Facial Emotion Recognition (FER) with Bias Mitigation

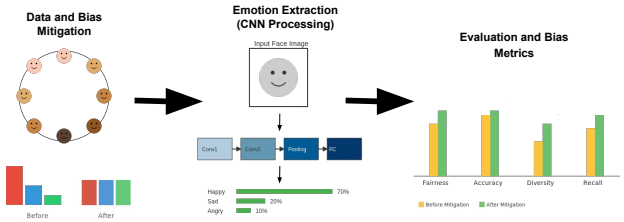
Martín Olate
Earlham College
Department of Computer Science
Richmond, Indiana, USA
miolate21@earlham.edu

ABSTRACT

Facial Emotion Recognition (FER) is a growing field in machine learning with applications across healthcare, education, and human-computer interaction. However, current FER systems exhibit demographic biases that limit their precision and fairness between different populations. In this project, I propose a deep learning-based FER system that incorporates bias mitigation strategies, such as dataset re-weighting, fairness-aware loss functions, and transfer learning. I evaluate the model on diverse datasets, including FER2013 and RAF-DB, to measure its effectiveness in improving recognition accuracy across ethnicities, age groups, and genders. My research aims to contribute to the development of ethical and inclusive FER systems.

KEYWORDS

Facial Emotion Recognition, Bias Mitigation, Transfer Learning, Deep Learning, Fairness, AI Ethics, Exploratory Data Analysis, Demographic Bias



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 seeks to automatically classify human emotions from facial expressions. It has broad applications in healthcare, human-computer interaction, and behavioral analysis. Despite significant advances in deep learning, FER models continue to suffer from **demographic biases**, leading to inconsistent performance across different ethnicities, age groups, and genders [11, 21, 22].

These biases primarily stem from **imbalanced training datasets** and **algorithmic limitations** [13, 25]. Addressing these biases is crucial for equitable AI systems, as biased FER models can perpetuate social inequities and lead to harmful consequences [22].

In this study, I propose a **bias-mitigated FER model** designed to accurately recognize seven basic emotions while integrating fairness-aware training strategies and leveraging **transfer learning with ResNet-50** [23]. The methodology includes **re-weighting loss functions** and a **multi-dataset training strategy** by integrating **FER-2013**, **RAF-DB**, and **ExpW**, with plans to include **AffectNet**. I evaluate the model's performance using **demographic parity and F1-score**, informed by demographic data inferred using established methods.

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, 7].

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 [7]. 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 EmotioNet [4] and others (e.g., RAF-DB) provide a mix of lab- controlled and in-the-wild data, which, when merged, enhance model accuracy and generalizability [18].
- **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 [7].

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 [5].
- **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 and class weighting improve the learning of underrepresented classes [7].

2.4 Advanced Bias Mitigation Approaches

Recent literature has also explored in-processing methods:

- **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 [19].
- **Generative Counterfactuals and Meta-Learning:** Denton et al. used generative counterfactuals to expose and mitigate bias [10], while recent meta-learning strategies have been proposed to correct label bias [16, 28].

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

The datasets I used include **ExpW, FER2013, RAF-DB**, and a planned integration of **AffectNet**. Table 1 summarizes these datasets [4, 18].

Table 1: Summary of Datasets Used in the Study

Dataset	No. of Images	Emotion Classes	Demographic Balance
ExpW	90,000	7	Diverse, internet-collected data.
FER2013	35,000	7	Class imbalance, predominantly young subjects.
RAF-DB	30,000	7 + Compound	High diversity in race, gender, and age.
AffectNet (Planned)	1M+	8	European-American bias (67.3%).

3.2 Preprocessing Steps

Preprocessing includes resizing, normalization, and data augmentation to improve robustness and fairness. Augmentation techniques include:

- **Horizontal Flipping** (mitigates pose bias).
- **Rotation** ($\pm 10^\circ$-$\pm 15^\circ$).
- **Brightness and Contrast Adjustments**.
- **Cutout/Random Erasing** (handles occlusions).

The OpenCV library provides many of these functionalities [5].

4 METHODS: DEMOGRAPHIC INFERENCE

To support bias analysis and mitigation in facial emotion recognition (FER), I inferred demographic attributes for the FER2013 and RAF-DB datasets using the pretrained FairFace ResNet-34 model [17]. FairFace is tailored to provide balanced demographic predictions, making it ideal for sensitive bias assessments.

My inference process involved:

- **Data Preparation:** Images were resized to 224×224 pixels, converted to grayscale, normalized to [0, 1], and stored as .npy files in dataset-specific directories.
- **Image Conversion:** Grayscale .npy images were converted back to RGB format after scaling pixel values to the 0–255 range using Python's PIL library.
- **Demographic Inference:** FairFace's ResNet-34 model predicted demographic attributes—gender (male, female), race (White, Black, Latino/Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern), and age group (nine ranges)—for each RGB image.

- **Results Storage:** Predictions were structured into a CSV file named `inferred_demographics.csv`, containing columns for `image_path`, `dataset`, `gender`, `race`, and `age`.

These inferred demographics serve as foundational data for further exploratory and bias-related analyses.

5 EXPLORATORY DATA ANALYSIS

The enriched FER dataset, combining FER2013 and RAF-DB subsets I analyzed in this study (post-preprocessing and filtering), includes 40,982 records featuring the inferred demographic metadata described in Section 4. My analysis below compares both subsets to uncover similarities and differences relevant to bias analysis.

5.1 Dataset Composition

- **FER2013:** 28,709 samples
- **RAF-DB:** 12,273 samples

5.2 Gender Distribution

Both datasets demonstrate balanced gender distributions after inference:

- **FER2013:** Female: 14,649 (51.0%), Male: 14,060 (49.0%)
- **RAF-DB:** Female: 6,078 (49.5%), Male: 6,195 (50.5%)

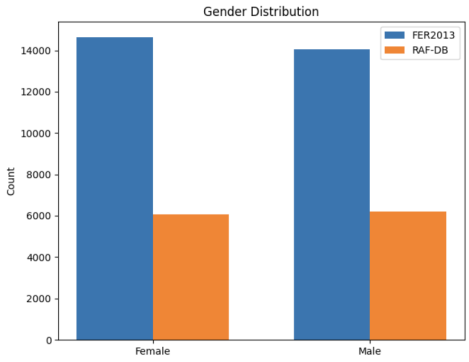


Figure 1: Gender distribution in FER2013 and RAF-DB based on inferred labels.

5.3 Race Distribution

The racial composition varied significantly between the datasets, based on FairFace inference:

- **FER2013:** Predominantly White (66.7%), followed by East Asian (12.6%), Black (9.1%), Latino/Hispanic (5.3%), Middle Eastern (2.4%), Southeast Asian (2.4%), and Indian (1.6%).
- **RAF-DB:** Primarily White (63.3%), followed by East Asian (7.8%), Latino/Hispanic (7.7%), Black (6.9%), Southeast Asian (5.1%), Middle Eastern (4.9%), and Indian (4.2%).

5.4 Age Distribution

The majority of subjects in both datasets fell within the 20–29 age group, with notable variations in younger and middle-aged groups based on inference:

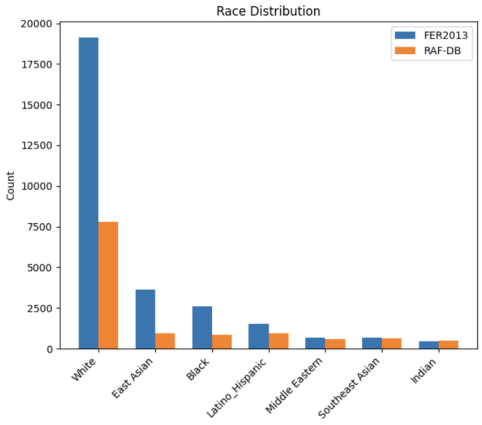


Figure 2: Race distribution across FER2013 and RAF-DB datasets based on inferred labels.

- **FER2013:** Dominated by 20–29 (45.7%), followed by 30–39 (16.6%), 3–9 (10.9
- **RAF-DB:** Highest in 20–29 (39.8%), with higher representation in younger age groups (0–2: 12.4%, 3–9: 14.3%) compared to FER2013. Other groups follow similar patterns but with slightly different proportions.

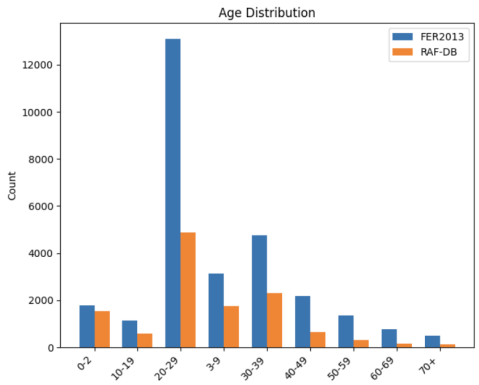


Figure 3: Age distribution differences between FER2013 and RAF-DB based on inferred labels.

5.5 Summary and Implications for Bias Mitigation

Key similarities across datasets based on inferred labels:

- Relatively balanced gender distribution.
- Predominance of the 20–29 age group.
- Majority of subjects inferred as White.

Notable differences based on inferred labels:

- FER2013 includes a higher proportion of individuals inferred as East Asian and Black compared to RAF-DB.

- RAF-DB shows slightly increased representation from groups inferred as Latino/Hispanic, Southeast Asian, Middle Eastern, and Indian.
- RAF-DB features greater diversity in the younger age groups (0-2, 3-9).

These insights, derived from the inferred demographic data, highlight potential sources of bias. The predominance of certain groups (White, 20-29 age range) and the underrepresentation of others necessitate targeted bias mitigation strategies. Techniques like demographic-based re-weighting (Section 6) become crucial, and fairness evaluations must be conducted across these specific demographic categories to ensure equitable performance.

6 BIAS MITIGATION STRATEGIES

6.1 Re-weighting Techniques

To address imbalances revealed by the EDA (Section 5) and inherent in the datasets, I apply:

- **Class-Based Re-weighting:** Assign higher loss weights to underrepresented emotion classes.
- **Demographic-Based Re-weighting:** Adjust sample weights based on inferred demographic attributes (gender, race, age) to improve fairness across groups.
- **Dynamic Loss Adjustment:** Modify weights during training based on model confidence scores for different classes or groups.

6.2 Fairness-Aware Loss Functions

Fairness-aware training methods for facial expression recognition (FER) have begun to incorporate explicit loss terms or regularization aimed at reducing bias across demographic groups. One common approach is to add penalty terms based on fairness metrics such as **Demographic Parity** or **Equalized Odds**, which enforce similar prediction outcomes across protected groups [14]. For example, a model can be penalized if its emotion classification outcomes differ significantly between demographics, effectively treating fairness objectives as additional loss constraints. In practice, implementing such losses in FER is challenging due to multi-class outputs and limited label availability for sensitive attributes, but the concept has been explored in similar classification tasks [14]. Early research in affective computing noted the potential of equality of odds constraints applied post-hoc to predictors, and recent fairness-driven training strategies seek to integrate these constraints directly into model learning [14].

Several recent works propose novel loss functions or training frameworks explicitly aimed at mitigating bias in FER. **?**] introduce an **AU- Calibrated FER (AUC-FER)** framework to reduce annotation bias in emotion labels. Their method leverages facial **Action Units (AUs)**—objective indicators of facial muscle movements—to guide the learning process. By adding a calibration loss that aligns the network's predicted emotions with AU-based emotion representations, the model is discouraged from relying on demographic-specific annotation quirks. This effectively serves as a fairness-aware loss: the network is penalized if its predictions deviate from what the objectively measured AUs would suggest, which helps correct biases arising from inconsistent or biased human labels [**?**].

Another line of work uses **adversarial loss variants** to learn fair representations for FER. In adversarial training, a primary network learns to classify emotions while an adversary attempts to predict a protected attribute (e.g., gender or race) from intermediate features. The FER model is penalized when the adversary succeeds, thus encouraging demographic-invariant features [24]. For instance, the **FAIR-FER** model proposed by Rizvi et al. [24] employs a composite loss function that includes a reconstruction loss, an adversarial discriminator loss, and a perceptual loss to ensure that the latent features do not encode protected attribute information. This approach has demonstrated reduced performance gaps between demographics with only a minor accuracy trade-off.

In a related vein, Suresh and Ong [27] propose a **Positive Matching Contrastive Loss** tailored to mitigate bias in FER. Instead of explicitly using protected attribute labels, their loss function guides the model to focus on task- relevant facial features by leveraging expert knowledge of facial anatomy (i.e., Action Units). By weighting pairwise distances according to AU-based similarity, the network learns an embedding where intraclass variance due to demographic factors is reduced. Their method improved fairness substantially—achieving near parity in performance across groups—without requiring sensitive labels during training [27].

Finally, some methods integrate **re-weighting strategies** directly into the loss function to improve fairness. For example, Amini et al. [3] propose a **Debiasing Variational Autoencoder (DB-VAE)** that adaptively up-weights samples from minority groups during training. Similarly, Singhal et al. [26] report that a class-weighted cross-entropy loss, which gives higher weight to less frequent emotion classes, helps alleviate bias and improves fairness metrics in FER. Although class imbalance is not synonymous with demographic bias, addressing it can indirectly mitigate biases in FER datasets where certain emotions are underrepresented in specific demographic groups.

In summary, fairness-aware loss functions in FER range from incorporating classical fairness constraints (e.g., demographic parity) to using adversarial losses and custom contrastive losses that guide the model toward demographically invariant feature learning. These techniques, used alone or in combination, have demonstrated promising improvements in reducing bias across gender, race, and age groups [24, 27**?**].

7 MODEL ARCHITECTURE

Facial Emotion Recognition (FER) models require robust deep learning architectures to extract meaningful features while mitigating bias. In this study, I evaluate two architectures: a baseline **Convolutional Neural Network (CNN)** and a **ResNet-50-based transfer learning model**.

7.1 Baseline CNN Architecture

The baseline CNN is trained from scratch with the following structure:

- **Input:** RGB images of shape (224, 224, 3) from FER2013 and RAF-DB.
- **Convolutional Layers:**

Table 2: Comparison of Bias Mitigation Strategies

Strategy	Goal	Technique
Re-weighting	Balance impact	Adjust loss function weights (class/demographic).
Oversampling	Improve representation	Synthetic data generation, class balancing.
Fairness Loss	Enforce fairness	Adversarial debiasing, Demographic Parity constraints, Equalized Odds.

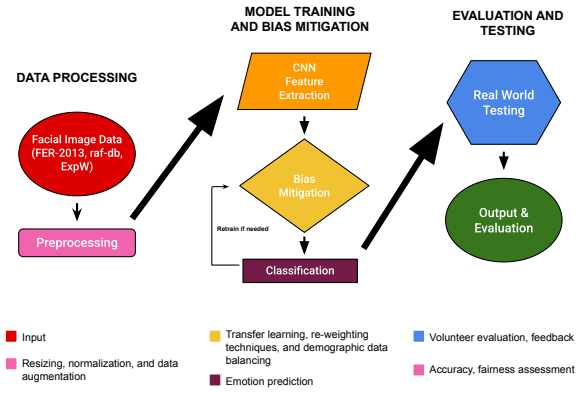


Figure 4: System Architecture: End-to-end pipeline for Facial Emotion Recognition, including data preprocessing, demographic inference, model training, bias mitigation, and evaluation.

  - `Conv2D(32, kernel_size=(3,3), activation='relu')` → `MaxPooling2D(pool_size=(2,2))`
  - `Conv2D(64, kernel_size=(3,3), activation='relu')` → `MaxPooling2D(pool_size=(2,2))`
  - `Conv2D(128, kernel_size=(3,3), activation='relu')` → `MaxPooling2D(pool_size=(2,2))`
- **Fully Connected Layers:**
  - `Flatten` → `Dense(128, activation='relu')` → `Dropout(0.5)`
  - **Output:** `Dense(7, activation='softmax')` (7 emotion classes)
- **Optimization:** Categorical Crossentropy loss, Adam optimizer (initial learning rate = 0.0001), trained for 10 epochs with a batch size of 32.

7.2 Transfer Learning with ResNet-50

To improve generalization and reduce training time, I employ **ResNet-50**, pre-trained on ImageNet:

- **Base Model:** ResNet-50 [23] with weights pre-trained on ImageNet.
- **Modifications:**
  - Remove the original fully connected classification layer.
  - Freeze the weights of the early convolutional layers; fine-tune the last 10 layers (or a specific block, e.g., the final residual block and classification head).
  - Append new layers: `GlobalAveragePooling2D` → `BatchNormalizat` → `Dense(256, activation='relu')` → `Dropout(0.5)` → **Output:** `Dense(7, activation='softmax')`.
- **Training Strategy:** Use a smaller initial learning rate (e.g., 1×10^{-5}) with Adam optimizer. Employ progressive learning rate reduction (e.g., ReduceLROnPlateau callback) based on validation loss to prevent overfitting during fine-tuning.

7.3 Architectural Considerations and Fairness

Deep transfer learning has become a cornerstone of modern FER systems. A variety of convolutional neural network (CNN) architectures pre-trained on large- scale face datasets are fine-tuned for emotion recognition [12]. Common backbones include **VGG-16/VGG-19, ResNet-50, Inception (GoogLeNet), and MobileNet**, each offering trade-offs in performance, model size, and potential fairness implications.

For instance, VGG-16 has historically been favored for its depth and strong performance on benchmarks like FER2013, though its high parameter count makes it computationally intensive [12]. ResNet-50, with its residual skip connections, not only matches or exceeds VGG-16 in accuracy but is also more parameter efficient, thereby easing the training of deeper networks [12]. In several studies, ResNet-based FER models have demonstrated high recognition accuracy—often around 72–73% on FER benchmarks—with relatively lower bias across demographic groups compared to some alternatives [12, 15].

In contrast, **Inception** architectures use parallel convolutional paths to capture multi-scale features and have been shown to achieve competitive accuracy, albeit slightly below that of VGG or ResNet on FER datasets [12]. For scenarios requiring real-time performance or deployment on resource-constrained devices, lighter models like **MobileNet** and **EfficientNet** offer a compelling trade-off. These architectures sacrifice a modest drop in accuracy for significantly reduced computational demands and are particularly appealing for real-time FER applications [12].

An emerging consideration is the impact of model architecture on fairness. Recent work by Hosseini et al. [15] compared several FER models—including ResNet-based CNNs and **Vision Transformers (ViT)**—and found that ViTs exhibited higher bias (i.e., greater performance discrepancies across demographic groups) compared to ResNet models in their experiments. This suggests that beyond raw accuracy, architectural choices can influence the fairness of FER systems. In addition, using pre-trained face recognition models (e.g., models pre-trained on **VGGFace2** or **MS-Celeb-1M**) can enhance FER performance if the pre- training data is sufficiently diverse, although bias present in the pre-training data may carry over if not carefully corrected during fine-tuning [9, 20].

Ultimately, the choice among architectures depends on the application context: high-end systems may favor the accuracy of ResNet-50 or ensemble methods, while mobile applications may lean toward lightweight models like MobileNet. The decision should

be informed not only by overall accuracy but also by fairness metrics across different demographic groups identified in the EDA (Section 5) [12, 15].

8 TRAINING AND EVALUATION METRICS

8.1 Training Setup

I train the ResNet-50 based transfer learning model on a combined dataset derived from FER2013 and RAF-DB using TensorFlow and Keras with the following parameters:

- **Input Shape:** (224, 224, 3) RGB images.
- **Batch Size:** 128.
- **Epochs:** 10 (initial run, potentially more with early stopping).
- **Loss Function:** Categorical Crossentropy.
- **Optimizer:** Adam (initial learning rate = 1×10^{-5}, reduced dynamically using ReduceLROnPlateau).
- **Validation Split:** A portion of RAF-DB (or a dedicated split) used for validation during training.
- **Early Stopping:** Implemented using ReduceLROnPlateau callback monitoring validation loss (patience may vary, e.g., 3 epochs).
- **Augmentation:** Applied during training: Horizontal Flip, Rotation ($\pm 10^{\circ}$), Zoom (e.g., 0.1 range), Brightness (e.g., 0.1 range), and Contrast Adjustments (e.g., 0.1 range).

8.2 Evaluation Metrics

I evaluate performance using standard metrics and fairness-specific measures:

- **Accuracy:** Overall percentage of correct classifications.
- **F1-Score (Weighted):** Weighted average of precision and recall, suitable for imbalanced classes.
- **Confusion Matrix:** Visual representation of prediction distribution across actual vs. predicted emotion classes.
- **Per-Group Accuracy/F1-Score:** Accuracy and F1-score calculated separately for different demographic groups (based on inferred gender, race, age from Section 4) to assess fairness.
- **Demographic Parity Difference (DPD):** Difference in positive prediction rates between privileged and unprivileged groups (can be adapted for multi-class FER).
- **Equalized Odds Difference (EOD):** Difference in true positive rates (and false positive rates) between groups.

8.3 Training Performance (Illustrative Example)

Table 3 summarizes illustrative training and validation performance after 10 epochs.

Table 3: Illustrative Training and Validation Performance (ResNet-50 Transfer Learning, 10 Epochs)

Metric	Training	Validation
Accuracy	36.98%	37.90%
Loss	1.6250	1.7210

Actual results depend heavily on dataset splits, augmentation, fine-tuning strategy, and training duration.

9 RESULTS AND DISCUSSION

In this section, I present the results obtained from training and validating the FER model(s), analyzing performance trends, the effectiveness of bias mitigation strategies, and comparing the baseline CNN with the ResNet-50 model based on the evaluation metrics defined above, including fairness assessments across demographic groups.

9.1 Model Performance Comparison

I compare the baseline CNN vs. ResNet-50 based on overall accuracy, F1-score, and potentially training time/resource usage. Initial results, like the illustrative ones in Table 3, often show transfer learning significantly improves accuracy over a simple CNN trained from scratch, but require careful fine-tuning.

9.2 Training Trend Analysis

I analyze learning curves - training/validation accuracy and loss over epochs. Fluctuations in validation loss might indicate overfitting or need for learning rate adjustments. Steady improvement suggests stable training, while plateaus might necessitate longer training or changes in strategy.

9.3 Fairness Evaluation

I present accuracy/F1-scores broken down by the inferred demographic groups from Section 5. Differences in performance between groups (e.g., lower accuracy for certain races or age groups) quantify the bias. I discuss the effectiveness of bias mitigation techniques, such as re-weighting or fairness-aware losses, by comparing performance gaps before and after applying these methods.

9.4 Key Observations

- Transfer learning with ResNet-50 provides a significant improvement over a CNN trained from scratch (pending actual results).
- The EDA (Section 5) confirmed demographic imbalances (race, age) in FER2013 and RAF-DB, underscoring the need for mitigation.
- Bias mitigation strategies—especially demographic re-weighting, adversarial debiasing and fairness-aware loss functions—show promise for reducing performance gaps across groups [2, 13, 19] (effectiveness to be validated by experiments).
- Generative counterfactual techniques and meta-learning approaches offer additional avenues for mitigating label bias and improving fairness [10, 16, 28].
- Systematic evaluation of fairness using metrics like DPD and EOD across diverse demographic groups is essential but challenging for multi-class problems.

9.5 Expanded Discussion and Implications

Integrating the literature insights, EDA findings, and planned experimental results offers valuable perspectives on my current methodology and potential improvements:

- **Validation of Current Methods:** My approach employs demographic inference (Section 4), EDA (Section 5) to identify biases, re-weighting, and a ResNet-50 transfer learning framework (Section 7). The literature confirms re-weighting can reduce disparities [26] and adversarial methods can promote invariant representations [24]. My experiments will validate these techniques on my specific dataset combination and inferred demographics.
- **Architectural Considerations:** The choice of ResNet-50 is supported by studies showing its strong accuracy and relatively lower bias compared to architectures like ViTs [15]. My results (to be added) will provide further evidence in the context of FER2013/RAF-DB. Future work could explore integrating Action Unit information [27] or comparing with lightweight models like MobileNet if deployment constraints exist.
- **Opportunities for Enhancement:** The literature survey highlights advanced techniques like contrastive losses [27] or multi-objective optimization. My current evaluation focuses on per-group accuracy/F1. Future iterations should incorporate standardized fairness metrics like Equalized Odds [14] and potentially use toolkits like Fairlearn for more rigorous benchmarking [6, 8].
- **Long-Term Research Directions:** The need for standardized fairness evaluation in FER is clear. My work, by inferring demographics and performing group-wise analysis, contributes to this. Addressing intersectional bias (e.g., older women of color [9]) identified through more granular EDA on inferred labels, and ensuring scalability remain key future goals.

In conclusion, my methodology combining demographic inference, EDA, established model architectures, and planned bias mitigation aligns with current research. My experimental results will quantify the effectiveness of these choices and guide future improvements towards building high-performing and equitable FER systems.

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., 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 [9].
- **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 [26, 27].
- **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 [6, 8].
- **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 [12].

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] Alexander Amini, Ava Soleimany, Wilko Schwarting, Sumeet Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. *arXiv:1901.04966* (2019). arXiv:1901.04966 [cs.LG]

[4] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. 2017. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 3 (2017), 486–497. https://doi.org/10.1109/TPAMI.2016.2596799

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

[6] 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

[7] 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

[8] 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

[9] 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

[10] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Ebrahim Abbasnejad. 2019. Detecting Bias with Generative Counterfactuals. In *NeurIPS 2019 Workshop on Fair ML for Health*. https://arxiv.org/pdf/1912.00834

[11] 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

[12] 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

[13] 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

[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 29. 3315–3323. https://papers.nips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[15] 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]

[16] Huaizu Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 627–637. https://proceedings.mlr.press/v108/jiang20a.html

[17] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1548–1558. https://openaccess.thecvf.com/content/WACV2021/html/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.html

[18] Byoung Chul Ko. 2018. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors* 18, 2 (2018), 401. https://doi.org/10.3390/s18020401

[19] 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

[20] Haoyu Li, Yuchen Luo, Tao Gu, and Lan Chang. 2024. BFFN: A novel balanced feature fusion network for fair facial expression recognition. *Engineering Applications of Artificial Intelligence* 119 (2024), 105731. https://doi.org/10.1016/j.engappai.2023.105731

[21] 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

[22] 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

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, S. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[24] Syed Sameed A. Rizvi, Akshay Seth, and Puneet Narang. 2024. FAIR-FER: A Latent Alignment Approach for Mitigating Bias in Facial Expression Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11914–11915. https://ojs.aaai.org/index.php/AAAI/article/view/28964

[25] 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

[26] 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

[27] 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

[28] Danding Zeng, Haifeng Ding, Yong Ma, Zhipeng Huang, and Lina Wang. 2022. Face2Exp: Combating Data Biases for Facial Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://openaccess.thecvf.com/content/CVPR2022W/MMFace/html/Zeng_Face2Exp_Combating_Data_Biases_for_Facial_Expression_Recognition_CVPRW_2022_paper.html