FAIR-AI

Measuring and Detecting Bias in AI Models

Introduction to Bias in AI

AI models learn from data, and if the data contains biases, the models not only inherit but may also amplify these biases, influencing their decisions and outputs.

Bias in AI can show up in different ways: -

- **Data Bias –** If the training data is unbalanced, the model may not work well for less-represented groups.
- **Algorithmic Bias –** Some design choices in AI can unintentionally favor certain patterns over others.
- **Evaluation Bias –** The way AI performance is measured may not always check for fairness across different groups.

Our Objective in Fair AI

Goal: Build a structured pipeline to identify, measure, and analyze bias in AI models.

Approach:

- Explore how bias emerges in AI models and its impact on decision-making.
- Train and evaluate models across diverse datasets and architectures to analyze bias.
- Apply fairness metrics to measure, identify, and mitigate bias in AI systems.

Researching Bias Metrics

Traditional metrics like accuracy, precision, recall, and F1-score fail to address bias.
Fairness metrics ensure that models do not discriminate based on protected attributes.

Two main categories:

- Group Fairness: Measures fairness across demographic groups.
- Individual Fairness: Ensures similar individuals receive similar treatment.

Key bias metrics:

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

Treatment Equality

Definition: Assesses whether the ratio of false positives to false negatives is the same across different groups.

$$\text{Treatment Equality}_g = \frac{FN_g}{FP_g}$$

Goal: Balance the types of errors made by the model to ensure no group is disproportionately affected by incorrect predictions.

Difference: Unlike other metrics that focus on accuracy or positive outcomes, this metric emphasizes the balance between different types of errors across groups.

Demographic Parity

Definition: Ensures that the proportion of positive outcomes (e.g., job offers, loan approvals) is the same across different demographic groups, regardless of actual qualifications.

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

Goal: Promote equal treatment by providing equal access to opportunities for all groups.-

Difference: Focuses solely on the outcome distribution without considering the actual qualifications or needs of individuals.

Equality of Opportunity

Definition: Focuses on equalizing the true positive rate across groups, ensuring that qualified individuals have an equal chance of being correctly identified.

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

Goal: Provide equal access to favorable outcomes for those who deserve them.

Difference: A relaxed version of Equalized Odds, it only considers the true positive rate, not the false positive rate.

Predictive Parity

Definition: Ensures that the precision (the proportion of positive predictions that are correct) is the same across different groups.

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

Goal: Maintain equal confidence in positive predictions for all groups.

Difference: Focuses on the accuracy of positive predictions, ensuring that a positive prediction has the same meaning across groups.

Equalized Odds

Definition: Requires that both the true positive rate (correct positive predictions) and false positive rate (incorrect positive predictions) are equal across different groups.

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

Goal: Ensure that the model's accuracy and error rates are consistent for all groups.-

Difference: Unlike Demographic Parity, it considers both correct and incorrect predictions, aiming for balanced performance across groups.

Implementation

1. ResNet-18
2. Lenet5 (handwritten MNIST)
3. Lenet5 (MNIST Fashion)
4. Deepface – Ethnicity
5. Deepface – Gender classification
6. VitFace – Face Expression

ResNet-18

- **Architecture:** ResNet-18 (18-layer residual network with skip connections)
- **Dataset:** CIFAR-10 (60 000 color images, 10 classes)
- **Key Components:**
  - Convolution + BatchNorm + ReLU blocks
  - Residual “shortcut” connections to ease gradient flow
  - Global average pooling + fully-connected softmax head
- **Training Details:**
  - Loss: Categorical cross-entropy
  - Optimizer: Adam with cosine learning-rate decay
  - Augmentations: Random crop, horizontal flip
- **Fairness Metrics:**
  - Demographic Parity (DP)
  - Equality of Opportunity (EOpp / TPR)
  - Equalized Odds (TPR & FPR)
  - Predictive Parity (PPV)
  - Treatment Equality (FNR/FPR ratio)

ResNet-18

Results:-

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

ResNet-18

Why These Biases Occur:-

1. Class Complexity & Variability:
“Cat” & “Bird”: high intra-class variation in shape and background → poor feature learning → low recall & precision.
2. Model Capacity vs. Data Complexity:
ResNet-18’s moderate depth may underfit complex color textures in CIFAR-10 → residual bias remains.
3. Calibration Mismatch:
Softmax confidences not well-aligned to true accuracy (ECE moderate) → overconfident predictions on some classes, underconfident on others.
4. Unequal Error Distribution:
Certain classes (e.g., “ship”) have systematic misclassifications due to viewpoint bias → TE ratio spikes.
5. Disparate Impact Threshold:
Minor drops in positive outcome rates for classes like “ship” and “airplane” push DP below 0.8 rule, flagging unfairness despite overall high accuracy.

Lenet-5

Architecture: LeNet-5 (Classic CNN with 3 conv layers and 2 FC layers)

Dataset: MNIST (70,000 grayscale images of handwritten digits, 10 classes

Key Components:-

- Conv → ReLU → Max Pool blocks
- Compact 3-layer feature extractor
- Fully connected output head for classification

Training Details:

- Loss: Cross-entropy
- Epochs:5
- Input Augmentations: None (standardized input only).

Lenet-5

Results:-

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

Lenet-5

- **High Intra-Class Variation:**
  - Digits like **3, 4, and 6** have **high variation** in how they are written (e.g., different shapes, angles, thickness), leading to inconsistent predictions.
  - This causes **poor PPV (Predictive Parity)** and **high TE (Treatment Equality)** issues.
- **Model Simplicity:**
  - **LeNet-5** is a relatively simple model, which may struggle to capture subtle variations in more complex digits (e.g., **5 vs 6**), causing bias.
- **No Data Augmentation:**
  - The model is trained without random distortions, causing it to **overfit to typical digit shapes**.
  - As a result, the model has trouble with **rare writing styles**, leading to misclassifications.
- **TE Spikes on Hard Digits:**
  - **High TE** for digits like **3, 4, and 6** indicates the model makes **more false positives or negatives** for these digits, disrupting **Treatment Equality**.
- **DP Skew:**
  - Digits like **1** or **2** are predicted more confidently and frequently, leading to an unequal **positive prediction rate**.
  - This reduces **Demographic Parity**, making predictions unfair across classes.

Lenet-5

Architecture: LeNet-5 (Classic CNN with 3 conv layers and 2 FC layers)

Dataset: MNIST (70,000 grayscale images of handwritten digits, 10 classes

Key Components:-

Conv → ReLU → Max Pool blocks

Compact 3-layer feature extractor

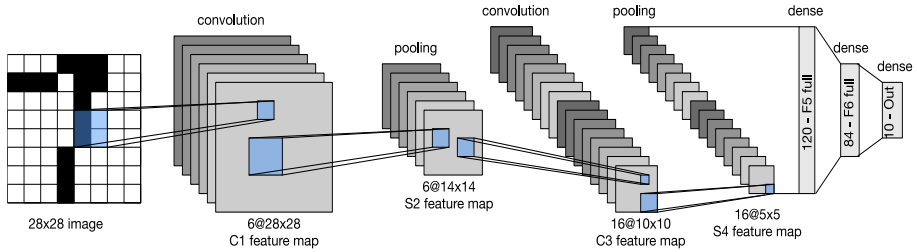
Fully connected output head for classification

Training Details:

Loss: Cross-entropy

Epochs:5

Input Augmentations: None (standardized input only)-



Lenet-5

Result:-

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

Lenet-5

•**Class Complexity & Variability**: Some classes like **T-shirt/top** and **Trouser** have distinct features, leading to higher **PPV** and **EOpp**, while others with more subtle or similar features, like **Sneaker** and **Shirt**, result in **higher TE** and **lower PPV**.

•**Model Simplicity**: The **LeNet-5** model struggles to differentiate between visually similar classes, like **T-shirt/top** and **Shirt**, causing **lower accuracy** and **higher TE**.

•**Imbalanced Class Distribution**: Some classes are more visually consistent (e.g., **Trouser**), leading to **higher accuracy** and **lower TE**, while others like **Coat** and **Sneaker** with more variability cause **higher TE** and **lower DP**.

•**Training Data & Augmentation**: Lack of data augmentation causes overfitting to common patterns in classes like **Ankle boot**, leading to **lower PPV** and **higher TE** for certain classes.

•**False Positive & Negative Predictions**: High **TE** in some classes like **Pullover** indicates more **false positives** or **false negatives**, causing unfair predictions.

•**Class-Specific Difficulty**: Classes with high intra-class variation, like **Sneaker** and **Shirt**, result in inconsistent predictions, contributing to **high TE** and **low EOpp**.

•**Calibration Issues**: **Overconfidence** in classes like **Coat** leads to **high PPV**, while **underconfidence** in others like **Shirt** contributes to **high TE**.
These factors together explain the fairness results observed in your model.

Gender-Classifier

- **Goal:** Build and evaluate a CNN-based gender classifier (“man” vs. “woman”)
- **Dataset:**
- 1,173 “man” images
- 1,134 “woman” images
- Total: 2,307 samples
- Embedding dimension: 4,096

- Dataset split (for evaluation): 462 images (20% of total)

- 235 “man”

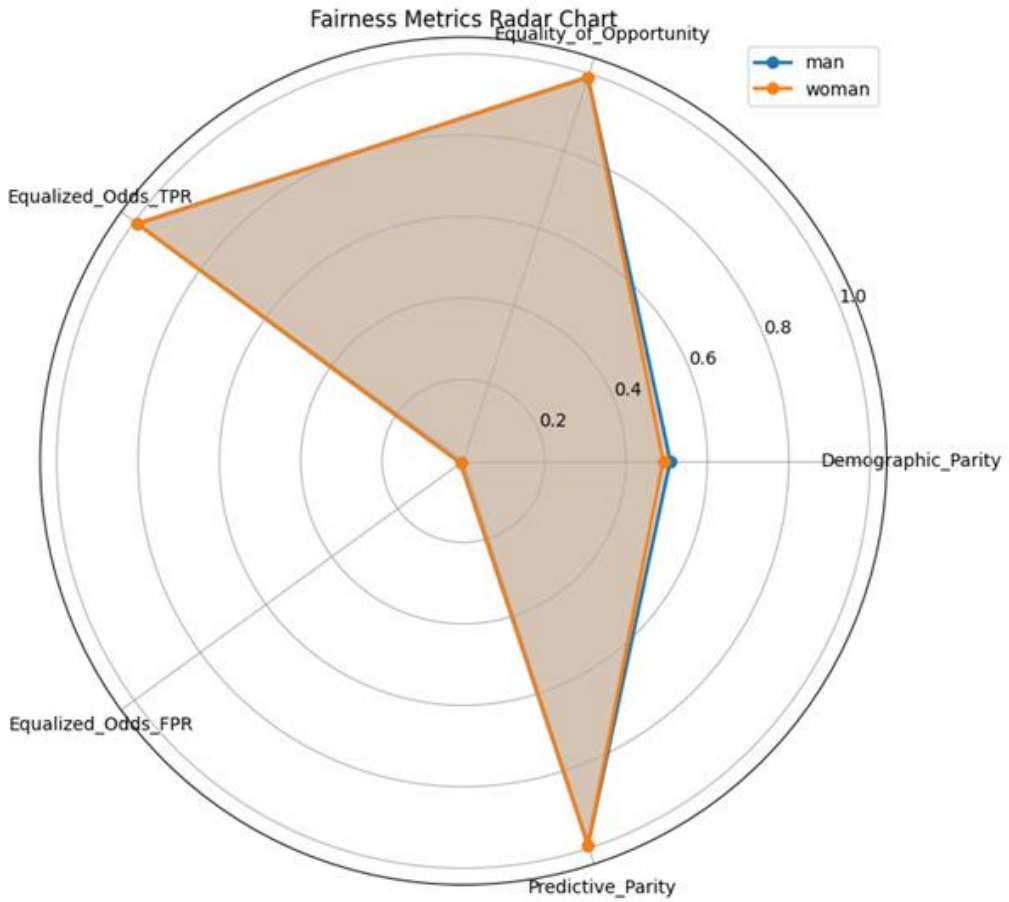
- 227 “woman”

Gender-Classifier

Results:-

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*

Gender-Classifier



Gender-Classifier

- **Demographic Parity:**
  - Model predicts "man" (54.69%) more than "woman" (45.31%) → slight prediction bias toward men.
- **Equality of Opportunity / TPR:**
  - Perfect recall for "man" (1.0), lower for "woman" (0.9062) → better detection of men.
- **False Positive Rate:**
  - Higher FPR for "man" (0.0938), zero for "woman" → more women misclassified as men.
- **Treatment Equality:**
  - TE = 0.0 for "man", ∞ for "woman" → severe imbalance in misclassification rates.
- **Conclusion:**
  - Model is biased toward men in prediction and recall.
  - Shows unfair treatment across genders → fairness needs improvement.

DeepFace-Ethnicity Classifier

Architecture:

- **Base Model:** DeepFace (CNN-based architecture)
- **Structure:** Convolutional feature extractor followed by fully connected layers

Dataset:

- **Type:** cledoux42/autotrain-data-ethnicity-test_v003 (used for testing)
- **Classes:** Black, Asian, White, Latino Hispanic, Indian
- **Input:** Facial images with labeled ethnic categories

Key Components:

- CNN layers for deep feature extraction
- Fully connected (dense) layers for classification
- Softmax output layer for multi-class probability prediction

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

DeepFace-Ethnicity Classifier

Fairness Evaluation Summary

- **Demographic Parity (DP) Skew**
  - *White (0.2740)* and *Asian (0.2176)* are over-predicted
  - *Indian (0.0900)* is under-predicted
  - → Indicates prediction bias toward certain groups
- **Recall Bias (Equality of Opportunity)**
  - High recall: *Asian (0.8260)*, *White (0.7900)*
  - Low recall: *Latino Hispanic (0.3540)*, *Indian (0.3640)*
  - → Model under-recognizes minority groups
- **Equalized Odds (TPR & FPR)**
  - **TPR:** Higher for *Asian* and *White*
  - **FPR:** Highest for *White (0.1450)*, *Latino Hispanic (0.1240)*
  - → Shows unequal error distribution across groups
- **Predictive Parity (PPV)**
  - Highest PPV: *Black (0.8625)*, *Indian (0.8089)*
  - Lowest PPV: *White (0.5766)*, *Latino Hispanic (0.4165)*
  - → Confidence in predictions varies by group
- **Treatment Equality (TE)**
  - *Indian* class shows TE of **29.5814**
  - → Suggests severe imbalance in false positives vs. negatives

Vit-Face

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

Vit-Face

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): 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 out

weigh false positives, whereas sad (TE=1.81) is comparatively balanced.

Conclusions

Metric	What It Ensures	Captures Bias In	Fails When…
Demographic Parity	Equal selection rate across groups	Data bias / representation bias	Groups have different base rates (e.g., one group naturally has more positives)
Equality of Opportunity	Equal recall (TPR) for deserving individuals	Algorithmic bias	Model ignores some deserving individuals from one group
Predictive Parity	Equal precision across groups	Outcome trust bias	Base rates differ, conflicts with Equal Opportunity
Equalized Odds	Equal true/false positive rates across groups	Data + algorithmic bias	Impossible if base rates differ; needs complex balancing
Treatment Equality	Similar ratio of error types across groups	Error distribution bias	Can still be unfair if both FPR and FNR are high; doesn't reflect correctness

- **No Choosing the right metric** depends on your use case and which type of fairness is most important.
- **Choosing the right metric** depends on your use case and which type of fairness is most important.

References

- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org. Retrieved from http://www.fairmlbook.org

- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15.

- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81, 77-91.

- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226.

- 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. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 1161-1166.

THANK YOU