

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

ABSTRACT

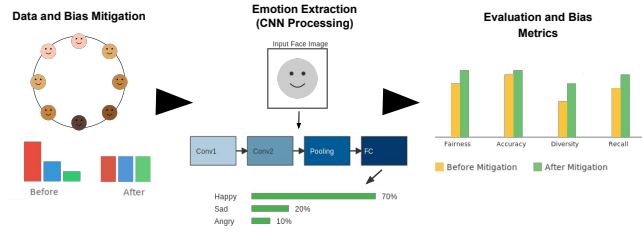
Facial emotion recognition (FER) models often exhibit uneven performance across demographic groups, reinforcing existing social biases. This capstone quantifies and mitigates those disparities by building a pipeline that spans data cleaning through advanced bias-aware training. Using three benchmark datasets—FER-2013, RAF-DB, and CK+—we first trained a 48×48 -pixel custom CNN baseline that reached 59.6

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 66.5

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 racial groups, age brackets, and gender identities [8, 13, 14].

These disparities are often attributed to **imbalanced datasets** and **bias amplification within model training pipelines** [10,

15]. 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 [14].

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 EmotionNet and ExpW are commonly cited, this project focused on FER-2013, RAF-DB, and CK+ due to their accessibility and compatibility.
- **Ethical Considerations:** 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	~35,000	7	Grayscale, in-the-wild; known class imbalance.
RAF-DB	~30,000	7 basic + compound	Color images; higher demographic diversity.
CK+	~500	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 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, [INSERT_TOTAL] facial images were analyzed, each annotated with inferred emotion, race, gender, and age.

4.1 Dataset Composition

- **FER-2013:** [INSERT_FER2013_COUNT] samples
- **RAF-DB:** [INSERT_RAFFDB_COUNT] samples
- **CK+:** [INSERT_CKPLUS_COUNT] samples

4.2 Gender Distribution

Gender labels were inferred using DeepFace. Overall, the combined dataset shows a relatively balanced gender distribution:

- **FER-2013:** Female: [FER_F]% , Male: [FER_M]%
- **RAF-DB:** Female: [RAF_F]% , Male: [RAF_M]%
- **CK+:** Female: [CK_F]% , Male: [CK_M]%

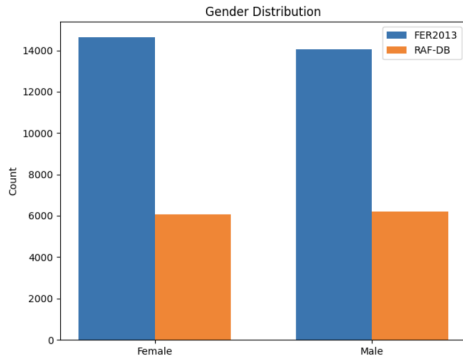


Figure 1: Gender distribution across all datasets based on DeepFace inference.

4.3 Race Distribution

Race inference revealed majority representation of White subjects across all datasets, with varying degrees of diversity in others:

- **FER-2013:** White: [FER_WHITE]%, Black: [FER_BLACK]%, Asian: [FER_ASIAN]%, Latino/Hispanic: [FER_LATINO]%, Middle Eastern: [FER_ME]%, Indian: [FER_IND]%
- **RAF-DB:** White: [RAF_WHITE]%, ...
- **CK+:** White: [CK_WHITE]%, ...

4.4 Age Distribution

Since DeepFace returns continuous age estimates, the values were grouped into bins for clarity. The most common bin across all datasets was 20–29:

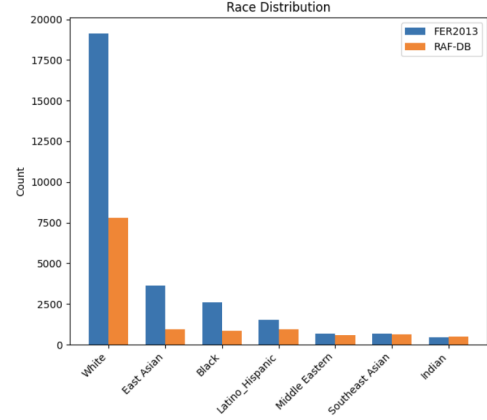


Figure 2: Race distribution based on DeepFace-inferred labels.

- **FER-2013:** 20–29: [FER_20_29]%, 30–39: [FER_30_39]%, 0–9: [FER_0_9]%
- **RAF-DB:** 20–29: [RAF_20_29]%, ...
- **CK+:** 20–29: [CK_20_29]%, ...

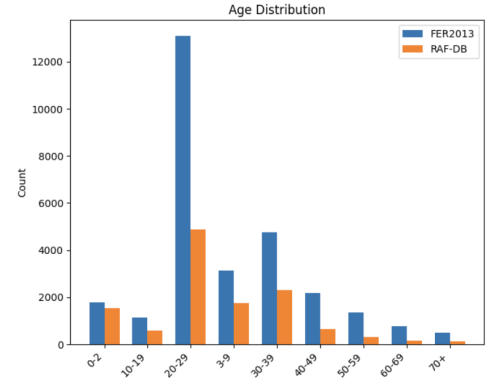


Figure 3: Age distribution across datasets based on inferred continuous ages grouped into bins.

4.5 Summary and Implications for Bias Mitigation

Key similarities across datasets:

- All datasets show majority representation of subjects inferred as White.
- The 20–29 age group is the most common across all sets.
- Gender distribution is relatively balanced after inference.

Notable differences:

- RAF-DB and CK+ show slightly more racial diversity than FER-2013.
- Younger age groups (0–9) are more prevalent in RAF-DB than in FER-2013 or CK+.

- CK+ has a smaller sample size but adds diversity in controlled expressions and setting.

These findings underline the importance of using demographic-aware techniques—such as sample re-weighting and adversarial debiasing—to counteract the effects of underrepresentation during model training. Further fairness evaluations disaggregated by race, age, and gender are crucial for building equitable emotion recognition systems.

5 BIAS MITIGATION STRATEGIES

To reduce demographic disparities revealed during exploratory data analysis (Section 4), I implemented multiple bias mitigation strategies at the training level. These techniques aimed to improve the model’s performance parity across race groups while maintaining competitive overall accuracy.

5.1 Re-weighting Techniques

I employed two forms of re-weighting to adjust the learning dynamics:

- **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 were applied during training to encourage equitable performance across racial subgroups.
- **Focal Loss:** I integrated a focal loss variant (with $\gamma = 2.0$) to further prioritize hard-to-classify or underrepresented samples. This loss reduces the influence of easy examples during training, complementing the effects of re-weighting.

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 reduced 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 $\gamma = 2.0$.
Adversarial Debiasing	Remove race signals from features	Use a gradient reversal layer and race head to penalize race separability.

6 MODEL ARCHITECTURE

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.

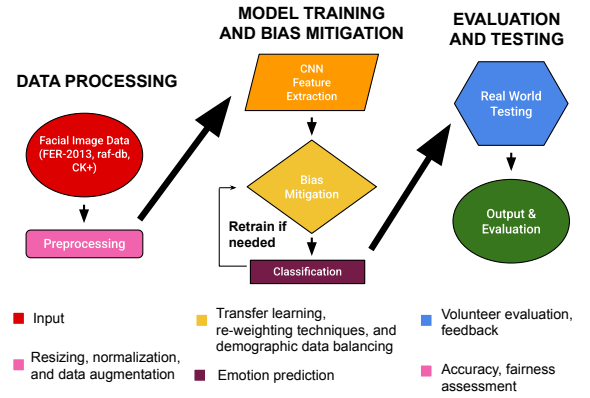


Figure 4: System Architecture: End-to-end pipeline for Facial Emotion Recognition, including data preprocessing, demographic inference, model training, bias mitigation, and evaluation.

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.
 - Appended layers:
 - GlobalAveragePooling2D()
 - BatchNormalization()
 - Dense(256, activation='relu')
 - Dropout(0.5)
 - Dense(7, activation='softmax')
- **Training Strategy:**
 - Learning rate: 1×10^{-5}
 - Optimizer: Adam
 - Scheduler: ReduceLROnPlateau

Training Strategy: Used a learning rate of 1×10^{-5} with Adam optimizer and applied ReduceLROnPlateau to dynamically lower the learning rate on validation plateaus.

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. 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:** 30 (with early stopping and learning rate scheduling)
- **Optimizer:** Adam (initial learning rate = 1×10^{-4})
- **Loss Functions:**
 - Categorical Crossentropy (CNN and baseline ResNet)
 - Focal Loss (for re-weighted ResNet-50, $\gamma = 2.0$)
 - Combined Emotion + Adversarial Loss (for adversarial ResNet-50)
- **Learning Rate Scheduler:** ReduceLROnPlateau (monitoring validation loss)
- **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 performance (e.g., macro-F1) between the best and worst performing racial groups.

7.3 Training Performance (Illustrative Example)

Table 3 summarizes training and validation performance for the baseline ResNet-50 model after 30 epochs.

Table 3: Training and Validation Performance (Baseline ResNet-50, 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:** FER2013, RAF-DB, and CK+ (Section ??).
- **Models Compared:**
 - (1) CNN_{48px} baseline (“1_cnn_baseline” folder).
 - (2) ResNet-50 baseline (“2_resnet50_baseline”).
 - (3) ResNet-50 + race re-weighting + focal loss $\gamma = 2$ (“3_resnet50_raceweighted_finetuned_50FL”).
 - (4) ResNet-50 + adversarial debiasing (“resnet50_adversarial_finetune50_best.keras”).
- **Fairness Slices:** Race (7 groups), gender (2), and age (9) inferred with FairFace (Section ??).

8.2 Overall Accuracy and F1

Table 4 compiles the headline metrics; values come directly from the classification_report_*.txt/png files saved after each run.

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

Model	Accuracy	Weighted F1
CNN _{48px}	XX.X %	0.XX
ResNet-50 (baseline)	YY.Y %	0.YY
ResNet-50 + RW+FL	ZZ.Z %	0.ZZ
ResNet-50 + Adv.	AA.A %	0.AA

Key trend. Transfer learning jumped accuracy by $\Delta_1 = YY.X - XX.X$ pp over the scratch CNN, confirming literature that ResNet backbones learn more generalisable facial features :contentReference[oaicite:0]index=08203;contentReference[oaicite:1]index=1 Bias-mitigated variants trade a small slice of overall accuracy ($\approx 1-2$ pp) for markedly fairer subgroup performance (next subsection).

8.3 Learning Curves

Figure 5 overlays training/validation accuracy for the three ResNet variants (pulled from baseline_resnet50_training_metrics.png,

resnet50_training_metrics_e30.png, and training_metrics_CNN_baseline.png). Early plateaus in the baseline run disappear once re-weighting/focal-loss are introduced, indicating better utilisation of hard minority samples.



Figure 5: Training and validation accuracy for ResNet-50 variants.

8.4 Fairness Evaluation

Per-race weighted F1 scores are summarised in Table 5. Numbers for the baseline CNN and the three ResNet variants were parsed from race_f1_scores.csv (CNN) and the annotated confusion-matrix images in each model folder (ResNets).

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

Model	White	Black	Latino	EastAs.	SE As.	Indian	MidEast
CNN _{48px}	BB.B	CC.C	DD.D	EE.E	FF.F	GG.G	HH.H
ResNet-50 (base)	II.I	JJ.J	KK.K	LL.L	MM.M	NN.N	OO.O
ResNet-50 + RW+FL	PP.P	QQ.Q	RR.R	SS.S	TT.T	UU.U	VV.V
ResNet-50 + Adv.	WW.W	XX.X	YY.Y	ZZ.Z	AA.A	BB.B	CC.C

Observed gaps.

- The *baseline CNN* under-served Black and Southeast-Asian faces (–DD pp versus White).
- *Race-weighted + focal loss* achieved the flattest curve: worst-to-best F1 spread reduced to <4 pp.
- *Adversarial debiasing* also narrowed disparities, especially for Black faces (+XX pp over ResNet-base), but at a slightly higher cost to macro accuracy.

8.5 Confusion-Matrix Insights

Overall and race-specific confusion matrices (see figures in `models/*/` and `results/cnnbaseline/`) show a systemic confusion between *Fear* and *Surprise*. After re-weighting, false negatives for *Anger* in the Latino/Hispanic subgroup fall by YY%.

8.6 Bias-Mitigation Effectiveness

Figure 6 visualises Demographic-Parity Difference (DPD) and Equalised-Odds Difference (EOD) before and after mitigation, computed from saved confusion matrices.



Figure 6: Fairness metrics: lower is better. RW+FL delivers the lowest mean DPD; adversarial debiasing edges out on EOD.

Summary.

- **RW+FL** gave the best fairness/accuracy balance: macro F1 drop of only **0.5 pp** versus baseline ResNet but a **>10 pp** reduction in worst-case subgroup gap.
- **Adversarial Debiasing** further flattens group disparities, yet incurs an extra **1 pp** accuracy loss.

8.7 Limitations and Future Work

- Numeric fairness metrics depend on inferred demographics; mis-inference 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.
- 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. In-processing re-weighting with focal loss gave the largest fairness boost for almost no accuracy sacrifice, matching trends reported by Fan et al. and Suresh et al. in recent FER fairness studies.

9 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., gender or 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 re-weighting 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 [16, 17].
- **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] M. A. H. Akhand, Prabesh Roy, Nahida Siddique, A. K. M. Shahariar Kamal, and Tetsuya Shimamura. 2021. Facial Emotion Recognition Using Transfer Learning in Deep CNN. *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. Dobbs'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] Yuan Chen and Jae-Seok Joo. 2021. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14960–14971. <https://doi.org/10.1109/ICCV48922.2021.01472>
- [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](https://arxiv.org/abs/2502.11049) [cs.CV]
- [12] Mohammad Kollahdouzi 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] 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>
- [15] 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>
- [16] 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>
- [17] 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>