

Abstract

My project focuses on building a **Facial Emotion Recognition (FER)** system that not only performs well but also **addresses demographic bias**. I trained a baseline CNN model and implemented bias mitigation techniques, including **data reweighting, focal loss, and adversarial debiasing**. I also applied transfer learning with ResNet-50 to boost accuracy. The system was evaluated for fairness across demographic lines using inferred metadata. **This work contributes toward building ethical and fair systems.**

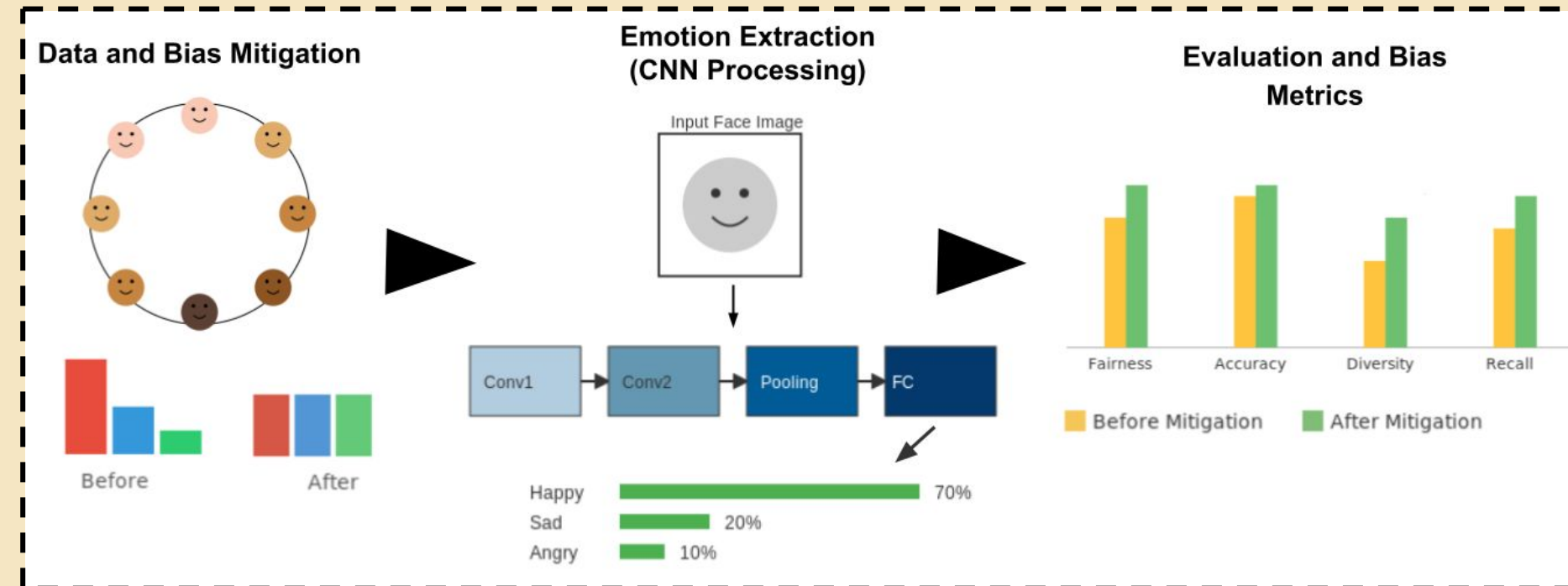


Figure 1: Graphical Abstract

Datasets

I used three publicly available datasets: **FER-2013, RAF-DB, and CK+**. To ensure diverse **representation of facial expressions**. These datasets vary in image resolution, capture conditions, and labeling consistency. To enhance demographic awareness, I applied **demographic inference on each image to identify race, gender, and age**. All images were preprocessed to maintain consistency across datasets during model training.

Methodolog

To assess fairness in facial emotion recognition, I built a pipeline that **preprocesses facial images from FER2013, RAF-DB, and CK+**, then I inferred race, gender, and age using **DeepFace**. I trained a baseline CNN and a ResNet-50 model with transfer learning, applying **bias mitigation techniques including class reweighting, focal loss, and adversarial debiasing**. Performance was evaluated across demographic groups using **accuracy, F1-scores, and fairness metrics**.

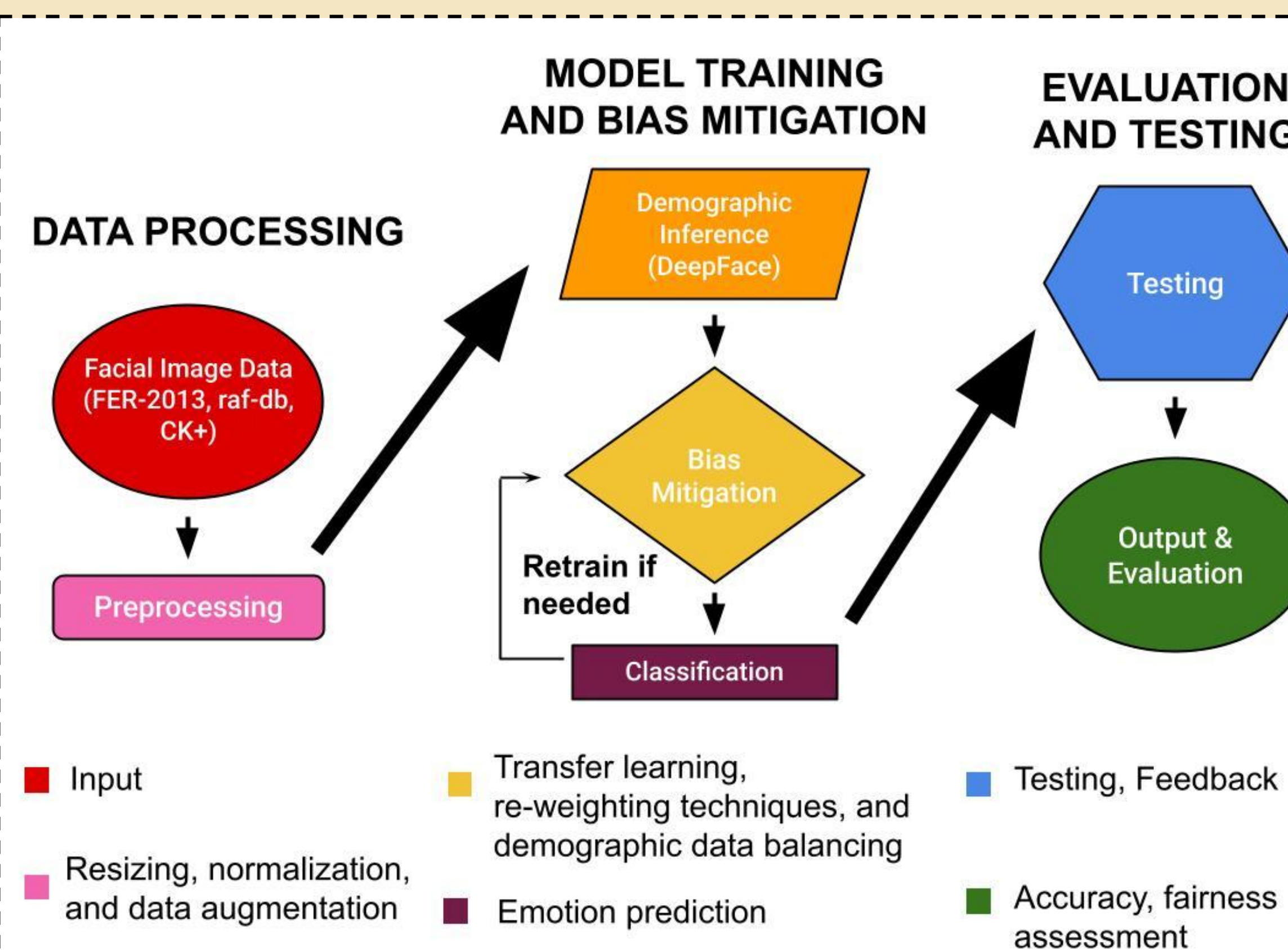


Figure 2: Data Architecture Diagram



Figure 3: Sample images from FER-2013, RAF-DB, and CK+ datasets

Results

I compared a **baseline CNN and ResNet-50** model using accuracy and macro F1-score. The CNN achieved 59% accuracy, but underperformed on minority emotions like **disgust and fear** ($F1 \approx 0.27$). ResNet-50 slightly improved accuracy (60%) but worsened fairness, with **disgust** recall dropping to 9%.

Focal loss addressed **class imbalance**, raising **macro F1 and accuracy**. Adversarial debiasing **achieved the best fairness-performance**.

The table below summarizes each model's performance. These results show that combining transfer learning with bias mitigation **improves both fairness and generalization**.

Model	Accuracy	Macro F1	Weighted F1
CNN Baseline	0.59	0.50	0.58
Resnet-50	0.60	0.50	0.58
+ Focal-Loss	0.67	0.58	0.65
+ Adversarial Debias	0.66	0.59	0.65

Table 1: Validation performance of CNN and ResNet-50 models with and without bias mitigation.

Future Work and Acknowledgment

Future work will explore additional bias mitigation strategies, to improve fairness without sacrificing accuracy. I also plan to expand demographic evaluation to intersectional groups (e.g., age $\times$ race).

Special thanks to Yunting and Charlie for their guidance.