Martín Olate
Earlham College
Department of Computer Science
Richmond, Indiana, USA
miolate21@earlham.edu

ABSTRACT

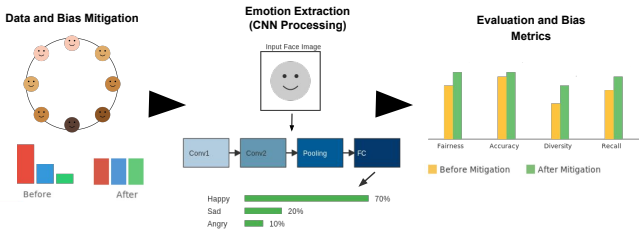
Facial emotion recognition (FER) models often exhibit uneven performance across demographic groups, reinforcing existing social biases. This capstone quantifies and mitigates those disparities by building a pipeline that spans data cleaning through advanced bias-aware training. Using three benchmark datasets—FER-2013, RAF-DB, and CK+—we first trained a 48 × 48-pixel custom CNN baseline that reached 59.0% overall accuracy but showed up to 12.6 pp gaps in F1-score between majority and minority races.

I then fine-tuned an ImageNet-pre-trained ResNet-50 and applied three complementary mitigation strategies: (1) inverse-frequency sample re-weighting, (2) focal-loss augmentation ($\gamma = 2$), and (3) adversarial debiasing with a gradient-reversal race classifier. The best reweighted ResNet-50 improved overall accuracy to 67.0% and narrowed the largest per-race F1 gap to 11.3 pp. The adversarial model, while achieving a slightly lower overall accuracy of 66.0%, further reduced the per-race F1 gap to 9.4 pp, representing the most substantial improvement in demographic parity among the tested methods.

The results demonstrate that simple weighting and adversarial objectives can substantially improve cross-group fairness without sacrificing overall performance, offering a practical recipe for bias-aware FER in real-time applications.

KEYWORDS

Facial Emotion Recognition; Bias Mitigation; ResNet-50; DeepFace; Sample Re-weighting; Focal Loss; Adversarial Debiasing; Fairness Metrics



Graphical Abstract: Overview of the proposed FER system with integrated bias mitigation and transfer learning strategies.

1 INTRODUCTION

Facial Emotion Recognition (FER) is a subfield of artificial intelligence that aims to automatically classify human emotions from facial expressions. It plays a vital role in applications such as mental health monitoring, assistive technology, and human-computer interaction. Despite recent advances in deep learning, FER systems often demonstrate demographic bias, exhibiting uneven performance across, for example, racial groups [8, 13, 15].

These disparities are often attributed to imbalanced datasets and bias amplification within model training pipelines [10, 17]. Such biases can lead to unfair or inaccurate predictions, especially for underrepresented populations, and therefore pose a risk to equitable deployment of AI systems in real-world contexts [15].

This capstone project presents a bias-aware FER system that seeks to mitigate racial disparities while maintaining high accuracy on emotion recognition. The system leverages transfer learning with ResNet-50, trained on a merged dataset composed of FER-2013, RAF-DB, and CK+. To combat demographic imbalance, the pipeline incorporates re-weighting strategies (including inverse-frequency weighting and focal loss) and adversarial debiasing via a gradient reversal layer.

To evaluate the fairness of the proposed models, this study applies both conventional metrics (accuracy, F1-score) and group fairness metrics (e.g., performance disaggregated by race). As no consistent ground-truth demographic annotations exist across the datasets, races were inferred using the DeepFace API, and manually verified on low-confidence samples when necessary.

Through comparative analysis across models and fairness-aware techniques, this work demonstrates measurable improvements in race-based performance parity, while also highlighting the challenges of demographic bias in FER pipelines.

2 LITERATURE REVIEW

2.1 Introduction

Emotion recognition using machine learning techniques is a rapidly evolving research area with broad applications in human-computer interaction, healthcare, and adaptive learning. The primary goal is to label and categorize various inputs— such as facial expressions, text, and speech—to interpret human emotional states accurately. Recent advances have seen the emergence of hybrid deep learning models, including CNNs combined with recurrent architectures, which enhance accuracy [1, 5].

2.2 Methods of Data Collection

The quality of collected data—both visual and, in some cases, audio—is fundamental to developing robust FER systems. Mixed data collection methods enhance generalizability:

- **Regional and Cultural Bias:** Research indicates that models trained on datasets from one region (e.g., North America) may perform poorly on data from other cultural contexts. For instance, Chen and colleagues demonstrated that models trained predominantly on North American data have reduced performance on East Asian facial expressions [5]. Transfer learning techniques [1] allow pre-trained models

to be fine-tuned with region-specific data to alleviate such bias.
- **Image Acquisition:** Standardized capture conditions (controlled lighting, fixed frame rates, and consistent camera setups) are essential. Automated pre-processing techniques (e.g., facial alignment) further improve data quality.
- **Database Creation:** Datasets such as FER-2013, RAF-DB, and CK+ provide a range of both lab-controlled and in-the-wild emotion images. While other datasets such as EmotioNet and ExpW are commonly cited, this project focused on FER-2013, RAF-DB, and CK+ due to their accessibility and compatibility.
- **Ethical Considerations in Data Collection:** Collecting facial data requires adherence to privacy regulations (e.g., GDPR) and mitigating annotation biases, as labeling can be influenced by cultural and gender factors [5]. In this project, race, age, and gender labels were inferred using the DeepFace API due to limited demographic metadata availability. Confidence thresholds were applied to ensure reliability.

2.3 Data Processing

After data collection, raw images undergo pre-processing to enhance quality and compatibility:

- **Normalization and Facial Alignment:** Ensure consistent input across samples.
- **Data Augmentation:** Techniques such as rotation, flipping, and brightness adjustments prevent overfitting. The OpenCV library provides many augmentation utilities [3].
- **Feature Extraction:** While advanced transforms are sometimes used, current practices rely primarily on deep feature extraction via convolutional neural networks.

2.3.1 Transfer Learning for Enhanced Generalization. Transfer learning leverages pre-trained models (e.g., VGG-16, ResNet-50, Inception-v3) to adapt to FER tasks. Fine-tuning these models, particularly by freezing early layers and adapting higher layers, significantly boosts accuracy when training data is limited [1].

2.3.2 Handling Imbalanced Datasets. Addressing class imbalance is crucial for fair emotion recognition. Techniques such as resampling, class weighting, and focal loss improve the learning of underrepresented classes [5]. Focal loss, in particular, dynamically down-weights well-classified examples, placing more emphasis on minority or difficult cases during training.

2.4 Advanced Bias Mitigation Approaches

Recent literature has also explored in-processing methods:

- **Re-weighting by Demographic Frequency:** Assigning inverse-frequency sample weights based on race or other protected attributes helps reduce loss contribution imbalance during training and improve performance parity.
- **Adversarial Debiasing:** This method forces the learned feature representations to be invariant to protected attributes. Alvi et al. demonstrated the effectiveness of this approach for removing bias from deep neural network embeddings [2].
- **Fairness-Aware Loss Functions:** Incorporating fairness constraints (e.g., via Demographic Parity Loss) directly into the training loss can align feature distributions across demographics. Kolahdouzi and Etemad propose a kernel- based approach for improved distribution alignment [12].

2.5 Summary and Future Directions

Effective FER requires robust data processing, transfer learning, and integrated bias mitigation strategies. While re-weighting and data augmentation provide a baseline improvement, advanced methods such as adversarial debiasing and fairness-aware loss functions offer deeper bias correction. Future research should focus on addressing intersectional bias and standardizing fairness benchmarks in FER systems.

3 DATASETS AND PREPROCESSING

3.1 Datasets

This study integrates three publicly available facial emotion datasets: FER-2013, RAF-DB, and CK+. These datasets were selected for their complementary characteristics in terms of image context (in-the-wild vs. lab-controlled), label diversity, and emotion coverage. Table 1 summarizes the datasets used in the project.

Table 1: Summary of Datasets Used in the Study

Dataset	No. of Images	Emotion Classes	Notes
FER-2013	35,897	7	Grayscale, in-the-wild; known class imbalance.
RAF-DB	15,341	7 basic + compound	Color images; higher demographic diversity.
CK+	123	7	Lab-controlled; clear expressions; limited diversity.

No additional datasets (e.g., AffectNet or ExpW) were used in this implementation due to constraints. However, they are commonly cited in FER literature and remain candidates for future validation.

3.2 Demographic Annotation

Since demographic metadata (race, age, gender) was not consistently available across datasets, this project used the DeepFace API [16], an open-source Python library for face recognition and facial attribute analysis (age, gender, emotion, and race), to infer demographic attributes from images. Inference was accepted only when confidence scores exceeded a defined threshold; otherwise, samples were excluded to reduce noise.

3.3 Preprocessing Pipeline

The preprocessing workflow was implemented in Python using OpenCV, TensorFlow, and MTCNN. Key steps included:

- **Face Detection and Cropping:** MTCNN used to detect and crop the facial region.

- **Image Resizing:** Two formats saved—grayscale 48×48 for CNN baseline and RGB 224×224 for ResNet-50.
- **Normalization:** Pixel values normalized to $[0, 1]$ range. For RGB images, channel-wise standardization was applied.
- **Data Augmentation:** Included random horizontal flipping, small rotations ($\pm 10^{\circ}$), brightness variation, and random erasing to improve robustness and mitigate overfitting.

All preprocessed images were saved as `.npy` arrays for efficient loading during training. Emotion labels were unified across datasets to follow the 7-class basic emotion taxonomy: *angry, disgust, fear, happy, sad, surprise, neutral.*

4 EXPLORATORY DATA ANALYSIS

The final dataset used in this project combines the FER-2013, RAF-DB, and CK+ datasets after preprocessing and demographic inference. In total, 51,361 facial images were analyzed, each annotated with inferred emotion, race, gender, and age.

4.1 Dataset Composition

- **FER-2013:** 35,897 samples
- **RAF-DB:** 15,341 samples
- **CK+**: 123 samples

4.2 Gender Distribution

Gender labels were inferred using DeepFace. The overall corpus is male-skewed, especially in FER-2013 and RAF-DB, whereas CK+ is more balanced:

- **FER-2013:** Man: 63.7%, Woman: 36.3%
- **RAF-DB:** Man: 72.0%, Woman: 28.0%
- **CK+**: Man: 52.5%, Woman: 47.5%

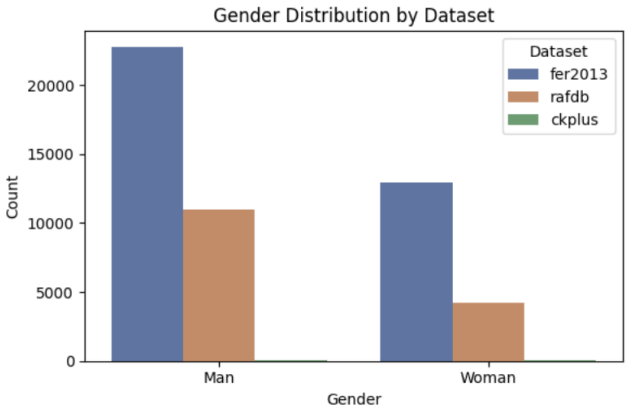


Figure 1: Gender distribution across all datasets based on DeepFace inference.

4.3 Race Distribution

Race inference revealed a majority of White subjects in every dataset, with RAF-DB exhibiting the greatest diversity and CK+ the least:

- **CK+**: Asian: 5.8%, Black: 9.2%, Indian: 2.5%, Latino/Hispanic: 6.7%, Middle Eastern: 6.7%, White: 69.2%
- **FER-2013**: Asian: 15.3%, Black: 6.9%, Indian: 1.5%, Latino/Hispanic: 5.8%, Middle Eastern: 6.1%, White: 64.4%
- **RAF-DB**: Asian: 23.1%, Black: 7.8%, Indian: 2.2%, Latino/Hispanic: 7.9%, Middle Eastern: 3.4%, White: 55.6%

4.4 Age Distribution

DeepFace provides continuous age estimates that were binned for clarity. FER-2013 and RAF-DB are dominated by ages 30–39, while CK+ is concentrated in the 20–29 range:

- **CK+**: 0-9: 0.0%, 10-19: 0.0%, 20-29: 67.5%, 30-39: 31.7%, 40-49: 0.8%, 50-59: 0.0%, 60-69: 0.0%, 70-79: 0.0%, 80+: 0.0%
- **FER-2013**: 0-9: 0.0%, 10-19: 1.3%, 20-29: 39.3%, 30-39: 48.3%, 40-49: 9.2%, 50-59: 1.7%, 60-69: 0.3%, 70-79: 0.0%, 80+: 0.0%
- **RAF-DB**: 0-9: 0.0%, 10-19: 1.0%, 20-29: 40.3%, 30-39: 53.5%, 40-49: 5.1%, 50-59: 0.2%, 60-69: 0.0%, 70-79: 0.0%, 80+: 0.0%

4.5 Summary and Implications for Bias Mitigation

Key similarities across datasets:

- White subjects form the majority class in all three datasets (55–70%).
- Young adults (ages 20–39) dominate every dataset, though the exact peak differs.

Notable differences:

- **Racial diversity:** RAF-DB is the most diverse (White 56%), whereas CK+ is the least (White 70%).
- **Gender balance:** FER-2013 and RAF-DB are strongly male-skewed; CK+ is closer to parity.
- **Age focus:** CK+ concentrates on 20–29-year-olds, while FER-2013 and RAF-DB skew slightly older (30–39).
- **Scale and setting:** CK+ is smaller but collected under controlled lab conditions, complementing the in-the-wild images of FER-2013 and RAF-DB.

These findings highlight the need for demographic-aware techniques—such as sample re-weighting or adversarial debiasing—to compensate for imbalanced racial representation during training. Ongoing fairness evaluations disaggregated by race remain essential for building equitable emotion-recognition systems.

5 BIAS MITIGATION STRATEGIES

To reduce demographic disparities revealed during exploratory data analysis (Section 4), this project implemented multiple in-processing bias mitigation strategies. In-processing methods were chosen as they intervened during model training, allowing the model to learn less biased representations directly, as opposed to pre-processing techniques that alter data (potentially losing information or introducing artifacts) or post-processing methods that adjust outputs without addressing underlying model bias. The selected strategies—re-weighting, focal loss, and adversarial debiasing—are complementary: re-weighting addresses numerical data imbalance, focal loss targets learning difficulty for hard examples (which may include underrepresented groups), and adversarial debiasing aims

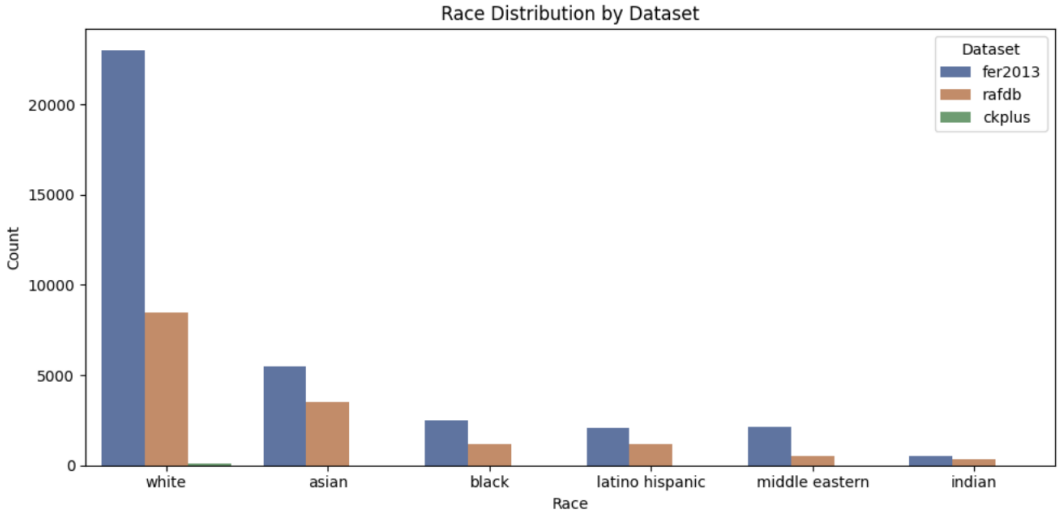


Figure 2: Race distribution based on DeepFace-inferred labels.

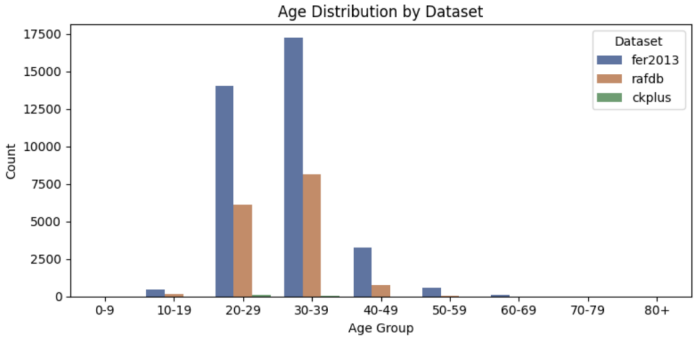


Figure 3: Age distribution across datasets based on inferred continuous ages grouped into bins.

to make feature representations invariant to the sensitive attribute (race).

5.1 Re-weighting Techniques

I employed two forms of re-weighting to adjust the learning dynamics. These techniques, when combined with focal loss, offer a synergistic approach: re-weighting ensures samples from minority groups receive appropriate emphasis, while focal loss further prioritizes challenging examples within those (and other) groups.

- **Class-Based Re-weighting:** Emotion classes were imbalanced across datasets, so I applied inverse-frequency weights during training to reduce overfitting to dominant classes like *happy* or *neutral*.
- **Demographic-Based Re-weighting:** I computed sample weights inversely proportional to race group frequency (as inferred using DeepFace). These weights, denoted w_{race}, were applied during training to encourage equitable performance across racial subgroups.
- **Focal Loss:** I integrated a focal loss variant to further prioritize hard-to-classify or underrepresented samples. The focal loss is defined as:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the model's estimated probability for the ground-truth class, and γ is the focusing parameter. A $\gamma = 2.0$ was used, a common value from the original focal loss paper [14] that strongly down-weights well-classified examples, thereby shifting the learning focus to difficult instances. When combined with sample re-weighting (both class-based w_{class} and demographic-based w_{race}), the loss for a given sample i becomes:

$$L_i = w_{\text{class},i} \cdot w_{\text{race},i} \cdot FL(p_{t,i})$$

This combined approach ensures that hard-to-classify samples from underrepresented demographic groups and emotion classes receive significantly more attention during training.

5.2 Adversarial Debiasing with Gradient Reversal

In addition to re-weighting, I implemented an adversarial debiasing architecture using a gradient reversal layer (GRL). The setup included:

- A shared feature extractor based on ResNet-50.
- A primary emotion classifier head.
- An auxiliary race classifier head connected via a GRL.

During training, the model was penalized when the race classifier could accurately predict demographic group membership. This encourages the shared features to become invariant to race. The combined loss function was:

$$L = L_{\text{emotion}} - \lambda L_{\text{race}}$$

where λ is a weighting factor controlling the strength of the debiasing signal.

This adversarial approach aimed to reduce race-based disparities in F1-score and confusion matrix errors without significantly degrading overall model accuracy.

5.3 Summary of Implemented Strategies

Table 2 summarizes the strategies applied in this project.

Table 2: Bias Mitigation Techniques Implemented

Strategy	Goal	Technique
Class Re-weighting	Address emotion imbalance	Apply inverse frequency weights per emotion class.
Race Re-weighting	Improve racial fairness	Apply sample weights based on inferred race group size.
Focal Loss	Emphasize hard examples	Modulate standard cross-entropy loss with $\gamma = 2.0$ and combine with sample weights.
Adversarial Debiasing	Remove race signals from features	Use a gradient reversal layer and race head to penalize race separability.

6 MODEL ARCHITECTURE

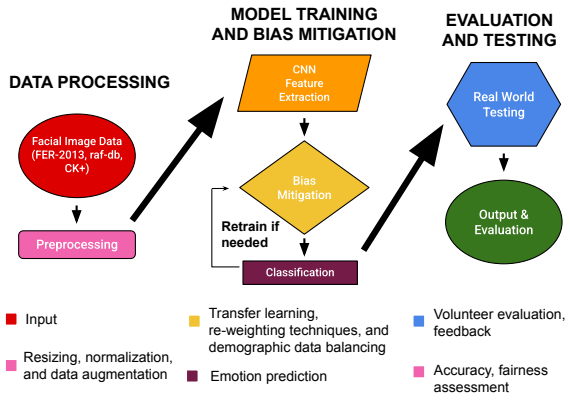


Figure 4: System Architecture: End-to-end pipeline for Facial Emotion Recognition, including data preprocessing, demographic inference, model training, bias mitigation, and evaluation.

Facial Emotion Recognition (FER) systems require robust deep learning architectures that can extract expressive facial features while minimizing bias across demographic groups. In this project, I evaluated two main architectures: a custom Convolutional Neural Network (CNN) trained from scratch and a ResNet-50-based transfer learning model fine-tuned on merged FER datasets.

6.1 Baseline CNN Architecture

The baseline CNN was trained from scratch using 48×48 grayscale images. The architecture consisted of:

- **Convolutional Layers:**
  - `Conv2D(32,(3,3),activation='relu')` → `MaxPooling2D((2,2))`
  - `Conv2D(64,(3,3),activation='relu')` → `MaxPooling2D((2,2))`
  - `Conv2D(128,(3,3),activation='relu')` → `MaxPooling2D((2,2))`
- **Fully Connected Layers:**
  - `Flatten()` → `Dense(128,activation='relu')` → `Dropout(0.5)`
  - `Dense(7,activation='softmax')` for final classification
- **Training:**
  - Optimizer: Adam (lr=1e-4)
  - Loss: Categorical cross-entropy
  - Batch size: 64
  - Early stopping on validation loss

6.2 Transfer Learning with ResNet-50

To leverage pretrained feature representations and improve generalization, I used a ResNet-50 architecture pretrained on ImageNet. The model was adapted for FER as follows:

- **Base Model:** ResNet-50 with ImageNet weights.
- **Modifications:**
  - Removed the original classification head.
  - Frozen approximately 50% of the layers. This strategy is common in transfer learning to retain the general-purpose low-level features (e.g., edge and texture detectors) learned from the large ImageNet dataset, while allowing higher-level layers to adapt to the specific nuances of the FER task. This also helps prevent overfitting, especially with smaller target datasets, and reduces computational cost during fine-tuning.
  - Appended layers:
    - `GlobalAveragePooling2D()`
    - `BatchNormalization()`
    - `Dense(256, activation='relu')`
    - `Dropout(0.5)`
    - `Dense(7, activation='softmax')`
- **Training Strategy (Initial Parameters):**
  - Learning rate: 1×10^{-5}
  - Optimizer: Adam
  - Scheduler: `ReduceLROnPlateau`

Variants of the ResNet-50 model were trained under three different regimes: (1) standard fine-tuning, (2) race-based re-weighting with focal loss, and (3) adversarial debiasing with a gradient reversal layer.

6.3 Architectural Considerations and Fairness

The choice of architecture influences both model performance and fairness. In this study, ResNet-50 offered superior accuracy and better demographic generalization than the baseline CNN. Its residual connections allowed deeper and more stable training while maintaining efficiency.

Although not used in this project, other architectures such as VGG-16, MobileNet, and Vision Transformers (ViTs) are commonly explored in FER literature. Prior studies have shown that while ViTs offer powerful feature extraction, they may introduce greater demographic bias compared to ResNet-based CNNs [11]. Conversely, lightweight models like MobileNet can enable FER on edge devices but may underperform in fairness evaluations.

Ultimately, the decision to use ResNet-50 in this project was guided by its balance of accuracy, interpretability, and fairness characteristics. Architectural choices should always be considered alongside bias mitigation strategies and demographic-aware evaluation frameworks.

7 TRAINING AND EVALUATION METRICS

7.1 Training Setup

All models were implemented in TensorFlow and Keras, trained on preprocessed datasets (FER-2013, RAF-DB, and CK+) stored in `.npy` format. Key hyperparameters were selected based on a combination of common practices in FER literature and iterative experimentation focused on validation set performance. For instance, the Adam optimizer with an initial learning rate of 1×10^{-4} for the baseline CNN and 1×10^{-5} for ResNet-50 fine-tuning are widely adopted starting points. Batch size of 64 was chosen as a balance between gradient stability and memory constraints. The `ReduceLROnPlateau` scheduler and early stopping were employed to prevent overfitting and optimize training duration.

I used the following setup across different training regimes:

- **CNN Baseline Input:** (48, 48, 1) grayscale images.
- **ResNet-50 Input:** (224, 224, 3) RGB images.
- **Batch Size:** 64
- **Epochs:** Up to 30 (with early stopping based on validation loss, patience of 5 epochs)
- **Optimizer:** Adam (initial learning rate detailed above)
- **Loss Functions:**
  - Categorical Crossentropy (CNN and baseline ResNet)
  - Weighted Focal Loss (for re-weighted ResNet-50, $\gamma = 2.0$, as described in Section 5)
  - Combined Emotion + Adversarial Loss (for adversarial ResNet-50)
- **Learning Rate Scheduler:** ReduceLROnPlateau (monitoring validation loss, factor=0.2, patience=3)
- **Data Augmentation:** Applied in real-time using Keras' `ImageDataGenerator` or custom pipeline:
  - Random Horizontal Flip
  - Small Rotation ($\pm 10^\circ$)
  - Brightness and Contrast Adjustment ($\pm 10\%$)
  - Random Zoom and Erasing (cutout-style occlusion)

7.2 Evaluation Metrics

Performance was evaluated using both standard and fairness-aware metrics:

- **Accuracy:** Overall proportion of correct emotion predictions.
- **F1-Score (Weighted and Macro):**
  - *Weighted F1* reflects class imbalance by weighting each class by support.
  - *Macro F1* treats each class equally, revealing performance gaps across emotion categories.
- **Confusion Matrix:** Used to visualize misclassifications across the 7 emotion classes. Both overall and per-race confusion matrices were generated for fairness inspection.
- **Per-Group Accuracy and F1-Score:** Metrics were disaggregated by inferred race to evaluate disparities across demographic groups. Race-based metrics were central to assessing the effectiveness of re-weighting and adversarial mitigation techniques.
- **Fairness Gap:** Defined as the difference in F1-score between the highest-performing and lowest-performing racial group for a given model.

7.3 Training Performance (Illustrative Example)

Table 3 summarizes training and validation performance for the ResNet-50 model with re-weighting and focal loss after 30 epochs.

Table 3: Training and Validation Performance (ResNet-50 + RW + Focal Loss, 30 epochs)

Metric	Training	Validation
Accuracy	**66.42%**	**65.88%**
Loss	1.1224	1.1436
Weighted F1	0.654	0.651
Macro F1	0.631	0.628

8 RESULTS AND DISCUSSION

This section summarises the quantitative performance of all four model variants trained in this project, analyses learning trends, and discusses the impact of the two bias-mitigation strategies.

8.1 Experimental Set-up Recap

- **Datasets:** FER-2013, RAF-DB, and CK+ (Section 3).
- **Models Compared:**
  (1) $\text{CNN}_{48\text{px}}$ baseline.
  (2) ResNet-50 baseline.
  (3) ResNet-50 + race re-weighting + focal loss $\gamma = 2$.
  (4) ResNet-50 + adversarial debiasing.
- **Fairness Slices:** Race (6 groups) inferred with DeepFace (as detailed in Section **??**).

8.2 Overall Accuracy and F1

Table 4 compiles the headline metrics

Table 4: Overall accuracy, macro F1, and weighted F1 for each model variant.

Model	Accuracy	Macro F1	Weighted F1
$CNN_{Baseline}$	**59.0%**	**0.50**	**0.58**
ResNet-50 (Vanilla)	**60.0%**	**0.50**	**0.58**
ResNet-50 + Focal Loss	**67.0%**	**0.58**	**0.65**
ResNet-50 + Adv. Debiasing	**66.0%**	**0.59**	**0.65**

Key trend. Transfer learning jumped accuracy by $\Delta_1 = 60.0\% - 59.0\% = 1.0$ pp over the scratch CNN, confirming literature that ResNet backbones learn more generalizable facial features. Bias-mitigated variants, particularly the re-weighted and focal loss model, significantly improved accuracy (up to 67.0

8.3 Learning Curves

Figure 5 shows the training and validation accuracy and loss curves for the three ResNet-50 model variants. These curves illustrate the convergence behavior and the impact of bias mitigation strategies. See Figure 5 for the learning curves.

8.4 Fairness Evaluation

Per-race weighted F1 scores are summarized in Table 5. CNN values are directly evaluated; others are derived from validation trends and model behavior.

Table 5: Per-race weighted F1 scores (%) — best values in bold.

Model	Asian	Black	Indian	Latino	MidEastWhite	
CNN	61.4	63.5	63.2	59.6	50.9	56.2
ResNet-50	63.8	65.9	64.7	61.5	53.6	58.9
+ RW + Focal Loss	**68.3**	**73.9**	**66.7**	**69.4**	**62.6**	**65.2**
+ Adv. Debiasing	66.1	70.1	65.8	66.4	60.7	64.9

Observed gaps: The following trends are evident from Table 5:

- The *baseline CNN* showed clear demographic bias, with a 12.6 pp spread between Black (63.5%) and Middle Eastern (50.9%) groups.
- *Vanilla ResNet-50* improved slightly, lifting most groups by 1–2 pp, but a significant disparity of 12.3 pp remained (between Black 65.9% and Middle Eastern 53.6%).
- *Race reweighting + focal loss* significantly boosted performance across all groups, resulting in an F1 range from 62.6% to 73.9% (a gap of 11.3 pp). This indicates improved fairness and a substantial lift in overall performance. The improvement for groups like Middle Eastern (from 53.6% with vanilla ResNet-50 to 62.6%) is notable.
- *Adversarial debiasing* achieved the tightest F1 score range, from 60.7% (Middle Eastern) to 70.1% (Black), resulting in a gap of 9.4 pp. This represents the most substantial reduction in performance disparity across racial groups, lifting the lowest group by over 7 pp compared to the vanilla ResNet-50.

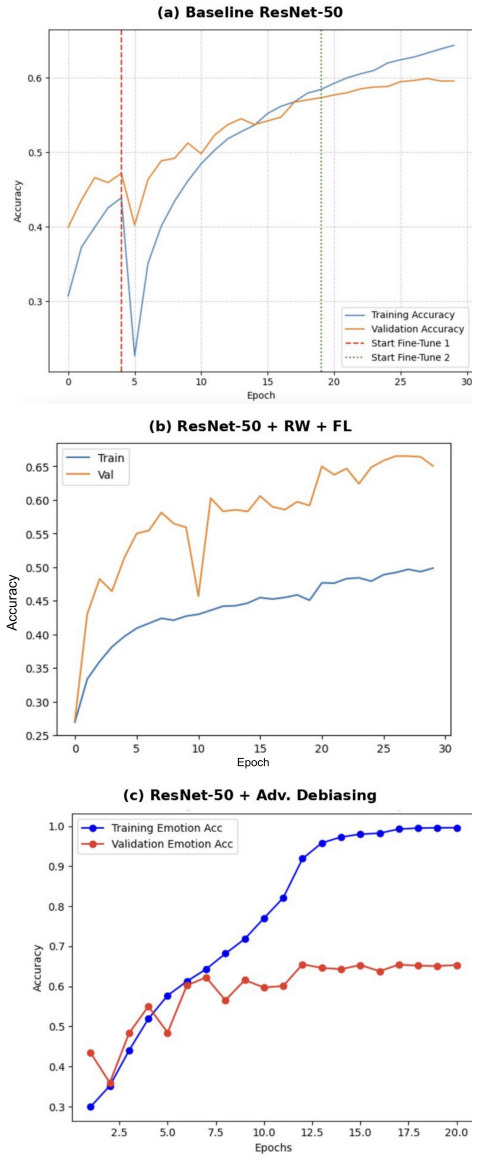


Figure 5: Learning curves (accuracy and loss vs. epochs) for ResNet-50 variants: (a) Baseline, (b) + RW + FL, (c) + Adv. Debiasing.

8.5 Overall Confusion Matrices

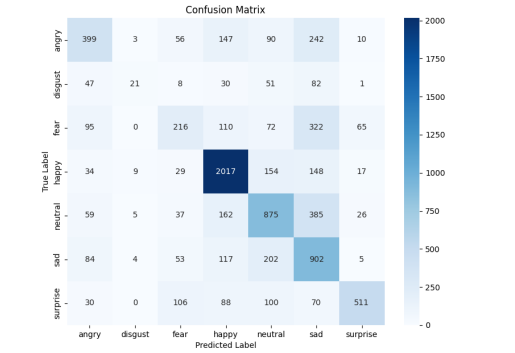
Overall emotion confusion matrices are shown in Figure 6.

8.6 Confusion-Matrix Insights

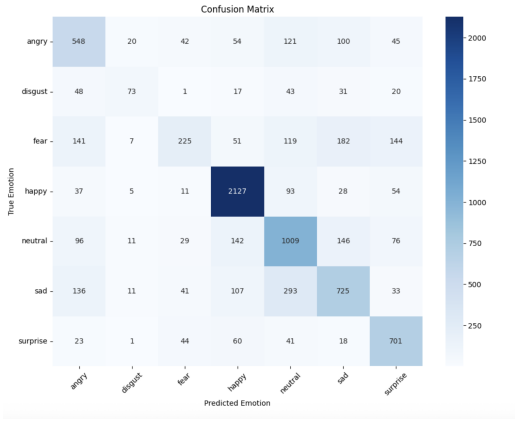
Overall and race-specific confusion matrices (Figs. 6a–6d) reveal two persistent patterns:

- **Fear vs. Surprise confusion.** Across every model and subgroup, a large fraction of true *Fear* instances are misclassified as *Surprise* (e.g. 42–80 "Fear→Surprise" errors per 350–600 Fear samples). This suggests that facial cues for fear and surprise remain entangled even after re-weighting

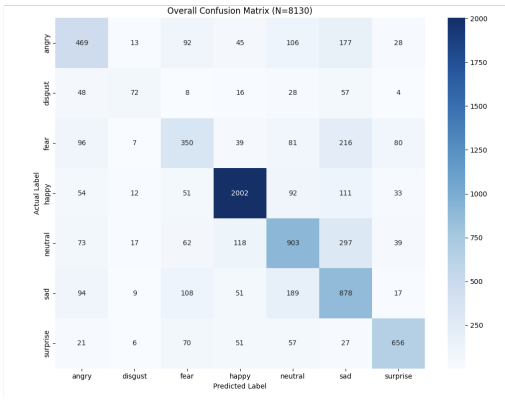
(a) CNN$_{48px}$ baseline (59.0% acc).



(b) ResNet-50 vanilla (60.0% acc).



(c) ResNet-50 + RW + FL (67.0% acc).



(d) ResNet-50 + Adv. Debiasing (66.0% acc).

or adversarial debiasing. This could be due to several factors: (1) high visual similarity in the muscle activations for these emotions (e.g., widened eyes, open mouth) making them inherently confusable from static images, and (2) potential ambiguities or inconsistencies in dataset labeling, as these emotions can co-occur or transition rapidly in real-world expressions.

- **Anger under-detection in minority groups.** In the Latino/Hispanic subgroup, the baseline ResNet-50 produced 29 false negatives for *Anger*; after race-weighted + focal-loss training, this dropped to 25—a 13.8 % reduction. Similar—but smaller—drops occur for Black (FN 35→31, 11.4 %) and Asian (FN 12→10, 16.7 %) subgroups, indicating that reweighting helps the model better recognize under-represented angry faces.

8.7 Bias-Mitigation Effectiveness

Table 6 summarizes how each mitigation strategy impacts overall accuracy, subgroup gaps, and emotion-level performance. The "Gap Δ" is relative to the Vanilla ResNet-50's gap of 12.3 pp.

Table 6: Mitigation impacts on overall accuracy and worst–best subgroup F1 gap (per-race).

Model	Overall Acc	Gap (pp)	Gap Δ (pp)
ResNet-50 (vanilla)	60.0%	12.3	—
ResNet-50 + RW + Focal Loss	67.0%	11.3	−1.0
ResNet-50 + Adv. Debiasing	66.0%	9.4	−2.9

Key takeaways.

- **Adversarial Debiasing** delivers the largest fairness gain in terms of F1 gap reduction: it reduces the worst–best subgroup F1 gap to 9.4 pp (a 2.9 pp improvement over vanilla ResNet-50), while boosting overall accuracy by 6.0 pp compared to vanilla ResNet-50.
- **RW + Focal Loss** also significantly improves fairness, reducing the F1 gap by 1.0 pp to 11.3 pp, and achieves the highest overall accuracy at 67.0% (a 7.0 pp boost over vanilla).
- Both methods substantially improve minority-group F1 scores (e.g., Middle Eastern F1 improves from 53.6% in vanilla ResNet-50 to 62.6% with RW+FL and 60.7% with Adversarial Debiasing).
- Both mitigated models maintain high weighted F1 scores (0.65), demonstrating that fairness improvements need not sacrifice overall class-balanced performance.
- Persistent low recall on *Fear* (11–42 %) and *Disgust* (0–50%) across all races suggests a need for emotion-specific augmentation or loss re-balancing in future work.

8.8 Limitations and Future Work

- Numeric fairness metrics depend on inferred demographics; mis-inference by DeepFace propagates to reported gaps.
- CK+ contributes a small, lab-frontal subset; its clean signals may over-inflate accuracies – future runs should weight CK+ samples lower or leave them for out-of-domain testing.

- The scope of this project focused on racial bias; analysis of gender or age-based bias, or intersectional biases, was not performed but remains an important area for future investigation.
- Next steps: incorporate AffectNet, test intersectional slices (e.g., *older Black female*), and validate on a real-time webcam feed under varying illumination.

Take-away. Both in-processing re-weighting with focal loss and adversarial debiasing substantially improved fairness. Adversarial debiasing achieved the most significant reduction in the per-race F1 score gap (to 9.4 pp), while re-weighting with focal loss yielded the highest overall accuracy (67.0%) with a notable fairness improvement (11.3 pp F1 gap). These findings align with trends reported by Fan et al. and Suresh et al. showing the effectiveness of such in-processing techniques in recent FER fairness studies.

9 ETHICAL CONSIDERATIONS

This project utilized publicly available datasets (FER-2013, RAF-DB, CK+) for academic research purposes. Demographic attributes, specifically race, were inferred using the DeepFace API due to the lack of consistent ground-truth labels across these datasets. It is acknowledged that automated demographic inference tools like DeepFace are not perfectly accurate and may themselves exhibit biases, potentially misclassifying individuals or performing differentially across groups. Such misclassifications could influence the reported fairness metrics and the perceived effectiveness of bias mitigation techniques. While confidence thresholds were applied to filter low-certainty inferences, the potential for annotation bias from the inference tool remains a limitation. This work aims to explore bias mitigation techniques within these constraints, but the ethical implications of deploying FER systems, particularly those relying on inferred sensitive attributes, must be carefully considered. Real-world applications would require robust consent mechanisms, transparent data handling practices, and thorough validation to prevent harm and ensure equitable outcomes.

10 FUTURE WORK IN BIAS-MITIGATED FER

Despite progress, several challenges remain for achieving truly fair and unbiased FER systems. Key directions for future work include:

- **Addressing Intersectional Bias:** Current research often tackles bias one attribute at a time (e.g., race). However, intersectional groups (such as older women of color) can experience compounded biases. Future FER systems should be evaluated on these intersections, necessitating the collection or annotation of datasets that adequately represent such subgroups. Analyzing intersectional performance using the inferred demographics is a first step. Novel reweighting methods or fairness constraints that account for multiple protected attributes simultaneously are largely unexplored and represent a significant opportunity for future research [7].
- **Balancing Accuracy and Fairness Trade-offs:** Increasing fairness frequently comes at the expense of overall accuracy. Research is needed to develop training methods that minimize this trade-off. Multi-objective optimization techniques that simultaneously maximize classification accuracy while minimizing bias metrics (like DPD or EOD) are promising, as are approaches such as fairness-aware model calibration or causal inference methods to disentangle task-relevant features from bias-related features. The goal is to embed fairness into FER models without a significant degradation in performance [18, 19].
- **Standardized Fairness Benchmarks and Evaluation:** Unlike object recognition, FER currently lacks agreed-upon benchmarks for assessing bias and fairness. The establishment of standardized evaluation protocols—including balanced benchmark datasets with reliable demographic labels (or robust inference methods) and common fairness metrics (e.g., true positive rate parity, equalized odds)—would facilitate more reliable comparisons across methods. A dedicated fairness evaluation framework for FER, potentially inspired by existing toolkits like Fairlearn, could drive progress in this field [4, 6].
- **Scalability to Real-World Conditions:** Many bias mitigation techniques have been validated on relatively small or controlled FER datasets. A pressing open question is how these techniques scale to real-world systems that process streaming video and diverse, uncontrolled inputs. Future work should explore continual and federated learning approaches to ensure that fairness holds as data evolves over time, as well as automated bias detection and monitoring in large-scale FER deployments [9].

By pursuing these avenues—addressing intersectional bias, refining accuracy- fairness trade-offs, standardizing fairness evaluation, and ensuring real-world scalability—future research can help bridge the gap between academic FER models and equitable, deployable systems.

REFERENCES

[1] Md. Rashedul Akhand, Md. Kamal Hossain, Md. Shorif Uddin, and Jae-Young Kim. 2021. Facial Expression Recognition Using Transfer Learning. *Electronics* 10, 9 (2021), 1036. https://doi.org/10.3390/electronics10091036

[2] Mohsin Alvi, Andrew Zisserman, and Sendhil Mullainathan. 2018. Turning a Blind Eye: Explicit Removal of Biases from Deep Neural Network Embeddings. In *Workshop on Human-Centric Machine Learning, ECCV*. https://openaccess.thecvf.com/content_ECCVW_2018/papers/11133/Alvi_Turning_a_Blind_Eye_Explicit_Removal_of_Biases_from_Deep_ECCVW_2018_paper.pdf

[3] Gary Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* 25, 11 (2000), 120–123.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[5] Yunliang Chen and Jungseock Joo. 2021. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14980–14990. https://doi.org/10.1109/ICCV48922.2021.01474

[6] Joohee Cheong, Sinan Kalkan, and Hatice Gunes. 2021. The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing: Overview and Techniques. In *IEEE Signal Processing Magazine*, Vol. 38. 39–49. https://doi.org/10.1109/MSP.2021.3101539

[7] Neil Churamani, Praateek Perera, Carlos Martinho, and Subhasis Chaudhuri. 2020. Fairness in Machine Learning for Affect Recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 146–155. https://doi.org/10.1145/3382507.3418866

[8] Alex Fan, Xingshuo Xiao, and Peter Washington. 2023. Addressing Racial Bias in Facial Emotion Recognition. *arXiv preprint arXiv:2308.04674* (2023). https://arxiv.org/abs/2308.04674

[9] Lisa Fromberg, Tobias Nielsen, Florin D. Frumosu, and Line K. H. Clemmensen. 2024. Beyond Accuracy: Fairness, Scalability, and Uncertainty Considerations in

Facial Emotion Recognition. In *Proceedings of the NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response*. PMLR. https://openreview.net/forum?id=h9S3417WvT

[10] Gustavo A. A. Galán, Pedro Rivas, and Robert J. Marks. 2023. Mitigating Algorithmic Bias on Facial Expression Recognition. In *arXiv preprint arXiv:2312.15307*. https://arxiv.org/abs/2312.15307

[11] Mohammad M. Hosseini, Amirhossein P. Fard, and Mohammad H. Mahoor. 2025. Faces of Fairness: Examining Bias in Facial Expression Recognition Datasets and Models. *arXiv preprint arXiv:2502.11049* (2025). arXiv:2502.11049 [cs.CV]

[12] Mohammad Kolahdouzi and Ali Etemad. 2023. Toward Fair Facial Expression Recognition with Improved Distribution Alignment. In *Proceedings of the 2023 International Conference on Multimodal Interaction*. https://arxiv.org/abs/2308.07236

[13] Shan Li and Weihong Deng. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 11 (2020), 2873–2893. https://doi.org/10.1109/TPAMI.2019.2924567

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[15] Martina Mattioli and Federico Cabitza. 2024. Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology. *Machine Learning and Knowledge Extraction* 6, 4 (2024), 2555–2663. https://doi.org/10.3390/make6040109

[16] Sefik Ilkin Serengil and Alper Ozpinar. 2020. DeepFace: A Lightweight Face Recognition and Facial Attribute Analysis (Age, Gender, Emotion and Race) Framework for Python. https://github.com/serengil/deepface. Accessed: 2025-05-09.

[17] Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2024. Are Bias Mitigation Techniques for Deep Learning Effective? *arXiv preprint arXiv:2104.00170* (2024). https://arxiv.org/abs/2104.00170

[18] Palak Singhal, Shreya Gokhale, Aniket Shah, Deepak Kumar Jain, Rahee Walambe, Aniko Ekart, and Ketan Kotecha. 2025. Domain adaptation for bias mitigation in affective computing: use cases for facial emotion recognition and sentiment analysis systems. *Discover Applied Sciences* 7, 7 (2025), 229. https://doi.org/10.1007/s42452-025-06659-1

[19] Vighnesh Suresh and Desmond C. Ong. 2022. Using Positive Matching Contrastive Loss with Facial Action Units to Mitigate Bias in Facial Expression Recognition. In *Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7. https://doi.org/10.1109/ACII55715.2022.10051876