Fair AI : Estimation of biasness in AI

KESHAV LOHANI
dept. name of organization (of Aff.)
Indian Institute of Information Technology, Allahabad
City, Country
email address or ORCID

TRILOK MEENA
dept. name of organization (of Aff.)
Indian Institute of Information Technology, Allahabad
City, Country
email address or ORCID

DARSHAN VANJARA
dept. name of organization (of Aff.)
Indian Institute of Information Technology, Allahabad
City, Country
email address or ORCID

GORANSH BARDE
dept. name of organization (of Aff.)
Indian Institute of Information Technology, Allahabad
City, Country
email address or ORCID

AMAN RAJ
dept. name of organization (of Aff.)
Indian Institute of Information Technology, Allahabad
City, Country
email address or ORCID

***Abstract*—As Artificial Intelligence (AI) systems play an increasingly influential role in real-world decision-making, addressing issues of fairness and bias has become essential. This research investigates methods for identifying and estimating bias in AI models by first establishing a conceptual understanding of fairness, its definitions, and its practical implications. We examine how different types of bias can emerge from data, model design, and deployment practices.**

The study begins with simple convolutional neural networks and progresses to more complex models, evaluating their behavior across diverse datasets. A range of group-based fairness metrics is applied to assess disparities in model outcomes across different demographic groups. We explore the effectiveness of these metrics in capturing bias and analyze the impact of model architecture and dataset characteristics on fairness. In some experiments, models are retrained or adjusted to observe potential improvements in fairness outcomes.

Our results provide a comprehensive view of how bias can be systematically estimated and addressed in machine learning systems. This work aims to support the development of more equitable AI technologies by offering a structured framework for fairness evaluation and improvement.

***Index Terms*——Artificial Intelligence, Fairness, Bias Estimation, Convolutional Neural Networks, Fair AI, Demographic Disparity, Ethical AI, Group Fairness, Model Evaluation, Dataset Bias.**

I. Introduction

Artificial Intelligence (AI) has rapidly become a central part of decision-making systems in various fields, including finance,healthcare, education, and law enforcement. With the growing influence of these systems, it has become increasingly important to ensure that the decisions made by AI are fair and free from unintended biases. While AI models offer significant advantages in terms of speed and efficiency, they are not immune to the data they are trained on. In fact, if the training data contains historical or societal biases, the models are likely to reproduce and even amplify those biases in their predictions.

Bias in AI can arise from several sources—such as unbalanced datasets, errors in labeling, or assumptions made during the design of the model itself. When these biases go unchecked, they can lead to serious consequences, especially in applications where decisions directly affect individuals. This includes job selection, loan approvals, legal risk assessments, and access to healthcare services. Biased models may disproportionately misclassify or misrepresent certain groups, reinforcing social inequalities and leading to unfair treatment.

Addressing these challenges requires a clear understanding of how and where bias emerges in AI systems. Fairness, in this context, means that models should treat all individuals equally, without being influenced by factors such as race, gender, or ethnicity. On the other hand, bias refers to the consistent differences in performance or outcomes for different groups, often stemming from issues in data or design. Both concepts are closely related, and dealing with one often involves addressing the other.

This research focuses on studying bias in AI models by applying different evaluation techniques across multiple datasets and model architectures. We begin with simpler neural network models and gradually move toward more complex setups and real-world datasets, such as those used in facial recognition. Our approach involves measuring how the model behaves across different groups and using fairness-oriented metrics to understand whether some groups are systematically favored or disadvantaged. The goal is not only to identify bias but also to suggest practical ways to estimate and interpret

it, ultimately contributing to the development of fairer AI systems.

II. Problem Statement

Despite significant advancements in Artificial Intelligence (AI), ensuring fairness in AI-driven decision-making remains a persistent and complex challenge. As AI systems become integral to sensitive applications—ranging from hiring to healthcare diagnostics and financial risk assessment—the risk of perpetuating existing social biases becomes more than just a technical concern; it transforms into a societal and ethical imperative. Traditional approaches often overlook nuanced biases embedded not only in data but also in model architectures, interaction pipelines, and deployment contexts. Moreover, AI systems trained on high-dimensional embeddings or deep features, while promising in performance, can obscure subtle yet harmful biases that standard evaluation metrics fail to capture.

Emerging AI deployments demand more robust, interpretable, and multi-perspective fairness evaluation frameworks that go beyond simple accuracy metrics and basic demographic analysis. Existing models frequently underrepresent intersectional identities and overlook fairness trade-offs across subgroups. This introduces an urgent need to design methodologies that can expose, quantify, and mitigate biases systematically and transparently.

In this study, we aim to address these gaps by systematically evaluating gender bias in neural network-based image classifiers using established fairness metrics. By applying a range of fairness indicators—such as Demographic Parity, Equality of Opportunity, Equalized Odds, Treatment Equality, and Predictive Parity—this work provides an empirical analysis of how bias manifests across different fairness dimensions. The insights from this analysis aim to inform stakeholders on the effectiveness and limitations of current fairness evaluation methods, paving the way for more responsible and informed deployment of AI systems in sensitive domains.

III. Literature Review

The presence of bias in artificial intelligence (AI) systems has garnered significant attention due to its profound ethical, societal, and technical implications. As AI becomes increasingly prevalent in critical sectors such as healthcare, criminal justice, finance, and recruitment, concerns about fairness, accountability, and transparency have intensified. This section reviews key literature on the sources of bias, frameworks for understanding bias, fairness evaluation metrics, and bias mitigation techniques, providing a foundation for the present study.

A. Sources of Bias in AI

Bias in AI systems can originate from various stages of the machine learning lifecycle and is often a reflection of broader societal inequalities. Mehrabi et al. [1] categorized bias into three main types: representational bias, algorithmic bias, and societal bias. Representational bias pertains to the portrayal of groups, as seen in gender stereotypes embedded within image and language models. Algorithmic bias emerges from model design decisions, such as biased feature selection or decision thresholds, while societal bias reflects historical and systemic inequalities captured in the data itself.

Barocas et al. [2] emphasized that historical inequalities are frequently encoded in training datasets, leading to models that perpetuate and even amplify social prejudices. Buolamwini and Gebru [3], in their seminal "Gender Shades" study, demonstrated stark disparities in commercial facial analysis algorithms, reporting error rates of 34.7% for darker-skinned women compared to 0.8% for lighter-skinned men due to demographic underrepresentation.

Further work by Suresh and Guttag [4] presented a pipeline-level taxonomy encompassing historical, representation, measurement, aggregation, evaluation, and deployment biases. Hooker [5] argued for context-sensitive bias mitigation approaches, warning against naive de-biasing strategies. Additionally, Wang et al. [6] highlighted that deeper neural networks can amplify existing biases, while Geva et al. [7] showed the marginalization of minority patterns due to over-reliance on high-frequency examples.

B. Taxonomies and Frameworks for Understanding Bias

To systematically address bias, structured taxonomies have been proposed. Suresh and Guttag [4] developed a comprehensive framework identifying six categories of bias along the data-to-deployment pipeline. Similarly, Olteanu et al. [8] classified biases arising in social data into user bias, platform bias, sampling bias, and societal bias. Mitchell et al. [9] introduced the concept of "Model Cards," a standardized reporting format to enhance transparency and accountability in AI models by detailing performance across demographics, known limitations, and ethical considerations. These frameworks emphasize that bias is a systemic issue requiring interventions at multiple levels, from data collection to deployment, rather than isolated technical fixes.

C. Methods for Measuring Bias and Fairness

The development of fairness metrics has been instrumental in quantifying and evaluating bias in AI systems. Fairness metrics can be broadly categorized into group fairness and individual fairness approaches.

a) Group Fairness Metrics: Group fairness ensures equitable outcomes across demographic groups. Key metrics include:

- Demographic Parity: Requires that the probability of a favorable outcome is independent of the sensitive attribute [10].
- Equalized Odds and Equality of Opportunity: Proposed by Hardt et al. [11], these metrics enforce equal error rates (False Positive Rate and False Negative Rate) and equal True Positive Rates, respectively, across groups.
- Predictive Parity: Ensures equal precision across groups [12].

- Treatment Equality: Focuses on maintaining consistent ratios of false positives to false negatives across groups.
- Wasserstein Distance: A distributional measure used to quantify disparities between the outcome distributions of different groups [13].

b) Individual Fairness Metrics: Individual fairness, introduced by Dwork et al. [14], asserts that similar individuals should receive similar outcomes. Despite its conceptual appeal, defining and implementing individual fairness remains challenging due to the complexity of similarity metrics in high-dimensional spaces.

c) Fairness Toolkits: Several toolkits support the implementation of fairness evaluations:

- IBM AI Fairness 360 (AIF360): Provides fairness metrics, bias mitigation algorithms, and visualization tools [15].
- Microsoft Fairlearn: Offers model assessment and mitigation capabilities with support for multiple fairness metrics [16].

These toolkits facilitate systematic fairness analysis and transparency throughout the AI development process.

D. Bias Mitigation Techniques

Bias mitigation strategies are generally classified into pre-processing, in-processing, and post-processing approaches.

a) Pre-processing Approaches: Pre-processing techniques adjust datasets prior to model training. Kamiran and Calders [17] proposed reweighting and resampling strategies to balance sensitive attributes, while Zemel et al. [18] introduced fair representation learning, transforming data into an intermediate representation that obscures sensitive information.

b) In-processing Approaches: In-processing methods modify the learning algorithm during training to reduce bias. Zhang et al. [19] proposed adversarial debiasing, leveraging a secondary adversarial network to penalize the encoding of sensitive attributes. Other approaches involve incorporating fairness constraints directly into the loss function, ensuring models optimize for both accuracy and fairness.

c) Post-processing Approaches: Post-processing methods adjust model outputs to achieve fairness objectives. Hardt et al. [11] proposed methods to adjust decision thresholds post-training to meet Equalized Odds or Equality of Opportunity constraints.

Recent studies (e.g., Jung et al. [20], Kim et al. [21]) have explored hybrid strategies combining pre-, in-, and post-processing techniques, recognizing that no single approach suffices in all scenarios.

IV. Methodology

In our research on **Fair AI: Estimation of Bias in AI Models**, we systematically explored bias in AI systems through empirical studies using a simple Convolutional Neural Network (CNN) model. Our methodology was structured into several phases, including literature review, model selection, dataset experimentation, bias analysis, and fairness evaluation.

A. Understanding Bias and Fairness in Models

We began our research by exploring the concepts of bias and fairness in AI models through manual observation of model outputs. By prompting the model with various inputs, we examined the responses to identify potential biases. This preliminary phase provided a foundational understanding of how bias may manifest in model outputs, especially when exposed to diverse datasets.

B. Dataset Bias Evaluation

In the next phase, we trained a simple CNN model on multiple datasets to evaluate its performance across different data distributions. This experimentation aimed to identify biases arising from the dataset itself. By training the model on datasets with varying characteristics, such as imbalanced class distributions or skewed demographic representations, we observed how model performance varied based on dataset composition. These observations allowed us to isolate dataset-induced biases affecting model behavior.

C. Estimating Bias Using Fairness Metrics

AI bias refers to systematic errors or prejudices in model predictions that disadvantage certain groups or individuals. Biases can originate from several sources, including unrepresentative training data, subjective labeling processes, or algorithmic design choices. When datasets underrepresent specific demographic subgroups, models tend to underperform for those groups, resulting in higher error rates. Additionally, biases introduced by human annotators can propagate into the model during training. Algorithmic assumptions, such as optimizing solely for overall accuracy, may further exacerbate disparities by prioritizing the majority group's performance while neglecting minority groups. Evaluation practices that rely solely on aggregate metrics can obscure poor performance on underrepresented groups, masking unfair behavior.

To address these issues, bias mitigation requires a comprehensive approach involving: Preprocessing-Improving dataset representation. Training-Incorporating fairness constraints. Postprocessing - Adjusting outputs to equalize error rates. Continuous Monitoring-Detecting emerging biases during deployment.

In our study, we focused on the following group fairness metrics to evaluate model bias:

- Demographic Parity
- Equality of Opportunity
- Equalized Odds
- Predictive Parity
- Treatment Equality

1) 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.

2) 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) 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.

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.

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.

D. Using Wasserstein Distance for Bias Estimation

Beyond fairness metrics, we employed Wasserstein Distance to estimate model bias by comparing prediction distributions across demographic groups.

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

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

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.

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.

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

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) Advantages of Using Wasserstein Distance: Wasserstein Distance is particularly valuable as it provides a **global comparison between entire distributions**, rather than focusing solely on individual metrics such as true positive rate or false positive rate. Unlike traditional fairness metrics that may only capture specific aspects of model performance, the Wasserstein Distance is sensitive to the **entire distribution of predicted values**. This characteristic makes it especially effective in detecting more **subtle forms of bias** that might otherwise go unnoticed. By considering the overall shape and spread of predictions, the Wasserstein Distance helps reveal disparities that could be masked in aggregate metrics.

E. Conclusion

Incorporating the Wasserstein Distance into our analysis allows us to gain a **deeper and more holistic understanding of how the model's predictions vary across different demographic groups**. This metric enables us to **quantify the degree of bias** present in the model and provides insights into **specific areas where the model may favor or disadvantage certain groups**. When used in conjunction with other fairness metrics, the Wasserstein Distance **enhances our capability to comprehensively assess and address fairness in AI models**.

F. Future Directions

In the next phase of our research, we plan to explore **additional fairness evaluation methods and bias mitigation techniques**. This will include testing **more advanced models**, evaluating their bias using the **full set of fairness metrics discussed in this study**, and investigating **more effective strategies to mitigate bias in AI systems**. Our goal is to develop a **robust framework that not only detects but also reduces bias**, ensuring **greater fairness and ethical integrity in AI-driven decision-making systems**.

V. Results and Observations

In machine learning models, **bias refers to systematic and unfair disparities in model predictions that disproportionately affect certain groups or classes**. Such biases

can lead to unequal treatment or outcomes, raising significant ethical and societal concerns, particularly when models are deployed in sensitive applications. To address these issues, **fairness metrics are employed to evaluate and quantify potential biases**, ensuring that the model performs equitably across different demographic or categorical groups.

In this study, we employ several well-established fairness metrics:

A. 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.

TABLE I
RESULTS FOR LENET-5 (HANDWRITTEN DIGITS)

Digit	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE	WD
Digit 0	0.0983	0.9949	0.9949	0.0009	0.9919	5.7526	0.9872
Digit 1	0.1135	0.9956	0.9956	0.0006	0.9956	7.8106	0.9925
Digit 2	0.1030	0.9893	0.9893	0.0010	0.9913	10.6210	0.9848
Digit 3	0.1000	0.9851	0.9851	0.0006	0.9950	26.7030	0.9798
Digit 4	0.0978	0.9919	0.9919	0.0004	0.9959	18.3666	0.9828
Digit 5	0.0919	0.9966	0.9966	0.0033	0.9674	1.0211	0.9793
Digit 6	0.0951	0.9854	0.9854	0.0008	0.9926	18.8768	0.9871
Digit 7	0.1029	0.9883	0.9883	0.0014	0.9874	8.0563	0.9720
Digit 8	0.0974	0.9887	0.9887	0.0012	0.9887	9.2669	0.9793
Digit 9	0.1001	0.9841	0.9841	0.0009	0.9920	17.8216	0.9689

Observations

- Demographic Parity (DP): It 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): It show high values (mostly above 0.98), indicating the model's ability to correctly classify the majority of true positives across classes.
- Equalized Odds (EOdds FPR):False positive rates are very low across all digits($FPR < 0.004$), indicating minimal likelihood of incorrect positive predictions overall. Digit 5 has the highest FPR (0.0033), suggesting a slightly elevated chance of false alarms, while digit 4 has the lowest (0.0004), showing almost no false positives.
- 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.
- Predictive Parity (PPV): PPV is consistently high across digits (> 0.98), showing strong precision. The highest occurs for digit 4 (0.9959) with almost no false positives, while digit 5 has the lowest (0.9674), indicating slightly more false positives but still excellent accuracy overall.
- Wasserstein Distance (WD): The separability of predicted probabilities is consistently very high across all digits ($allWD > 0.96$). Digits *1* (0.9925) and *0* (0.9872) exhibit the greatest separability, indicating the model is highly confident distinguishing these digits from others. In contrast, digit *9* shows the lowest WD (0.9689), suggesting marginally more overlap in predicted probabilities for this class, though still indicating excellent overall discriminative ability.

TABLE II
RESULTS FOR LENET-5 (MNIST FASHION DATASET)

Class	DP	EOpp	EOdds TPR	EOdds FPR	PPV	TE	WD
0	0.0929	0.8020	0.8020	0.0141	0.8633	14.0315	0.7473
1	0.0972	0.9660	0.9660	0.0007	0.9938	51.0000	0.9732
2	0.1128	0.8810	0.8810	0.0274	0.7810	4.3360	0.7843
3	0.1069	0.9280	0.9280	0.0157	0.8681	4.5957	0.8453
4	0.0927	0.7900	0.7900	0.0152	0.8522	13.7956	0.7868
5	0.0990	0.9680	0.9680	0.0024	0.9778	13.0909	0.9452
6	0.0989	0.7060	0.7060	0.0314	0.7139	9.3498	0.6525
7	0.1064	0.9770	0.9770	0.0097	0.9182	2.3793	0.9287
8	0.0970	0.9560	0.9560	0.0016	0.9856	28.2857	0.9707
9	0.0962	0.9390	0.9390	0.0026	0.9761	23.8696	0.9513

Observations

- Demographic Parity (DP): Demographic Parity is low across all classes, indicating minimal bias in positive classification rates between classes.
- Equality of Opportunity (EOpp) and Equalized Odds (EOdds TPR): These are high in most classes, but class 0 and class 6 have lower values, suggesting poorer performance in these classes.
- Equalized Odds (EOdds FPR):FPR values remain low for most classes, reflecting low misclassification risk. Class 6 shows the highest FPR (0.0314), indicating more frequent false positives for this class. Conversely, class 1 (0.0007) and class 8 (0.0016) have the lowest FPRs, suggesting the model is especially conservative in predicting these classes incorrectly.
- Treatment Equality (TE): It shows a wide range, with class 1 having the highest TE value (51.0000), indicating potential disproportionate treatment for this class.
- Predictive Parity (PPV): PPV is generally high across classes, indicating strong precision overall. The model is most precise on class 1 (0.9938) and class 8 (0.9856), while class 6 (0.7139) and class 2 (0.7810) show comparatively lower precision, suggesting more false positives for these classes.
- Wasserstein Distance (WD): Wasserstein Distance varies noticeably across classes, reflecting differences in how confidently the model separates correct and incorrect predictions. The highest separability occurs in class 1 (0.9732), class 8 (0.9707), and class 9 (0.9513), indicating very clear discrimination. In contrast, class 6 (0.6525) and class 0 (0.7473) show lower WD, suggesting more overlap in predicted probabilities and less confident separation for these classes.

B. DeepFace Gender Classifier Fairness Evaluation

Model Description

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.

TABLE III
RESULTS FOR -DEEPFACE (GENDER CLASSIFIER)

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*
Wasserstein Distance	0.8793	0.8793

Observations

- Demographic Parity :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: It 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):It mirrors the Equality of Opportunity values, confirming a disparity in recall between men and women.
- Equalized Odds (False Positive Rate): It 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.
- Wasserstein Distance: It is identical across genders (0.8793), indicating that the separability between predicted probabilities for correct and incorrect classes does not differ between men and women. This suggests comparable model confidence in distinguishing predictions across both groups.

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.

C. 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.

TABLE IV
RESULTS FOR DEEPFACE (ETHNICITY CLASSIFIER)

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
Wasserstein Distance(WD)	0.6540	0.6472	0.5314	0.1450	0.2652

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
- Wasserstein Distance:Wasserstein Distance varies considerably across ethnicities, reflecting different levels of separability between correct and incorrect predictions. The model achieves the highest separability for Black individuals (0.6540) and Asian individuals (0.6472), indicating greater confidence distinguishing these groups. Conversely, predictions for Latino Hispanic (0.1450) and Indian (0.2652) individuals show much lower WD, suggesting higher overlap in predicted probabilities and less certainty in these classifications.

D. 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 V
RESULTS FOR RESNET-18 (CIFAR10)

Class	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE	WD
airplane	0.0980	0.7660	0.7660	0.0238	0.7816	9.8411	0.7123
automobile	0.0982	0.8440	0.8440	0.0153	0.8595	10.1739	0.7657
bird	0.1090	0.6890	0.6890	0.0446	0.6321	6.9800	0.6399
cat	0.1011	0.5890	0.5890	0.0469	0.5826	8.7654	0.5293
deer	0.0971	0.7160	0.7160	0.0283	0.7374	10.0235	0.6577
dog	0.0954	0.6230	0.6230	0.0368	0.6530	10.2508	0.6070
frog	0.1083	0.8440	0.8440	0.0266	0.7793	5.8745	0.7075
horse	0.0998	0.7970	0.7970	0.0223	0.7986	9.0896	0.7827
ship	0.0923	0.8070	0.8070	0.0129	0.8743	14.9741	0.8160
truck	0.1008	0.8180	0.8180	0.0211	0.8115	8.6211	0.7961

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.
- Wasserstein Distance:Separability of predicted probabilities varies across classes. Ship (0.8160) and truck (0.7961) exhibit the highest WD, indicating the model is highly confident distinguishing them. In contrast, cat (0.5293) and dog (0.6070) have the lowest WD, suggesting greater overlap and less discriminative certainty for these categories.

E. Vision Transformer–Based 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 multi-head self-attention 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 VI
RESULTS FOR EXPRESSION CLASSIFIER

Expression	DP	EOpp	EOdds_TPR	EOdds_FPR	PPV	TE	WD
angry	0.1793	0.6071	0.6071	0.0964	0.9926	4.0765	0.5665
fear	0.1376	0.4129	0.4129	0.0842	0.6750	6.9743	0.4504
happy	0.1608	0.8029	0.8029	0.0363	1.3126	5.4342	0.8285
neutral	0.1436	0.5643	0.5643	0.0620	0.9225	7.0239	0.5731
sad	0.2593	0.6771	0.6771	0.1783	1.1070	1.8103	0.5121
surprise	0.1100	0.6057	0.6057	0.0138	0.9903	28.4753	0.7603

Observations

- Demographic Parity (DP): The prediction rates vary noticeably—from 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) and Equalized Odds (TPR): 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.
- Wasserstein Distance (WD): The separability of predicted probabilities across classes is variable. The model shows **strongest separability for *happy* (0.8285) and *surprise* (0.7603)**, suggesting it is comparatively confident in distinguishing these expressions from others. In contrast, **fear exhibits the lowest WD (0.4504)**, indicating the model's predicted probabilities for *fear* samples are more similar to the probabilities assigned to non-*fear* samples, which likely contributes to its lower recall and precision.

VI. DISSCUSSIONS

A. Summary of Key Findings

In this work we evaluated bias in AI models of increasing complexity—starting 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 0.09–0.11) and high Equality of Opportunity (EOpp 0.98–0.99). However, **Treatment Equality (TE)** was inconsistent—e.g., TE = 1.02 for digit 5 vs. 26.70 for digit 3—highlighting 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 29.6—suggesting serious fairness gaps. .
- **ResNet (CIFAR-10)** maintained low DP (±0.01) but recall varied (0.59–0.84), and TE remained high for certain classes (e.g., 14.97 for "ship"), indicating class imbalance in treatment.

B. 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 imbalances—models often under-detect minority classes (e.g. cat in CIFAR-10, female in gender classification).
- **Equalized Odds (FPR):** Non-zero FPR for privileged groups (e.g. male) indicates unfair over-labeling, whereas zero FPR for others (e.g. female) may hide under-prediction 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 rates—classes with TE 1 suffer disproportionate misclassification cost.
- **Wasserstein Distance (WD):**Quantifies how well the model separates true positives from false positives. Higher WD values (closer to 1) indicate strong confidence and clear discrimination between correct and incorrect predictions, while lower WD values signal overlapping distributions and less reliable scoring.

C. 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 bias—gender bias in recall, and severe class bias in ethnicity prediction.
- **ResNet** performs better in parity, but still struggles with treatment fairness across classes.

D. 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.

E. 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.

F. 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.

REFERENCES

[1] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1-35.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

[3] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of FAT/ML.*

[4] Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Unintended Consequences of Machine Learning. *Communications of the ACM*, 64(3), 62-71.

[5] Hooker, S. (2021). Moving Beyond "Algorithmic Bias is a Data Problem". *Patterns*, 2(4), 100241.

[6] Wang, T., Zhao, J., & Russakovsky, O. (2020). Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. *CVPR Workshops*.

[7] Geva, M., Goldberg, Y., & Berant, J. (2019). Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. *EMNLP*.

[8] Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2, 13.

[9] Mitchell, M., et al. (2019). Model Cards for Model Reporting. *Proceedings of FAT/ML*.

[10] Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters*.

[11] Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *NeurIPS*.

[12] Berk, R., et al. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*.

[13] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206-215.

[14] Dwork, C., et al. (2012). Fairness Through Awareness. *Proceedings of ITCS*.

[15] Bellamy, R. K. E., et al. (2019). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*.

[16] Bird, S., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Research*.

[17] Kamiran, F., & Calders, T. (2012). Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, 33(1), 1-33.

[18] Zemel, R., et al. (2013). Learning Fair Representations. *ICML*.

[19] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of AIES*.

[20] Jung, C., et al. (2021). Fairness Beyond Demographics: Ranking Candidates on Dynamic Content with Hybrid Mitigation. *WWW 2021*.

[21] Kim, P. T., & Ghosh, S. (2023). Survey of Hybrid Fairness Interventions in Machine Learning. *ACM Computing Surveys*.