Fair AI

A thesis submitted in partial fulfillment of the requirements
for the award of the degree of

Bachelor of Technology

with specialization in

Information Technology

Submitted by

Keshav Lohani	IIT2022012
Goransh Barde	IIT2022027
Darshan Vanjara	IIT2022037
Trilok Meena	IIT2022038
Aman Raj	IIT2022050

Under the Supervision of

Dr. K.P Singh

to the

Department of Information Technology

भारतीय सूचना प्रौद्योगिकी संस्थान, इलाहाबाद

Indian Institute of Information Technology,
Allahabad

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद

Indian Institute of Information Technology Allahabad

An Institute of National Importance by Act of Parliament

Deoghat Jhalwa, Prayagraj 211015 (U.P.) India

Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CERTIFICATE

It is certified that the work contained in the thesis titled "Fair AI" has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

Dr. K.P Singh
Department of Information Technology
IIIT Allahabad

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CANDIDATE DECLARATION

I, Keshav Lohani, Roll no. IIT2022012, Goransh Barde roll no. IIT2022027, Darshan Vanjara Roll no. IIT2022037, Trilok Meena Roll no. IIT2022038, Aman Raj Roll no. IIT2022050, certify that this thesis work titled "Fair AI" is submitted by me towards partial fulfillment of the requirement of the Degree of Bachelor of Technology in the Department of Information Technology, Indian Institute of Information Technology, Allahabad.

I understand that plagiarism includes:

1. Reproducing someone else's work (fully or partially) or ideas and claiming it as one's own.
2. Reproducing someone else's work (verbatim copying or paraphrasing) without crediting.
3. Committing literary theft (copying some unique literary construct).

I have given due credit to the original authors/sources through proper citation for all the words, ideas, diagrams, graphics, computer programs, experiments, results, and websites that are not my original contributions. I have used quotation marks to identify verbatim sentences and given due credit to the original authors/sources.

I affirm that no portion of my work is plagiarized. In the event of a complaint of plagiarism, I shall be fully responsible. understand that my supervisor may not be in a position to verify that this work is not plagiarized.

Date: __________

Keshav Lohani(IIT2022012)
Goransh Barde(IIT2022027)
Darshan Vanjara (IIT2022037)
Trilok Meena (IIT2022038)
Aman Raj(IIT2022050)
Department of Information Technology
IIIT Allahabad
Prayagraj - 211015, U.P.

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I also got a good job in industry."

Keshav Lohani(IIT2022012)
Goransh Barde(IIT2022027)
Darshan Vanjara (IIT2022037)
Trilok Meena (IIT2022038)
Aman Raj(IIT2022050)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

Fair AI

ABSTRACT

Bachelor of Technology
Department of Information Technology

by Keshav Lohani(IIT2022012)
Goransh Barde(IIT2022027)
Darshan Vanjara (IIT2022037)
Trilok Meena (IIT2022038)
Aman Raj(IIT2022050)

Artificial Intelligence (AI) models are increasingly being deployed in various real-world applications, making it crucial to ensure their fairness and mitigate biases. AI models trained on biased datasets can perpetuate discrimination, leading to unfair and unethical decisions. This research aims to estimate bias in AI models, starting with simple architectures like LeNet-5 and extending the study to complex models and facial recognition datasets. We evaluate bias using different datasets, analyze performance disparities, and apply fairness evaluation metrics to quantify bias. The study provides insights into bias estimation and suggests improvements for fair AI model development.

Acknowledgements

We would like to express our heartfelt gratitude to **Dr. K.P Singh** for his invaluable guidance, constant encouragement, and insightful advice throughout the course of our mini project titled "Fair AI" His expertise and suggestions were pivotal in helping us navigate the challenges we encountered, and we deeply appreciate his support.

As 6th semester B.Tech students, we found the process both challenging and rewarding. The project allowed us to enhance our knowledge in Artificial Inteliigence and its applications , a field we had not previously explored. We thoroughly enjoyed working on this project and appreciate the learning experience it provided.

Once again, we thanks to Dr. K.P Singh for making this project a success.

Contents

Chapter 1

Introduction

1.1 Background

Artificial Intelligence (AI) is revolutionizing critical sectors such as healthcare, finance, hiring, and law enforcement. As AI technologies advance, they offer promising benefits such as increased efficiency and accuracy in decision-making. However, these systems are not without significant concerns, especially in terms of fairness and bias. AI systems often inherit and amplify biases present in the training data, leading to unfair or discriminatory outcomes. This phenomenon poses a serious ethical challenge as AI models are deployed in areas that can impact people's lives in profound ways.

1.2 Problem Statement

Bias in AI is a critical issue, particularly as AI systems are increasingly integrated into decision-making processes that affect peoples careers, health, financial stability, and legal outcomes. These biases can arise from various sources, including unbalanced datasets, flawed data labeling, and biased architectural assumptions. The consequences of biased AI can be severe, often disproportionately affecting marginalized groups and reinforcing existing societal inequalities. Thus, addressing fairness and mitigating bias is essential to ensure the ethical deployment of AI systems in high-stakes domains.

1.3 Core Concepts

This research focuses on two central concepts:

- **Fairness in AI**: The principle that AI models should treat all individuals equitably, without bias or discrimination based on sensitive attributes such as race, gender, age, or ethnicity. Fair AI models ensure that decisions or predictions are not influenced by irrelevant or discriminatory factors.

- **Bias in AI**: Systematic errors or prejudices that lead to unequal treatment or disadvantage to certain groups. These biases often stem from imbalanced or unrepresentative datasets, flawed labeling processes, or biased assumptions embedded within the algorithms themselves.

1.4 Origins of Bias in AI Models

Bias in AI can arise from several sources:

- **Unbalanced Datasets**: When datasets used to train AI models are not representative of the population or fail to include diverse demographic groups, the models may favor the majority group, leading to biased outcomes.
- **Flawed Labeling**: Incorrect or biased data labeling can introduce unintended prejudice into the models predictions.
- **Biased Model Architecture**: Assumptions or design choices in the model architecture may inadvertently reinforce biases present in the data.

Understanding these sources of bias is crucial for developing AI systems that are both accurate and fair.

1.5 Importance of Bias Estimation in AI Models

Accurately estimating and addressing bias in AI models is essential for several reasons:

- **Ethical Considerations**: Ensuring fairness in AI systems is critical to avoid discrimination against any group based on race, gender, age, or other protected attributes. Unchecked biases can raise significant ethical and legal concerns.
- **Regulatory Compliance**: With the introduction of laws and regulations such as the European Union's AI Act and the U.S. AI Bill of Rights, organizations must ensure that their AI systems are fair and compliant with these frameworks.
- **Improved Model Reliability**: Biased models may generalize poorly across diverse datasets, leading to inaccuracies and unreliable predictions. Identifying and addressing bias can improve the robustness and performance of AI models.
- **Public Trust and Acceptance**: Users are more likely to trust AI systems that demonstrate fairness and transparency. By addressing bias, AI systems can foster greater public confidence and acceptance.

1.6 Research Objectives

This study aims to:

1. Investigate bias in AI models through empirical evaluations on neural networks.
2. Train a baseline model on multiple datasets and analyze variations in bias using confusion matrices and fairness metrics.

3. Identify key factors contributing to bias in AI models and apply fairness metrics to quantify bias across different datasets.
4. Extend the study to more complex models and additional datasets, such as facial recognition datasets, to assess bias in broader AI applications.
5. Develop a framework that integrates multiple fairness evaluation metrics to enhance bias estimation and mitigation across AI models.

Chapter 2

Literature Review

The study of bias in Artificial Intelligence (AI) has become increasingly significant due to its ethical, societal, and technical implications. With the growing adoption of AI in domains such as healthcare, criminal justice, hiring, and financial services, the presence of bias in decision-making models raises critical concerns about fairness, accountability, and transparency. This chapter reviews existing literature on the **sources of bias**, **fairness evaluation metrics**, **bias mitigation techniques**, and **research gaps**, laying a foundation for our work on bias estimation and fairness evaluation.

2.1 Sources of Bias in AI

Numerous studies have categorized and analyzed the roots of AI bias, revealing that bias can manifest at multiple stages of model development and deployment. It is now widely accepted that bias is not solely a technical flaw but a reflection of deeper societal structures and choices made during system design.

- **Mehrabi et al. (2021)** identified three major types of bias:
  - **Representational Bias** refers to biases embedded in how groups are portrayed, such as reinforcing stereotypes in language models and image datasets (e.g., gender stereotypes in occupation-related images).
  - **Algorithmic Bias** arises from model design choices, including biased decision thresholds, inductive biases inherent in algorithms, or feature selection that inadvertently prioritizes majority group characteristics.
  - **Societal Bias** reflects real-world inequalities captured in the data, such as income disparities, racial profiling in policing data, or healthcare treatment biases.
- **Barocas et al. (2019)** further explained how historical and systemic inequalities become encoded into training datasets, leading models to inherit and even amplify existing social biases. They highlight how seemingly neutral data often embeds deep-rooted prejudices unless actively corrected.
- **Buolamwini & Gebru (2018)** in their seminal "**Gender Shades**" study demonstrated that commercial facial analysis algorithms from major tech

companies exhibited significant disparities in accuracy. They found error rates as high as 34.7% for darker-skinned women compared to 0.8% for lighter-skinned men, primarily due to the underrepresentation of certain demographic groups in the training data.

Recent research has expanded the understanding of bias sources further:

- **Suresh and Guttag (2021)** proposing a taxonomy covering historical, representation, measurement, aggregation, evaluation, and deployment biases.
- **Hooker (2021)** arguing that bias mitigation must be context-sensitive, warning against naive debiasing.
- **Wang et al. (2020)** highlighting amplification of bias in deeper layers of neural networks.
- **Geva et al. (2019)** showing how over-reliance on high-frequency examples marginalizes minority patterns.

Collectively, these findings emphasize that AI systems are susceptible to a wide range of biases arising not just from flawed data, but also from algorithmic choices, evaluation strategies, and deployment practices. Addressing bias, therefore, requires holistic interventions at multiple stages of the AI lifecycle.

2.2 Taxonomies and Frameworks for Understanding Bias

To systematically address bias in AI systems, several taxonomies and frameworks have been proposed. These structured approaches help in identifying, categorizing, and analyzing different types of biases at various stages of the data and model lifecycle. A strong understanding of these frameworks is essential for designing effective bias detection and mitigation strategies.

- **Suresh and Guttag (2021)** proposed a comprehensive **pipeline-level taxonomy** that categorizes sources of bias into six types: historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias. Their framework highlights that bias can be introduced at any stage from data collection to model deployment and stresses the need for interventions throughout the pipeline rather than solely at the data or algorithm level.
- **Olteanu et al. (2019)** focused on biases specifically arising in **social data** and introduced a classification system that identifies four major types:
  - **User Bias**: Bias introduced by the behavior, attributes, or preferences of users generating the data.
  - **Platform Bias**: Bias stemming from the way platforms (such as social media sites) design interactions, visibility algorithms, and data collection mechanisms.

  - **Sampling Bias**: Bias resulting from how data samples are selected, often leading to underrepresentation or overrepresentation of certain groups.
  - **Societal Bias**: Bias that reflects broader systemic inequalities embedded in societal structures.

- **Mitchell et al. (2019)** proposed the concept of **"Model Cards,"** a standardized reporting format aimed at promoting transparency and accountability in AI systems. Model Cards document important information about a models performance across different demographic groups, known limitations, intended uses, and ethical considerations. By including fairness evaluations and demographic breakdowns, Model Cards provide a structured way to communicate potential biases to users and stakeholders.

These taxonomies and frameworks offer critical lenses through which biases can be systematically uncovered and addressed. Rather than treating bias as an isolated defect, they encourage researchers and practitioners to recognize bias as a systemic issue requiring holistic solutions at multiple levels.

2.3 Methods for Measuring Bias and Fairness

The challenge of measuring fairness in AI systems has led to the development of various **fairness metrics**, which are designed to assess and quantify biases across different demographic groups. These metrics fall broadly into two categories: **group fairness** and **individual fairness**.

Group Fairness Metrics **Group fairness** metrics focus on ensuring that outcomes are balanced across demographic groups (e.g., based on gender, race, or other sensitive attributes). Several key metrics used to evaluate group fairness include:

- **Demographic Parity**:

  This metric requires that the **probability of a favorable outcome** (e.g., hiring, loan approval) should be independent of the sensitive attribute. In other words, a model should not systematically favor one group over another. For example, a job recruitment tool should not show bias towards one gender or ethnicity in the selection process (Dastin, 2018).

- **Equalized Odds** *(Hardt et al., 2016)*:

  Equalized Odds mandates that both the **false positive rate (FPR)** and the **false negative rate (FNR)** should be equal across groups. This ensures that the likelihood of making an error is consistent for all demographic groups. In contexts like predictive policing or credit scoring, this ensures no group is disproportionately misclassified, leading to fairer decision outcomes.

- **Equality of Opportunity** *(Hardt et al., 2016)*:

A more relaxed version of Equalized Odds, Equality of Opportunity focuses only on **equal true positive rates (TPR)** across groups. The core idea is that if individuals belong to a disadvantaged group, they should still have an equal chance of receiving a positive outcome if they meet the relevant criteria (e.g., creditworthiness or job qualifications).

- **Predictive Parity**:

  Predictive Parity focuses on **equal positive predictive values (precision)** across groups. This metric ensures that a positive prediction (e.g., loan approval or hiring decision) is equally accurate across groups. It is particularly relevant in domains where the stakes of predictions (e.g., financial decisions, medical diagnoses) are high.

- **Treatment Equality**:

  Treatment Equality ensures that the **ratios of false positives to false negatives** remain consistent across groups. This metric is critical when comparing how similarly different groups are treated in terms of both positive and negative predictions.

- **Distributional Measures** (e.g., Wasserstein Distance):

  While **Wasserstein Distance** is not a fairness metric in the strict sense, it is increasingly used as a **tool for comparing distributions** of predicted outcomes between groups. It measures the **distributional discrepancy** between groups' outcomes, providing insights into the disparity in predictions. Although not a direct fairness metric, it offers a quantitative measure of **fairness disparities**, which can be used alongside other metrics for a comprehensive fairness evaluation (Rudin, 2019).

Individual Fairness Metrics **Individual fairness** shifts the focus from group-level fairness to fairness at the level of individual predictions. The concept of individual fairness posits that **similar individuals** should be treated similarly, regardless of their group membership:

Individual Fairness Principle *(Dwork et al., 2012)*:

This foundational principle argues that "similar individuals should receive similar treatment" under a decision-making system. Defining "similarity" is challenging, particularly when considering complex, high-dimensional data, but the idea underpins much of the ongoing research in fairness.

Fairness Toolkits and Frameworks Several frameworks have emerged to help researchers and practitioners implement, evaluate, and mitigate fairness in AI systems. These tools integrate a variety of fairness metrics and bias mitigation strategies:

- **IBM AIF360 (AI Fairness 360)**:

  A comprehensive toolkit for fairness evaluation, AIF360 provides a collection of fairness metrics, bias mitigation algorithms, and data visualization tools. It allows users to analyze and mitigate biases across different stages of the AI pipeline, from data preprocessing to model evaluation (Bellamy et al., 2019).

- **Microsoft Fairlearn**:

  Fairlearn offers an interactive framework to assess model fairness, visualize fairness disparities, and implement fairness constraints during model training. It includes support for several fairness metrics, such as Equalized Odds, Demographic Parity, and other related measures (Mitchell et al., 2020).

These tools allow for the systematic evaluation of fairness across AI models, enabling researchers to analyze and mitigate biases in various domains. They also facilitate transparency in AI development, ensuring that fairness considerations are integrated into the decision-making process.

2.4 Bias Mitigation Techniques

Efforts to mitigate bias in AI can be broadly classified into three categories: **pre-processing**, **in-processing**, and **post-processing** approaches. Each category targets different stages of the AI model lifecycle and offers unique solutions for reducing bias.

a. Pre-processing Approaches Pre-processing techniques aim to modify the dataset before training to ensure fairer representation:

- **Kamiran & Calders (2012)**: They proposed techniques like **reweighting** and **resampling** to balance sensitive attributes, ensuring that underrepresented groups are better represented in the training data.
- **Zemel et al. (2013)**: Introduced the concept of **fair representation learning**, which transforms the input data into an intermediate space where sensitive attributes are obscured, thus reducing bias during training.

These techniques focus on altering the data distribution to achieve fairness before the model is trained, which is particularly useful when dealing with imbalanced datasets or when sensitive attributes are disproportionately represented.

b. In-processing Approaches In-processing methods involve adjusting the learning algorithm itself during training to encourage fairer predictions:

- **Zhang et al. (2018)**: Proposed **adversarial debiasing**, where a secondary network is introduced to penalize predictions that encode sensitive attributes, thereby forcing the model to learn less biased representations.
- Adding **fairness constraints** directly into the models loss function is another effective in-processing method. These constraints ensure that the model optimizes for both accuracy and fairness by incorporating fairness objectives into its training objective.

In-processing techniques focus on altering the training process itself, enabling the model to learn from biased data without perpetuating existing inequities.

c. Post-processing Approaches Post-processing techniques modify the outputs of the model after training, adjusting predictions to correct for bias:

- **Hardt et al. (2016)**: Introduced the **Reject Option Classification** method, which adjusts decisions near the decision boundary, ensuring that unfair outcomes are mitigated by altering classification thresholds.
- Emerging approaches such as **Causal Fairness** and **Counterfactual Fairness** focus on ensuring that decision outcomes are fair even in the presence of changing sensitive attributes. These methods use causal models and counterfactual reasoning to ensure fair outcomes across different scenarios.

Post-processing techniques are particularly useful when the model itself cannot be altered, but adjustments to the final predictions are still necessary to ensure fairness.

2.4 Challenges in Operationalizing Fairness

While the development of fairness metrics and mitigation techniques has made significant strides, several challenges remain in making fairness operational in AI systems:

1. Fairness-Accuracy Tradeoff There is often a tradeoff between fairness and accuracy. In many cases, improving fairness by balancing predictions across demographic groups may lead to a decrease in overall accuracy, particularly for groups that are underrepresented in the data.

2. Dynamic Bias Bias is not a static issue. As data distributions evolve over time, so too can the biases embedded in AI models. This **dynamic bias** presents a challenge in maintaining fairness over time, especially in real-world applications where data shifts are common.

3. Context-Dependence Fairness is highly context-dependent. Definitions of fairness can vary across domains, making it difficult to develop one-size-fits-all solutions. For instance, fairness in healthcare models may prioritize different outcomes than fairness in hiring models, requiring tailored fairness definitions for each use case.

4. Lack of Consensus Different fairness metrics can sometimes conflict with each other, leading to challenges in choosing the right metric for a given problem. This **lack of consensus** complicates the task of ensuring fairness across different applications, requiring careful balancing of fairness objectives.

5. Legal and Ethical Constraints Real-world AI applications must adhere to privacy laws, ethical standards, and industry regulations. These legal and ethical constraints can further complicate the operationalization of fairness, as they impose additional requirements beyond technical considerations.

Successfully addressing these challenges requires a balanced approach that incorporates not only technical and algorithmic solutions but also a strong understanding of societal, ethical, and legal factors.

2.5 Research Gaps and Future Directions

Despite significant advancements in bias mitigation and fairness research, there are still notable gaps that need to be addressed in future studies:

1. Standardized Fairness Pipelines Currently, there is a lack of **standardized fairness evaluation frameworks** that integrate various fairness metrics and mitigation techniques into a cohesive pipeline. The development of such frameworks would streamline the process of evaluating and ensuring fairness across different AI models.

2. Fairness in Complex Architectures Much of the research on fairness has focused on simpler models, but **deep learning models** like **CNNs** and **RNNs**, as well as **multimodal datasets** (e.g., combining vision and ECG data), have not been extensively studied in terms of fairness. Further research is needed to understand how bias manifests in these complex architectures and how fairness can be achieved in such models.

3. Bias in Generative Models Emerging models such as **diffusion models** and **large language models (LLMs)** present new challenges for bias and fairness. These models are capable of generating content that reflects societal biases, and there is a need for research into how to mitigate these biases in generative tasks.

4. Multi-objective Fairness Few studies focus on optimizing multiple fairness criteria simultaneously. Addressing **multi-objective fairness**where different fairness metrics must be balancedrepresents a promising research direction that could lead to more equitable AI systems across diverse applications.

5. Interpretability and Explainability Finally, as fairness becomes more integrated into AI systems, it is essential to develop models that are not only fair but also **interpretable** and **explainable** to stakeholders. Transparency in how decisions are made is crucial for fostering trust in AI systems, especially in high-stakes domains like healthcare and criminal justice.

Future research must address these gaps to create AI systems that are not only fair but also effective, interpretable, and robust in real-world settings.

2.6 Summary

This chapter provided an overview of the critical issues surrounding **bias** and **fairness** in AI. We explored the **sources of bias** in AI models, highlighting factors such as dataset imbalances, biased algorithms, and societal biases that can manifest in model outcomes. We also discussed a range of **fairness metrics**, including **group fairness** and **individual fairness**, which serve as tools to assess and quantify bias across different demographic groups.

Additionally, we examined various **bias mitigation techniques**, including **pre-processing**, **in-processing**, and **post-processing** methods, each addressing different stages of the AI model lifecycle. Despite the availability of these techniques, we also highlighted significant **challenges** in operationalizing fairness, including issues like the **fairness-accuracy tradeoff**, **dynamic biases**, and the **lack of consensus** on standardized fairness definitions.

Furthermore, we identified **research gaps** and areas for future exploration, such as the need for **standardized fairness pipelines**, **fairness in complex models** like deep learning, and methods for ensuring fairness in **generative models** and **multi-objective fairness**.

Our work aims to address these challenges by proposing more comprehensive **fairness evaluation frameworks**, extending fairness studies to **complex AI architectures**, and developing more **interpretable** and **effective** techniques for mitigating bias in real-world AI applications. By bridging these gaps, we hope to contribute to the development of **fairer, more transparent** AI systems that better align with ethical and societal standards.

Chapter 3

Methodology

In our research on **Fair AI: Estimation of Bias in AI Models**, we have systematically explored bias in AI systems by conducting empirical studies on a simple neural network model. Our investigation has been structured into multiple phases, including literature review, model selection, dataset experimentation, bias analysis, and fairness evaluation.

3.1 Understanding Bias and Fairness in Models

To begin our research, we first explored bias and fairness in AI models by manually observing their outputs. We prompted the model with different inputs to identify potential biases in the model's responses. This initial step helped us understand how biases may be reflected in model outputs, especially when exposed to diverse types of data.

3.2 Dataset Bias Evaluation

Next, we trained a simple Convolutional Neural Network (CNN) model on various datasets to observe the performance of the model across different types of data. This process aimed to identify dataset-related biases. By training the model on datasets with varying characteristics (e.g., imbalanced class distributions or skewed demographic representations), we observed how the model's performance fluctuated based on the dataset's composition. This helped us pinpoint the potential source of bias that may arise from the data itself.

3.3 Estimating Bias Using Fairness Metrics

Bias in AI refers to systematic errors or prejudices in a models predictions that disadvantage certain individuals or groups. These biases can originate from unrepresentative training data, flawed labeling processes, or design choices in the algorithm itself. When a dataset undersamples a demographic subgroup, the model tends to underperform for that group, leading to higher error rates. Human annotators may also introduce subjective biases during labeling, which then propagate into the learned model. Algorithmic assumptionssuch as optimizing solely for overall accuracycan amplify disparities by prioritizing majoritygroup performance. Evaluation practices that rely on aggregate metrics may conceal poor performance on minority groups, masking unfair behavior. In deployment, biased predictions can reinforce societal inequities, especially in highstakes domains like lending, hiring, or criminal justice. Addressing bias requires a multistage approach: preprocessing data to improve representation, integrating fairness constraints during training, and postprocessing outputs to equalize error rates. Continuous monitoring under realworld conditions is essential to detect and correct emerging biases over time. Designing AI systems with these safeguards helps ensure they are both effective and equitable. We focused on the following **group fairness metrics**:

- **Equality of Opportunity**
- **Demographic Parity (Statistical Parity)**
- **Equalized Odds**
- **Predictive Parity**
- **Treatment Equality**

These metrics provide us with the necessary tools to evaluate fairness across groups and determine where and why bias exists within models.

3.3.1 Equality of Opportunity

Equality of Opportunity is a fairness metric that ensures equal true positive rates (TPR) across different demographic groups. The true positive rate for a group g is defined as:

$$TPR_g = \frac{TP_g}{TP_g + FN_g}$$

where TP_g and FN_g represent the number of true positives and false negatives for group g, respectively. The equal opportunity condition implies that:

$$TPR_{g_1} = TPR_{g_2} \quad \text{for all groups } g_1, g_2$$

Implied Bias: If TPR differs significantly across groups, the model is not treating each group equally when predicting positive outcomes, indicating possible dataset imbalance or model bias.

Why Bias Occurs:

- **Dataset Bias:** If some groups are underrepresented, the model may perform worse on these groups.
- **Model Overfitting:** The model may overfit to the majority group, failing to generalize to minority groups.

3.3.2 Demographic Parity (Statistical Parity)

Demographic Parity, or Statistical Parity, ensures that the proportion of positive predictions is the same across different groups. The positive prediction rate for group g is given by:

$$P_{\text{pos},g} = \frac{TP_g + FP_g}{N_g}$$

where TP_g and FP_g are the true positives and false positives for group g, and N_g is the total number of individuals in group g. Demographic Parity requires that:

$$P_{\text{pos},g_1} = P_{\text{pos},g_2} \quad \text{for all groups } g_1, g_2$$

Implied Bias: If the positive prediction rate differs between groups, it suggests that the model is biased towards or against specific groups.

Why Bias Occurs:

- **Representation Bias:** The model may be trained on imbalanced data, leading to unequal prediction rates for different groups.
- **Data Imbalance:** One group may naturally have more positive cases, leading to an increased likelihood of positive predictions for that group.

3.3.3 Equalized Odds

Equalized Odds ensures that both the false positive rate (FPR) and the false negative rate (FNR) are the same across groups. Mathematically, this can be represented as:

$$FPR_g = \frac{FP_g}{FP_g + TN_g}, \quad TPR_g = \frac{TP_g}{FN_g + TP_g}$$

The condition for Equalized Odds requires that:

$$FPR_{g_1} = FPR_{g_2} \quad \text{and} \quad TPR_{g_1} = TPR_{g_2} \quad \text{for all groups } g_1, g_2$$

Implied Bias: If FPR or TPR differ across groups, the model is making more errors (either false positives or false negatives) for certain groups, indicating potential biases in the decision-making process.

Why Bias Occurs:

- **Class Imbalance:** An imbalanced dataset may cause higher error rates for underrepresented groups.
- **Model Overfitting:** The model may overfit to the majority group, leading to higher error rates for minority groups.

3.3.4 Predictive Parity

Predictive Parity ensures that the positive predictive value (PPV) is the same across all groups. The positive predictive value is given by:

$$PPV_g = \frac{TP_g}{TP_g + FP_g}$$

The condition for Predictive Parity is:

$$PPV_{g_1} = PPV_{g_2} \quad \text{for all groups } g_1, g_2$$

Implied Bias: If the predictive accuracy (PPV) differs between groups, it suggests that the model is less accurate for some groups when predicting positive outcomes.

Why Bias Occurs:

- **Data Imbalance:** Some groups may have more high-quality data, leading to better prediction accuracy for those groups.
- **Feature Imbalance:** The model may rely on features that are more predictive for certain groups.

3.3.5 Treatment Equality

Treatment Equality requires that the ratio of false negatives to false positives is equal across different groups. It ensures that the costs of misclassification (i.e., missing true positives or incorrectly predicting positives) are balanced between groups.

Mathematically, for each group g:

$$\text{Treatment Equality}_g = \frac{FNR_g}{FPR_g}$$

The condition for Treatment Equality is:

$$\text{Treatment Equality}_{g_1} = \text{Treatment Equality}_{g_2} \quad \text{for all groups } g_1, g_2$$

Implied Bias: If the ratio of false negatives to false positives differs significantly between groups, it suggests that the model imposes different misclassification costs on different groups.

Why Bias Occurs:

- **Model Sensitivity Issues:** The model may have different decision thresholds or sensitivities across groups.
- **Data Imbalance:** Different distributions of positive and negative labels across groups can lead to unequal error rates.

- **Historical Bias:** If past data encodes unequal misclassification rates, the model may inherit this bias.

3.4 Using Wasserstein Distance for Bias Estimation

In addition to fairness metrics, we employed the **Wasserstein Distance**, which, although not a direct fairness metric, helps estimate model bias by comparing the distribution of predictions across different groups. The Wasserstein Distance is a measure of the distance between two probability distributions, and in this context, it quantifies the disparity between the predicted probabilities (or scores) for different demographic groups.

3.4.1 Wasserstein Distance Overview

The Wasserstein Distance (also known as the Earth Mover's Distance) between two probability distributions P and Q is defined as:

$$W(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma} [\|x - y\|]$$

where $\Gamma(P,Q)$ is the set of all joint distributions with marginals P and Q, and $\|x - y\|$ is the distance between the points x and y. In simpler terms, the Wasserstein Distance computes the minimum "cost" of transforming one distribution into another.

For our case, we use the Wasserstein Distance to measure the disparity between the predicted probability distributions for different demographic groups, such as gender, ethnicity, or other sensitive attributes.

3.4.2 Process of Using Wasserstein Distance for Bias Estimation

To estimate bias using the Wasserstein Distance, we follow these steps:

1. **Obtain Predicted Probabilities:** For each group g, we extract the predicted probabilities P_g of positive predictions (or relevant outcome) from the model. This is typically the output of a classifier that produces probabilities for each class (e.g., for a binary classifier, this would be the probability of predicting the positive class).

2. **Compare Group Distributions:** We compute the Wasserstein Distance between the predicted probability distributions of two groups, P_g and P_h, for each pair of groups g and h. This gives a measure of how different the prediction distributions are between groups.

$$W(P_g, P_h) = \inf_{\gamma \in \Gamma(P_g,P_h)} \mathbb{E}_{(x,y)\sim\gamma} [\|x - y\|]$$

A smaller Wasserstein Distance indicates that the distributions of predicted probabilities between the groups are similar, suggesting less bias, whereas a larger value suggests greater disparity and potential bias.

3. **Interpret Wasserstein Distance Values:** The Wasserstein Distance value helps us interpret the degree of bias between the groups. If the Wasserstein Distance is large, it implies that the model is favoring one group over another in terms of prediction distributions. This can indicate bias towards the group with more favorable outcomes. A small Wasserstein Distance, on the other hand, implies that the model's predictions are more evenly distributed across the groups, indicating reduced bias.

4. **Bias Implications:** By analyzing the Wasserstein Distance across different group pairs, we can infer the following:

    - **Disparity in Predictive Outcomes:** A larger distance suggests that the model's predictions are not equally distributed among groups, indicating that the model is treating groups differently.

    - **Model Fairness:** If the Wasserstein Distance is large across certain groups, it might imply that the model is biased, favoring one group over others, either due to dataset biases, model architecture, or overfitting to certain features.

    - **Potential Areas for Improvement:** By understanding which groups show the largest Wasserstein Distance, we can target those areas for further model refinement and fairness improvement, such as by rebalancing the training data or applying fairness constraints during training.

5. **Experimentation and Future Work:** In future work, we will continue to investigate different methods to mitigate bias using the Wasserstein Distance. This could involve applying domain adaptation techniques, re-weighting the loss function to minimize Wasserstein Distance during training, or leveraging fairness-enhancing algorithms that focus on reducing disparity in prediction distributions between groups.

3.4.3 Advantages of Using Wasserstein Distance

The Wasserstein Distance is particularly useful because it provides a global comparison between two distributions, rather than just comparing individual metrics such as true positive rate or false positive rate. It is also sensitive to the entire distribution of predicted values, which makes it effective in detecting more subtle forms of bias that might not be captured by other fairness metrics.

3.4.4 Conclusion

By incorporating the Wasserstein Distance into our analysis, we gain a deeper understanding of how the model's predictions differ across demographic groups. This allows us to quantify the degree of bias in the model and provides insights into areas where the model may be favoring one group over another. In conjunction with the other fairness metrics, the Wasserstein Distance enhances our ability to assess and improve fairness in AI models.

3.5 Future Directions

In the next phase of our research, we aim to explore new fairness evaluation methods and mitigation techniques. This will involve testing more advanced models, evaluating their bias using the fairness metrics mentioned, and exploring more effective ways to mitigate bias in AI systems.

Chapter 4

Discussions

4.1 Summary of Key Findings

In this work we evaluated bias in AI models of increasing complexitystarting from LeNet-5 on handwritten and fashion datasets, through DeepFace classifiers (gender and ethnicity), to ResNet-based CIFAR-10. Across all experiments:

- **LeNet-5 (MNIST)** showed balanced performance with low Demographic Parity (DP 0.090.11) and high Equality of Opportunity (EOpp 0.980.99). However, **Treatment Equality (TE)** was inconsistente.g., TE = 1.02 for digit 5 vs. 26.70 for digit 3highlighting class-specific bias.
- **LeNet-5 (Fashion MNIST)** had similar trends. Despite good recall (EOpp), classes like 1 and 8 had extremely high TE (51.00, 28.28), indicating uneven misclassification.
- **DeepFace Gender** classifier showed perfect recall for males (EOpp = 1.00) but lower for females (0.91), and **zero false positives for females**, leading to undefined or infinite TE.
- **DeepFace Ethnicity** model favored some classes (Black: PPV = 0.86) but had low recall for others (Indian, Latino-Hispanic: EOpp 0.36), with TE as high as 29.6suggesting serious fairness gaps. .
- **ResNet (CIFAR-10)** maintained low DP (ś0.01) but recall varied (0.590.84), and TE remained high for certain classes (e.g., 14.97 for "ship"), indicating class imbalance in treatment.

4.2 Interpretation of Fairness Metrics

The suite of metrics highlighted different aspects of bias:

- **Demographic Parity (DP):** Low class- or group-level allocation bias in most models, though small DP gaps still translate into unequal access.
- **Equality of Opportunity (EOpp) / Equalized Odds (TPR):** Revealed sensitivity imbalancesmodels often under-detect minority classes (e.g. cat in CIFAR-10, female in gender classification).

- **Equalized Odds (FPR):** Nonzero FPR for privileged groups (e.g. male) indicates unfair over-labeling, whereas zero FPR for others (e.g. female) may hide underprediction bias.
- **Predictive Parity (PPV):** Precision disparities highlight where a positive label is more or less trustworthy across groups.
- **Treatment Equality (TE):** Amplified differences by combining FP and FN ratesclasses with TE 1 suffer disproportionate misclassification cost.

4.3 Cross-Model Comparison

Comparing across architectures and tasks shows:

- **Simple models** perform fairly on clean data (MNIST) but falter on complex sets (Fashion MNIST) with high TE.
- **DeepFace models** show demographic biasgender bias in recall, and severe class bias in ethnicity prediction.
- **ResNet** performs better in parity, but still struggles with treatment fairness across classes.

4.4 Implications for AI Deployment

Our findings suggest:

- **Ethical Risk:** Unmitigated bias in high-stakes domains (e.g. biometric screening) can perpetuate discrimination and unfair resource allocation.
- **Regulatory Compliance:** Models violating DI or EO may fail emerging fairness regulations (EU AI Act, US AI Bill of Rights).
- **Trust and Adoption:** Transparency in multi-metric fairness reporting can improve stakeholder confidence and guide mitigation.

4.5 Limitations of the Study

- **Dataset Scope:** We focused on canonical benchmarks and one facial dataset; real-world data may exhibit more complex biases.
- **Metric Trade-offs:** Some fairness criteria conflict (e.g. DP vs. EO); our aggregated score equally weights each metric, which may not reflect domain priorities.
- **Static Evaluation:** We evaluated trained models post-hoc without exploring in-processing or post-processing mitigation strategies.

4.6 Recommendations and Future Work

To build on this study, we propose:

- **Mitigation Experiments:** Integrate pre-, in-, and post-processing techniques (reweighting, adversarial debiasing, threshold adjustment) and re-evaluate.
- **Dynamic Fairness Monitoring:** Track fairness metrics over time under data drift to sustain equity in production.
- **Task-Specific Weighting:** Tailor metric weighting and thresholds to domain-specific risk profiles (e.g. higher weight on EO in healthcare).
- **Explainability Integration:** Combine fairness metrics with model interpretability tools to diagnose root causes of disparity and guide targeted fixes.

This discussion grounds our empirical findings in ethical, regulatory, and technical contexts, and charts a path toward more equitable AI systems.

Chapter 5

Results And Observations

5.1 Bias and Fairness in AI Models

In machine learning models, bias refers to systematic and unfair discrimination against certain groups or classes during prediction or classification. Fairness metrics aim to evaluate and quantify this bias to ensure equitable model performance across different demographic or categorical groups. We use several fairness metrics in this study:

Demographic Parity (DP): Measures if positive prediction rates are equal across groups.

Equality of Opportunity (EOpp): Measures if true positive rates are equal across groups.

Equalized Odds (EOdds TPR and EOdds FPR): Measures if both true positive and false positive rates are equal across groups.

Predictive Parity (PPV): Measures if precision (positive predictive value) is equal across groups.

Treatment Equality (TE): Measures the balance between false negatives and false positives across groups.

5.2 LeNet-5 Fairness Evaluation

Model Description

LeNet-5 ’s architecture consists of 7 layers in total, which includes 2 convolutional layers, 2 subsampling layers, and 3 fully connected layers. This design is highly effective for simple image classification tasks, such as digit recognition, making it one of the most widely referenced CNN architectures in deep learning research.

Results for MNIST Dataset(Handwritten digits

Digit	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE
Digit 0	0.0983	0.9949	0.9949	0.0009	0.9919	5.7526
Digit 1	0.1135	0.9956	0.9956	0.0006	0.9956	7.8106
Digit 2	0.1030	0.9893	0.9893	0.0010	0.9913	10.6210
Digit 3	0.1000	0.9851	0.9851	0.0006	0.9950	26.7030
Digit 4	0.0978	0.9919	0.9919	0.0004	0.9959	18.3666
Digit 5	0.0919	0.9966	0.9966	0.0033	0.9674	1.0211
Digit 6	0.0951	0.9854	0.9854	0.0008	0.9926	18.8768
Digit 7	0.1029	0.9883	0.9883	0.0014	0.9874	8.0563
Digit 8	0.0974	0.9887	0.9887	0.0012	0.9887	9.2669
Digit 9	0.1001	0.9841	0.9841	0.0009	0.9920	17.8216

Observations

Demographic Parity (DP) is consistently low across all classes, indicating that there is minimal disparity in the proportion of positive classifications for different classes.

Equality of Opportunity (EOpp) and Equalized Odds (EOdds TPR) show high values (mostly above 0.98), indicating the models ability to correctly classify the majority of true positives across classes.

Treatment Equality (TE) varies widely, with class 5 having the lowest TE, and class 3 showing a much higher TE, suggesting potential fairness concerns in some classes where the model may treat certain classes disproportionately in terms of False Positives and False Negatives.

Results for Fashion MNIST Dataset)

Class	DP	EOpp	EOdds TPR	EOdds FPR	PPV	TE
0	0.0929	0.8020	0.8020	0.0141	0.8633	14.0315
1	0.0972	0.9660	0.9660	0.0007	0.9938	51.0000
2	0.1128	0.8810	0.8810	0.0274	0.7810	4.3360
3	0.1069	0.9280	0.9280	0.0157	0.8681	4.5957
4	0.0927	0.7900	0.7900	0.0152	0.8522	13.7956
5	0.0990	0.9680	0.9680	0.0024	0.9778	13.0909
6	0.0989	0.7060	0.7060	0.0314	0.7139	9.3498
7	0.1064	0.9770	0.9770	0.0097	0.9182	2.3793
8	0.0970	0.9560	0.9560	0.0016	0.9856	28.2857
9	0.0962	0.9390	0.9390	0.0026	0.9761	23.8696

Observations

Demographic Parity is low across all classes, indicating minimal bias in positive classification rates between classes.

Equality of Opportunity and Equalized Odds are high in most classes, but class 0 and class 6 have lower values, suggesting poorer performance in these classes.

Treatment Equality shows a wide range, with class 1 having the highest TE value (51.0000), indicating potential disproportionate treatment for this class.

5.3 DeepFace Gender Classifier Fairness Evaluation

Model Description

he DeepFace Gender Classifier uses a deep convolutional neural network (CNN) to predict gender (Man/Woman) from facial images. It consists of convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. The output layer uses a sigmoid activation function for binary prediction. Trained with binary cross-entropy loss and optimized with the Adam optimizer, this architecture efficiently captures facial features for accurate gender classification.

Results

We evaluated the fairness of the DeepFace-based gender classification model using a dataset containing two demographic groups: Man and Woman. The fairness analysis was conducted using several group fairness metrics, namely Demographic Parity, Equality of Opportunity, Equalized Odds (True Positive Rate and False Positive Rate), and Treatment Equality.

TABLE 5.1: Fairness Metrics per Gender

Metric	**Man**	**Woman**
Demographic Parity	0.5469	0.4531
Equality of Opportunity	1.0000	0.9062
Equalized Odds (TPR)	1.0000	0.9062
Equalized Odds (FPR)	0.0938	0.0000
Treatment Equality	0.0000	*inf*

Observations

Demographic Parity shows a slight imbalance, with a higher prediction rate for male individuals (0.5469 for Men, 0.4531 for Women), indicating the model favors males.

Equality of Opportunity is perfect for men (1.0000) but lower for women (0.9062), suggesting better recall for men and a potential disadvantage for women.

Equalized Odds (True Positive Rate) mirrors the Equality of Opportunity values, confirming a disparity in recall between men and women.

Equalized Odds (False Positive Rate) reveals a higher FPR for men (0.0938) compared to women (0.0000), indicating a higher likelihood of incorrect positive predictions for males.

Treatment Equality shows no false positives for women (0.0000) but an undefined or infinite ratio for women due to false negatives, pointing to a potential misclassification bias against women.

Overall, the model performs more favorably for men, particularly in recall, while misclassifying females due to a higher rate of false negatives and no false positives, emphasizing the need for fairness-aware model development.

5.4 DeepFace Ethnicity Classifier Fairness Evaluation

Model Description

The DeepFace Ethnicity Classifier predicts one of five ethnicities (Black, Asian, White, Latino Hispanic, Indian) from facial images using deep learning. It leverages a convolutional neural network (CNN) to automatically extract features from facial images, followed by fully connected layers that classify the image into one of the ethnicity categories. The output layer employs a softmax activation function to produce probabilities for each ethnicity. The model is trained using categorical cross-entropy as the loss function and optimized with the Adam optimizer to ensure efficient learning and accurate predictions.

Results

The fairness of the DeepFace-based ethnicity classification model was evaluated using multiple fairness metrics, including Demographic Parity (DP), Equality of Opportunity (EOpp), Equalized Odds (True Positive Rate and False Positive Rate), Predictive Parity (PPV), and Treatment Equality (TE). The model was tested across five ethnicity groups: Black, Asian, White, Latino Hispanic, and Indian.

TABLE 5.2: Fairness Metrics per Ethnicity

Metric	Black	Asian	White	Latino Hispanic	Indian
Demographic Parity (DP)	0.1716	0.2176	0.2740	0.1700	0.0900
Equality of Opportunity (EOpp)	0.7400	0.8260	0.7900	0.3540	0.3640
Equalized Odds (TPR)	0.7400	0.8260	0.7900	0.3540	0.3640
Equalized Odds (FPR)	0.0295	0.0655	0.1450	0.1240	0.0215
Predictive Parity (PPV)	0.8625	0.7592	0.5766	0.4165	0.8089
Treatment Equality (TE)	8.8136	2.6565	1.4483	5.2097	29.5814

Observations

Demographic Parity: The model predicts White individuals with the highest probability (0.2740), followed by Asian (0.2176), Black (0.1716), Latino Hispanic (0.1700), and Indian (0.0900), suggesting a slight bias towards White ethnicity.

Equality of Opportunity: The model performs best for Asian (0.8260) and White (0.7900) individuals in recall, with significantly lower recall for Latino Hispanic (0.3540) and Indian (0.3640) individuals, indicating underperformance for these groups.

Equalized Odds (True Positive Rate): TPR values align with Equality of Opportunity results, with higher performance for Asian and White individuals and lower performance for Latino Hispanic and Indian groups.

Equalized Odds (False Positive Rate): The highest FPR occurs for White individuals (0.1450) and Latino Hispanic (0.1240), indicating a higher false positive rate for these groups, suggesting misclassification bias.

Predictive Parity (PPV): Precision is highest for Black individuals (0.8625), followed by Indian (0.8089) and Asian (0.7592), while it is lowest for White (0.5766) and Latino Hispanic (0.4165), showing weaker performance for these latter groups.

Treatment Equality: The TE ratio is highest for Indian individuals (29.5814), indicating significant misclassification issues for this group, whereas White individuals have a lower TE ratio (1.4483), suggesting more balanced performance.

5.5 Resnet-18 (Cifar10)

Model Description

ResNet-18 for CIFAR-10 is a deep convolutional neural network consisting of **18 layers**, organized into **8 residual blocks** with **skip (identity) connections**. These skip connections help mitigate the **vanishing gradient problem**, allowing the training of deeper and more accurate networks.

The model is trained on the **CIFAR-10** dataset, which contains **10 object classes** (e.g., airplane, car, bird, cat). It uses **ReLU activations**, **batch normalization**, and a **softmax output layer** for classification.

The network is optimized using **categorical cross-entropy loss** and the **Adam optimizer**, facilitating efficient and stable convergence.

TABLE 5.3: Fairness Metrics per CIFAR-10 Class

Class	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE
airplane	0.0980	0.7660	0.7660	0.0238	0.7816	9.8411
automobile	0.0982	0.8440	0.8440	0.0153	0.8595	10.1739
bird	0.1090	0.6890	0.6890	0.0446	0.6321	6.9800
cat	0.1011	0.5890	0.5890	0.0469	0.5826	8.7654
deer	0.0971	0.7160	0.7160	0.0283	0.7374	10.0235
dog	0.0954	0.6230	0.6230	0.0368	0.6530	10.2508
frog	0.1083	0.8440	0.8440	0.0266	0.7793	5.8745
horse	0.0998	0.7970	0.7970	0.0223	0.7986	9.0896
ship	0.0923	0.8070	0.8070	0.0129	0.8743	14.9741
truck	0.1008	0.8180	0.8180	0.0211	0.8115	8.6211

Observations

Demographic Parity (DP): Bird (0.1090) and frog (0.1083) had the highest DP values, while ship (0.0923) and dog (0.0954) had the lowest, indicating varying prediction rates across classes.

Equality of Opportunity (EOpp): Best recall was achieved for automobile (0.8440), frog (0.8440), and truck (0.8180), while cat (0.5890) and bird (0.6890) had the lowest recall, showing underperformance for these classes.

Equalized Odds (TPR): Trends matched EOpp, with automobile and frog showing the highest TPR (0.8440) and cat showing the lowest (0.5890), indicating recall disparities.

Equalized Odds (FPR): Lowest FPR was for ship (0.0129) and airplane (0.0238), while cat (0.0469) and bird (0.0446) had the highest, showing a bias in false positives for these classes.

Predictive Parity (PPV): Highest precision for ship (0.8743) and truck (0.8115), while cat (0.5826) and bird (0.6321) showed lower precision, reflecting disparities in prediction accuracy.

Treatment Equality (TE): Ship (14.9741) had the highest TE, indicating significant imbalance, while frog (5.8745) had the lowest TE, suggesting more balanced performance.

5.6 Vision TransformerBased Expression Classifier

Model Description

We used the `trpakov/vit-face-expression` Vision Transformer, fine-tuned on a balanced six-class facial expression dataset (angry, disgust, fear, happy, neutral, sad, surprise). The model employs multihead selfattention over image patches, followed by a classification head with softmax activation. Training was done with categorical cross-entropy loss, the Adam optimizer (learning rate 3×10^{-5}), and standard data augmentations (random crop, horizontal flip).

TABLE 5.4: DP, EOpp, EOdds_TPR, EOdds_FPR, PPV, and TE for Facial Expressions

Expression	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE
angry	0.1793	0.6071	0.6071	0.0964	0.9926	4.0765
fear	0.1376	0.4129	0.4129	0.0842	0.6750	6.9743
happy	0.1608	0.8029	0.8029	0.0363	1.3126	5.4342
neutral	0.1436	0.5643	0.5643	0.0620	0.9225	7.0239
sad	0.2593	0.6771	0.6771	0.1783	1.1070	1.8103
surprise	0.1100	0.6057	0.6057	0.0138	0.9903	28.4753

Observations

- **Demographic Parity (DP):** The prediction rates vary noticeablyfrom lowest for *fear* (0.1376) to highest for *sad* (0.2593)indicating that some expressions are far more likely to be predicted than others.
- **Equality of Opportunity (EOpp):** Recall is weakest for *fear* (0.4129) and strongest for *happy* (0.8029), suggesting the model struggles to correctly identify fearful expressions.
- **Equalized Odds (FPR):** False positive rates are highest for *sad* (0.1783) and lowest for *surprise* (0.0138), showing uneven misclassification costs across classes.
- **Predictive Parity (PPV):** The model is most precise on *happy* (PPV=1.3126 ratio indicates overprediction benefit) and least on *fear* (PPV=0.6750), reflecting unequal confidence in positive predictions.

- **Treatment Equality (TE):** Extreme imbalance appears for *surprise* (TE=28.48), meaning its false negatives greatly outweigh false positives, whereas *sad* (TE=1.81) is comparatively balanced.

References

OTAVIO PARRAGA, MARTIN D. MORE, CHRISTIAN M. OLIVEIRA, NATHAN S. GAVEN SKI, LUCAS S. KUPSSINSKÜ, ADILSON MEDRONHA, LUIS V. MOURA, GABRIEL S. SIMÕES, and RODRIGO C. BARROS, "Machine Learning Theory and Applications (MALTA) Lab," PUCRS, Brazil.

Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019. Available: http://www.fairmlbook.org.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 135, 2021.

Dastin, J., "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters*, 2018.

Suresh, H., & Guttag, J. V., "A Survey of Bias in Machine Learning through the Lens of Fairness and Ethics," *Communications of the ACM*, vol. 63, no. 5, pp. 110, 2020.

Pessach, I., & Shmueli, E., "Bias in AI: A Survey on Mitigation Techniques," *Journal of Artificial Intelligence Research*, vol. 69, pp. 243283, 2020.