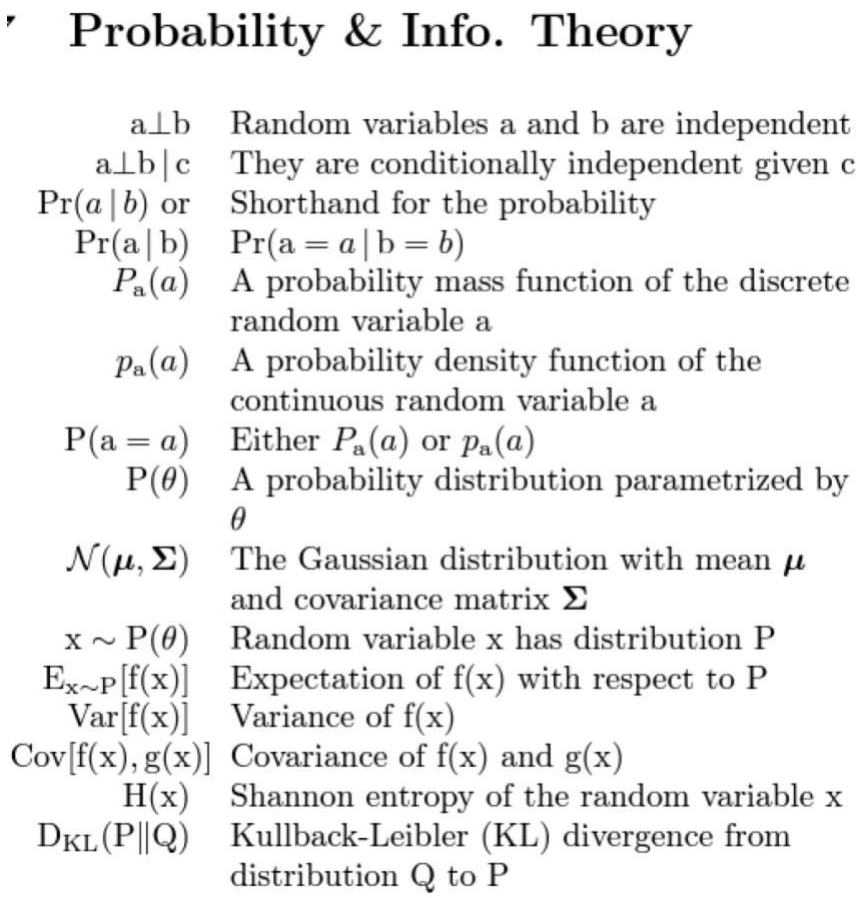


Figure A-2: Probability Notations

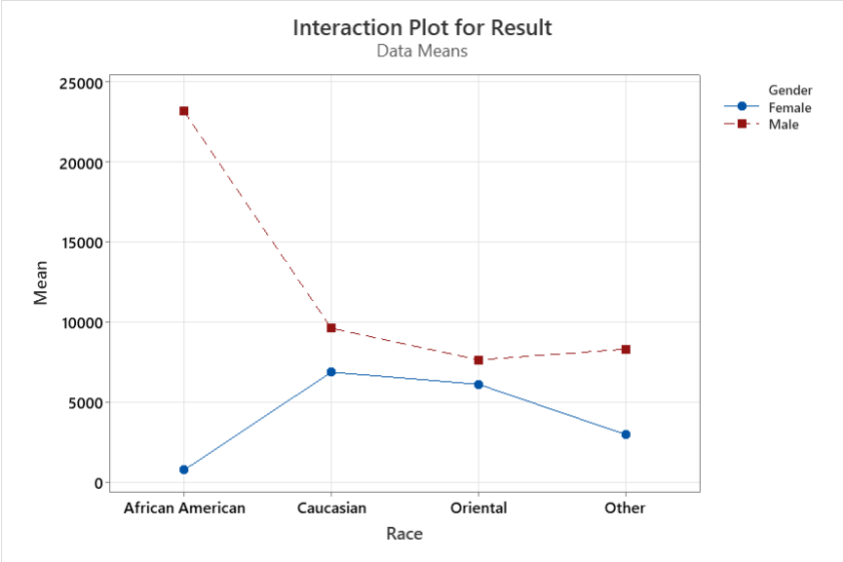


Probability & Info. Theory

$\mathrm{a} \perp \mathrm{b}$	Random variables a and b are independent
$\mathrm{a} \perp \mathrm{b} \mid \mathrm{c}$	They are conditionally independent given c
$\Pr(a \mid b)$ or	Shorthand for the probability
$\Pr(\mathrm{a} \mid \mathrm{b})$	$\Pr(\mathrm{a} = a \mid \mathrm{b} = b)$
$P_{\mathrm{a}}(a)$	A probability mass function of the discrete random variable a
$p_{\mathrm{a}}(a)$	A probability density function of the continuous random variable a
$\mathrm{P}(\mathrm{a} = a)$	Either $P_{\mathrm{a}}(a)$ or $p_{\mathrm{a}}(a)$
$\mathrm{P}(\theta)$	A probability distribution parametrized by θ
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{x} \sim \mathrm{P}(\theta)$	Random variable x has distribution P
$\mathrm{E}_{\mathrm{x} \sim \mathrm{P}}[\mathrm{f}(\mathrm{x})]$	Expectation of f(x) with respect to P
$\mathrm{Var}[\mathrm{f}(\mathrm{x})]$	Variance of f(x)
$\mathrm{Cov}[\mathrm{f}(\mathrm{x}), \mathrm{g}(\mathrm{x})]$	Covariance of f(x) and g(x)
$\mathrm{H}(\mathrm{x})$	Shannon entropy of the random variable x
$\mathrm{D}_{\mathrm{KL}}(\mathrm{P} \Vert \mathrm{Q})$	Kullback-Leibler (KL) divergence from distribution Q to P

Appendix B

Figure B-1 Results



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Gender	1	128640800	128640800	*	*
Race	3	45081602	15027201	*	*
Gender*Race	3	142673490	47557830	*	*
Error	0	*	*		
Total	7	316395892			

Factor Information

Factor	Type	Levels	Values
Gender	Fixed	2	Female, Male
Race	Fixed	4	African American, Caucasian, Oriental, Other

Regression Equation

P(Male) = exp(Y')/(1 + exp(Y'))

Race	
African American	Y' = -98.04 + 0.008190 Result
Caucasian	Y' = -67.54 + 0.008190 Result
Oriental	Y' = -56.31 + 0.008190 Result
Other	Y' = -46.18 + 0.008190 Result

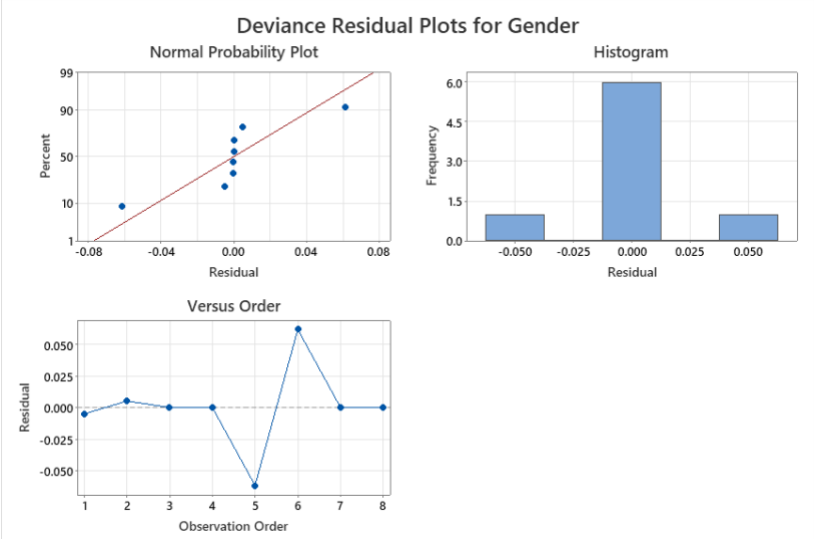
Regression Equation

P(Female) = exp(Y')/(1 + exp(Y'))

Race	
African American	Y' = 98.04 - 0.008190 Result
Caucasian	Y' = 67.54 - 0.008190 Result
Oriental	Y' = 56.31 - 0.008190 Result
Other	Y' = 46.18 - 0.008190 Result

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Result	0.9918	(0.9674, 1.0169)



Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve
99.93%	63.86%	10.01	40.01	10.40	1.0000

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	3	0.01	1.000
Pearson	3	0.00	1.000
Hosmer-Lemeshow	6	0.00	1.000

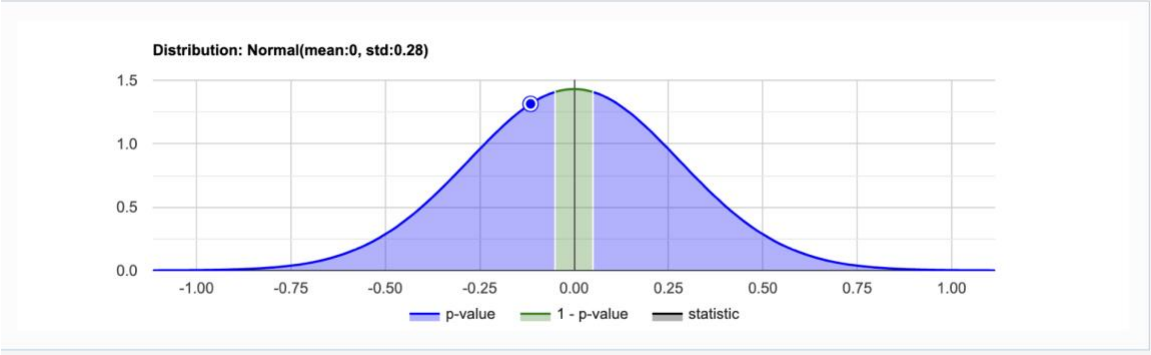
Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	4	0.41	0.981
Result	1	0.41	0.520
Race	3	0.02	0.999

Predicted \ Actual	Positive	Negative
Positive	23943 (TP)	38 (FP)
Negative	10 (FN)	61 (TN)

In this matrix, we observe 23943 True Positives, 61 True Negatives, 38 False Positives, and 10 False Negatives. These figures enable us to calculate various performance metrics:

- Accuracy: 99.80%
- Precision: 99.84%
- F1-Score: 99.90%



$$S.E = \frac{\sigma}{\sqrt{n}} \qquad Z = \frac{MOE}{S.E}$$

$$S.E = \frac{9.25}{\sqrt{1102}} = 0.2786 \qquad Z = \frac{0.05}{0.2786} = 0.1794$$

P = P(x ≤0.1794) = 0.5712

- Data is normally distributed