4ML3 Final Report: Fairness in Facial Emotion Recognition

Andre Menezes, Ahren Chen
{meneza3, chena125}@mcmaster.ca

1 Introduction

Facial emotion recognition (FER) refers to the automatic inference of human affective states—such as happiness, fear, anger, or disgust—from visual facial cues. It plays an important role in human–computer interaction, with applications in areas such as adaptive user interfaces, digital advertising, education, and healthcare monitoring (Fasel and Luettin, 2003), (Abdat et al., 2011). Over the past two decades, substantial progress has been made in recognizing facial expressions under controlled laboratory conditions, to the point that many such scenarios are considered largely solved (Sariyanidi et al., 2015). However, deploying FER systems "in the wild" remains challenging due to large intra-class variation (changes in pose, illumination, occlusion, and expression intensity) and subtle inter-class differences between emotions (Sariyanidi et al., 2015), (Mehendale, 2020).

With the advent of deep learning, Convolutional Neural Networks (CNNs) have become the dominant paradigm for FER, owing to their strong representation learning capabilities and computational efficiency (Jain et al., 2019), (Pramerdorfer and Kampel, 2016). Building on early successes in large-scale image recognition (Krizhevsky et al., 2017), a wide range of CNN-based FER architectures have been proposed, often tailored to handle unconstrained facial images captured in naturalistic environments (Jain et al., 2019), (Pramerdorfer and Kampel, 2016). Among the benchmark datasets used to compare such approaches, FER2013 has emerged as one of the most widely adopted. It contains 35,888 grayscale face images labeled with seven emotion categories and is designed to capture many of the difficulties present in real-world settings (Goodfellow et al., 2015). Human accuracy on FER2013 has been estimated to lie in the mid-60% range (Goodfellow et al., 2015), and a sequence of deep models has steadily pushed machine performance beyond this level (Pramerdorfer and Kampel, 2016), (Khaireddin and Chen, 2021).

In recent work, VGG-style CNN architectures combined with aggressive data augmentation and careful hyperparameter tuning have achieved state-of-the-art single-network accuracy on FER2013 (Khaireddin and Chen, 2021). Khaireddin and Chen, in particular, report a VGGNet-based model that reaches 73.28% test accuracy on FER2013 without using any additional training data, highlighting the impact of optimizer choice, learning rate scheduling, and fine-tuning strategies on overall performance (Khaireddin and Chen, 2021). Their results illustrate how methodical optimization of the training pipeline can yield significant gains over earlier CNN-based FER systems.

Despite these advances in accuracy, most FER2013 studies focus primarily on aggregate metrics such as overall test accuracy, with comparatively little attention to how performance may vary across demographic groups or other sensitive subpopulations. Recent work in facial expression recognition has begun to address this gap by explicitly quantifying how demographic imbalances and stereotypes in large-scale FER datasets transfer to trained models, revealing substantial demographic dependent disparities in recall for systems trained on AffectNet and FER+ (Dominguez-Catena et al., 2022). Follow-up work further proposes a family of demographic bias metrics and applies them across multiple FER datasets, showing that representational and stereotypical biases are both widespread at the dataset level and liable to propagate into the models (Dominguez-Catena et al., 2023). In the context of FER, these findings show that misclassification is not merely a matter of overall accuracy, but can differentially affect demographic groups, raising fairness and ethical concerns whenever such systems are used in real world settings.

In this work, we adopt an operational notion of fairness that is explicitly group-based. Concretely,

we say that an FER model is *fairer* when its performance is more similar across demographic groups defined by apparent race, gender, and age. We focus on outcome metrics such as accuracy and recall computed separately for each demographic group and emotion class, and we view large systematic gaps in these quantities as evidence of unfair treatment. In line with prior work on demographic bias in FER (Dominguez-Catena et al., 2022, 2023), our analysis is therefore concerned less with the absolute value of the overall accuracy and more with whether any particular race, gender, or age group consistently receives worse predictions than others.

In this paper, we build directly on the VGG-based FER2013 model of Khaireddin and Chen (Khaireddin and Chen, 2021) to examine these issues more closely. Our objectives are threefold. First, we provide a structured summary of their approach, clarifying the key design choices in data preprocessing, augmentation, architecture, and optimization that contribute to its performance. Second, we offer a critical analysis of their methodology and evaluation protocol, highlighting limitations and open questions that are not fully addressed in the original work. Third, and most importantly, we extend their evaluation framework to investigate model behavior across different demographic groups. By quantifying performance disparities and potential biases, we aim to complement existing accuracy-focused benchmarks with a fairness-aware perspective, showcasing FER2013 results not only in terms of how well the model performs overall, but also in terms of for whom it performs well for.

2 Summary of Findings

Khaireddin and Chen present one of the strongest single-network baselines on FER2013 to date, using a VGG-style CNN combined with an aggressively tuned training pipeline (Khaireddin and Chen, 2021). Their architecture consists of four convolutional stages with 3×3 filters and 2×2 max-pooling after each block, followed by three fully connected layers for classification. Compared to earlier VGG-based FER work, such as Pramerdorfer and Kampel (Pramerdorfer and Kampel, 2016), their variant includes an additional hidden fully connected layer and places dropout after the fully connected layers rather than after each convolutional block, increasing model capacity at the classification head. On the data side, they rely on an extensive augmentation pipeline: images are randomly scaled (up to $\pm 20\%$), shifted horizontally and vertically (up to $\pm 20\%$), and rotated (up to $\pm 10^{\circ}$), each transformation applied with probability 50%. The augmented images are then ten-cropped into 40×40 patches (four corners and center plus their mirrored counterparts), with random erasing applied to each crop with probability 50%, and pixel values normalized by division by 255. At test time, performance is reported using ten-crop averaging. Trained for 300 epochs with stochastic gradient descent with Nesterov momentum and weight decay, combined with a Reduce-on-Plateau learning rate scheduler and a subsequent cosine-annealing fine-tuning phase, their best model achieves 73.28% accuracy on the FER2013 public test set, surpassing the previously reported single-network VGG baseline of 72.7% (Pramerdorfer and Kampel, 2016; Goodfellow et al., 2015). Saliency-map visualizations further show that the network primarily attends to internal facial regions (eyes, eyebrows, mouth), while largely downweighting hair and background, and the confusion matrix indicates that "happiness" and "surprise" are classified most accurately, whereas "disgust" and "fear" remain the most challenging (Khaireddin and Chen, 2021).

Complementary to this accuracy-focused line of work, Dominguez-Catena et al. propose a metric-based framework to quantify demographic bias transfer from datasets to trained FER models (Dominguez-Catena et al., 2022). Their analysis, conducted on AffectNet and FER+, uses FairFace as a proxy demographic classifier to obtain apparent race and gender labels, and then defines three families of metrics: a representational bias metric based on the normalized standard deviation of demographic group proportions (capturing whether some groups are overrepresented or underrepresented); a pair of stereotypical bias metrics based on Normalized Mutual Information (NMI) and Normalized Pointwise Mutual Information (NPMI) between emotion labels and demographic groups (capturing label–demographic associations); and a model bias metric based on recall disparities across demographic groups for each class, summarized as an overall disparity score. Applied to AffectNet, these metrics reveal a strong representational imbalance toward subjects labeled as White, as well as stereotypical gender patterns such as an overrepresentation of men in the "angry" class and women in the "happy" class. Training VGG-style mod-

els on various balanced and artificially biased subsets, they find that balancing the dataset by gender substantially reduces gender-related disparity with only minor accuracy changes, whereas balancing by race does not significantly reduce race-related model bias (Dominguez-Catena et al., 2022). They also observe that increasing the total number of training examples tends to reduce measured bias scores, even when the underlying demographic distribution remains skewed.

In subsequent work, the same authors extend this analysis to a broader collection of FER datasets and refine their proposed metric suite (Dominguez-Catena et al., 2023). They systematically evaluate representational bias, global stereotypical bias, and local (class-specific) stereotypical bias across multiple FER corpora, showing that demographic skew and label–demographic associations are widespread and strongly dependent on the data source. Their experiments further suggest that many bias metrics are highly correlated, and they recommend a compact set of measures focusing on representational imbalance and global and local stereotypical bias as a practical basis for dataset auditing. Taken together, these findings provide evidence that popular FER benchmarks can encode substantial race and gender-related structure which, if unaddressed, is likely to propagate into models.

Overall, the literature we draw on provides two complementary perspectives. On the one hand, VGG-based CNNs with carefully engineered augmentation, optimization, and fine-tuning strategies can achieve state-of-the-art performance on FER2013 when evaluated solely in terms of overall accuracy (Pramerdorfer and Kampel, 2016; Khaireddin and Chen, 2021). On the other hand, recent work on demographic bias in FER datasets and models demonstrates that such aggregate metrics can mask systematic disparities across demographic groups, and that dataset composition and preprocessing choices play a central role in shaping these disparities (Dominguez-Catena et al., 2022, 2023). Our project is positioned at the intersection of these two strands: we take the high-performing FER2013 VGG architecture of Khaireddin and Chen as a starting point, and we adapt the bias-analysis methodology of Dominguez-Catena et al. to investigate how this model behaves across demographic subgroups, with the goal of jointly characterizing its accuracy and fairness.

3 Critiques

Our project relies primarily on FER2013 as the core benchmark and on the VGG-based architecture of Khaireddin and Chen (Khaireddin and Chen, 2021), so it is important to examine both the dataset and the reference model critically. FER2013 itself is designed as a challenging "in-the-wild" facial expression dataset, with low-resolution grayscale images collected from the web and labeled into seven basic emotion categories (Goodfellow et al., 2015). While this makes it suitable for stress-testing FER models, it also introduces several limitations. First, the dataset is known to contain noisy labels and ambiguous expressions, as annotations were obtained at scale and many images exhibit subtle or mixed emotions. Second, the class distribution is imbalanced, with some emotions (e.g., "disgust") being markedly underrepresented, which can bias models toward majority classes when trained with standard cross-entropy objectives. Third, FER2013 does not include any explicit demographic labels (e.g., age, race, or gender), which complicates fairness analysis: any demographic information must be inferred using external models, introducing additional noise and potential bias. Finally, FER2013 focuses on seven "basic" emotions and does not capture more nuanced or culturally specific affective expressions, limiting the scope of conclusions about real-world emotional communication.

The reference model of Khaireddin and Chen (Khaireddin and Chen, 2021) also has several methodological limitations when viewed from the perspective of reproducibility and fairness. On the positive side, the paper provides a clear high-level description of the VGG architecture, data augmentation pipeline, and training schedule. However, many of the design choices are reported as a fixed recipe without specific details of the components and its relation to the outcome. For example, the augmentation strategy combines scaling, shifting, rotation, ten-crop evaluation, and random erasing, each applied with a probability of 50%, but the paper does not quantify how much each component contributes to the final 73.28% accuracy. It is therefore unclear whether all steps are necessary, or whether a simpler augmentation scheme would perform comparably. Similarly, the authors train for 300 epochs with SGD with Nesterov momentum, Reduce-on-Plateau learning rate scheduling, and additional cosine-annealing fine-tuning, but there is no analysis of convergence behavior or sensitivity

to training duration.

Reproducibility is further complicated by missing implementation details. The paper does not report the random seeds used for initialization and data augmentation, making exact replication of the reported 73.28% result difficult. The ten-crop evaluation also implies an increase in effective test-time computation and parameter usage (due to repeated forward passes over multiple crops), yet the exact number of parameters and the computational cost are not reported, which hinders fair comparison with lighter architectures or real-time systems. Moreover, while the confusion matrix and saliency maps provide some interpretability, the evaluation remains entirely focused on aggregate accuracy and per-class performance, without any disaggregation by demographic attributes or discussion of potential fairness issues. As a result, we cannot tell from this work whether the performance gains benefit all user groups equally, or whether some groups are systematically disadvantaged.

The demographic bias framework of Dominguez-Catena et al. (Dominguez-Catena et al., 2022, 2023) also comes with important caveats that affect how we can adapt it to FER2013. Their metrics require demographic labels and thus rely on FairFace as a proxy model to infer apparent race and gender. While this is a pragmatic solution for unlabeled FER datasets, it introduces several layers of approximation: FairFace itself is imperfect and reflects its own biases, gender is treated as a binary attribute, and race is collapsed into a limited set of coarse categories. Any demographic analysis performed on top of these predictions will therefore be influenced by FairFace's misclassifications and by the underlying normative choices about how to categorize people. In addition, their work emphasizes that representational bias (imbalanced group frequencies) and stereotypical bias (label–demographic associations) are statistical properties of the dataset, not direct measures of normative fairness. A dataset can be demographically skewed without implying that any particular deployment is unfair, and conversely, even balanced datasets can yield unfair outcomes if the model or decision context is problematic.

Finally, there is a mismatch between the datasets and settings in which the bias metrics were originally validated and our target benchmark. Dominguez-Catena et al. focus on AffectNet and FER+, which differ from FER2013 in resolution, label distributions, and collection procedures (Dominguez-Catena et al., 2022, 2023). Their empirical findings, such as race-related model bias being relatively insensitive to dataset balancing, while gender bias is more easily mitigated, may not transfer directly to FER2013. Applying their framework to our setting therefore requires additional thought: we must account for the lack of ground-truth demographic labels, the small image size and grayscale format, and the specific class imbalances of FER2013. Our critiques highlight that both the accuracy-oriented FER2013 literature and the emerging demographic-bias analyses leave important gaps. These gaps motivate our proposed extension, in which we adapt the high-performing FER2013 model of Khaireddin and Chen and evaluate it through a demographic lens, while being explicit about the limitations of the underlying dataset and proxy demographic labels.

4 Proposed Extensions

Our primary extension is to take the high-performing FER2013 VGG model of Khaireddin and Chen (Khaireddin and Chen, 2021) and place it within a fairness-aware evaluation pipeline. Architecturally, we do not introduce new hand-crafted features or a separate feature-engineering stage: as in the original work, the model is a purely convolutional VGG-style network that learns feature representations end-to-end from raw pixels. To ensure comparability with their results, we reproduce their data preprocessing and augmentation strategy as closely as possible. Concretely, we use the official FER2013 training, validation, and test splits; apply random scaling (up to $\pm 20\%$), horizontal and vertical shifts (up to $\pm 20\%$), and rotations (up to $\pm 10^{\circ}$) during training, each with probability 50%; then perform ten-crop evaluation on 40×40 patches at test time, with random erasing applied to crops during training and all pixel intensities normalized (e.g., by division by 255). The model is trained for 300 epochs using stochastic gradient descent with Nesterov momentum, weight decay, and a Reduce-on-Plateau learning rate scheduler, optionally followed by a short cosine-annealing fine-tuning phase. This setup preserves the core architectural, optimization, and augmentation choices of (Khaireddin and Chen, 2021), yielding a baseline single-network accuracy on FER2013 that we can compare both to their reported results and to our subsequent fairness-aware analyses.

To incorporate demographic considerations, we then augment FER2013 with proxy demographic la-

bels in a manner inspired by Dominguez-Catena et al. (Dominguez-Catena et al., 2022, 2023). Specifically, we pass each face image through the DeepFace framework to obtain estimates of apparent race, gender, and age. These attributes are not treated as ground truth, but rather as noisy proxies that enable group-level analysis on a dataset that does not provide demographic metadata natively. Once FER2013 has been annotated in this way, we re-run the same VGG architecture on the dataset, preserving all modeling and training hyperparameters and using the same simple preprocessing (fixed resolution and normalization, with no geometric augmentation or ten-crop evaluation). The network is again trained for 300 epochs on the original training split, and evaluation is performed on the held-out test set in a single forward pass per image. The key difference is that, in addition to reporting a single aggregate test accuracy, we stratify the predictions by demographic group, computing accuracy, recall, and confusion matrices separately for each apparent race, gender, and age category.

Within this framework, fairness is assessed post hoc by comparing outcome metrics across demographic groups. In line with our operational definition of fairness, we treat the model as *fairer* when accuracy and recall are more similar across groups, and we view large systematic gaps as evidence of unfair treatment. Concretely, we plan to report (i) per-group overall accuracy, (ii) per-group, per-class recall, and (iii) simple disparity summaries such as the difference between the best and worst performing group for each emotion class. Where appropriate, we also draw on the recall-disparity style metrics proposed by Dominguez-Catena et al. (Dominguez-Catena et al., 2022) to summarize group differences into a single overall disparity score. Because the architecture, training procedure, and data splits are held fixed relative to (Khaireddin and Chen, 2021), any observed disparities can be attributed to interactions between the original FER2013 data distribution and the VGG-based model, rather than to changes in model design.

In summary, our proposed extension does not attempt to redesign the FER model itself, but rather to wrap an existing state-of-the-art FER2013 architecture in a demographic analysis layer. By first replicating the original training setup and then stratifying evaluation by apparent race, gender, and age, we aim to answer a question that the original work leaves open: not only how accurate the model is on FER2013 overall, but how that accuracy is distributed across different demographic groups.

5 Proxy Demographic Annotation via DeepFace

We encountered multiple difficulties using FairFace in a reproducible and computationally stable manner (issues related to CUDA, dlib, and pretrained model dependencies). To ensure that a demographic analysis was still conducted, we adopted *DeepFace* as an alternative.

Similar to FairFace, DeepFace is an open-source facial analysis toolkit that integrates multiple pretrained CNNs for face detection and attribute inference, including models for apparent gender, race, and age estimation. For each image in the FER2013, we used the tool to infer:

- Dominant apparent gender
- Dominant apparent race

To improve robustness and prevent failures from breaking the process, labels were assigned in batches of 1000 images, with intermediate results written incrementally to the csv file. This allowed recovering from crashes without losing large chucks of data.

We emphasize that these demographic attributes are *not treated as ground truth*, but as noisy proxy labels derived from a separate model. They reflect how another automated system categorizes individuals, not how individuals identify themselves. Consequently, any fairness analysis presented in this work should be interpreted as an audit of model *behavior under apparent demographic groupings*, rather than a claim about actual human populations.

6 Dataset Demographic Audit

Augmenting the dataset with demographic metadata also allowed us to analyze the dataset composition independently from any model predictions. This makes it possible to distinguish dataset imbalance from model-induced disparities.

6.1 Gender Distribution

FER2013 is demographically imbalanced with respect to apparent gender. Approximately 61% are classified as male and 39% female. This indicates that male-presenting faces are substantially overrepresented, which may affect the patterns learned during training and evaluation phases.

```
=== GENDER DISTRIBUTION ===
apparent_gender
Man      21487
Woman    14400
Name: count, dtype: int64

Proportions:
 apparent_gender
Man      0.59874
Woman    0.40126
Name: proportion, dtype: float64
```

Figure 1: FER2013 gender distribution

6.2 Race Distribution

When it comes to race distribution, FER2013 is also heavily skewed by apparent race. Images classified as white constitute the dominant group (61.7%), followed by Asian (19.8%), Black (6.3%), Latino/Hispanic (6.2%), and Middle Eastern categories (6.0%), with the Indian category accounting for only a very small fraction of the dataset. This imbalance in race distribution mirrors findings reported for other public FER benchmarks and reflects the lack of demographic control during dataset construction.

```
=== RACE DISTRIBUTION ===
apparent_race
white             22131
asian              7130
black              2250
latino hispanic    2208
middle eastern     1743
indian              425
Name: count, dtype: int64

Proportions:
 apparent_race
white             0.616686
asian             0.198679
black             0.062697
latino hispanic   0.061526
middle eastern    0.048569
indian            0.011843
Name: proportion, dtype: float64
```

Figure 2: FER2013 race distribution

6.3 Intersectional Distribution

Representation becomes even more uneven when we consider race and gender together. In most racial categories, male-presenting subjects dominate, resulting in poor representation for non-white women. This intersectional imbalance further limits the coverage of the feature space across demographic subgroups.

These results confirm that FER2013 exhibits substantial representational bias before any modeling, and that any fairness analysis must be interpreted in the context of this underlying skew.

```
=== RACE x GENDER ===
apparent_gender       Man     Woman
apparent_race
asian            0.625806  0.374194
black            0.822222  0.177778
indian           0.771765  0.228235
latino hispanic  0.595562  0.404438
middle eastern   0.836489  0.163511
white            0.545570  0.454430
```

Figure 3: FER2013 joint distribution of race and gender

7 Methods and Experimental Setup

All models were implemented in PyTorch and trained on the FER2013 dataset using the official training, validation, and test splits.

7.1 Model Architecture

We implement a VGG-style convolutional neural network following the design of Khaireddin and Chen (Khaireddin and Chen, 2021), adapted to FER2013 grayscale inputs and extended for fairness analysis. The architecture consists of four convolutional blocks, each comprising two 3×3 convolution layers followed by batch normalization, ReLU activation, and 2×2 max pooling.

The number of feature channels doubles at each stage:

$$1 \to 64 \to 128 \to 256 \to 512.$$

After the final block, adaptive average pooling is applied to produce a fixed 3×3 output regardless of input resolution. The resulting feature map is flattened and passed through three fully connected layers:

$$512 \cdot 3 \cdot 3 \to 512 \to 512 \to 7,$$

where 7 corresponds to the number of emotion categories in FER2013. ReLU nonlinearity is used after the first two fully connected layers, followed by dropout with probability $p = 0.5$. The final layer outputs class logits that are passed to a cross-entropy loss during training.

7.2 Data Preprocessing and Augmentation

FER2013 images are grayscale and provided at a resolution of 48×48 pixels. All images are converted to floating-point tensors and normalized

to the range $[0, 1]$ by dividing pixel intensities by 255.

Training transforms. To improve robustness to geometric variations, we apply stochastic data augmentation during training:

- Random transformation (applied with probability 0.5) including:
  - Rotation up to $\pm 10°$,
  - Horizontal and vertical translation up to $\pm 20\%$,
  - Scaling between 0.8 and 1.2.
- Random cropping to 40×40 pixels,
- Random erasing with probability 0.5, to simulate occlusion.

Validation and test preprocessing. No augmentation is applied during validation or testing. Instead, center cropping is used for validation, and ten-crop evaluation is applied at test time. Ten crops are extracted from each test image (four corners, center, and their mirrored versions), and predictions are averaged over all crops.

7.3 Training Configuration

The model is trained using stochastic gradient descent (SGD) with Nesterov momentum under the following configuration:

- Optimizer: SGD with Nesterov acceleration,
- Momentum: 0.9,
- Weight decay: 10^{-4},
- Initial learning rate: 0.01,
- Loss function: cross-entropy.

Training is performed for 300 epochs with batch size 128. The model checkpoint achieving the highest validation accuracy is saved and used for final testing.

7.4 Learning Rate Scheduling

We apply a Reduce-on-Plateau learning rate scheduler that monitors validation accuracy:

- Reduction factor: 0.75,
- Patience: 5 epochs,
- Mode: maximize validation accuracy.

When validation performance plateaus, the learning rate is reduced, improving convergence during later training stages.

7.5 Evaluation

The final model is evaluated on the FER2013 private test split using ten-crop inference. Predictions across all crops are averaged to produce final probability estimates. We report:

- Overall test accuracy,
- Confusion matrix,
- Per-class recall.

8 Results and Fairness Evaluation

8.1 Confusion Matrix and Per-Class Performance

Table 1 reports precision, recall, and F1-score for each emotion class. The confusion matrix in Figure 4 visualizes common misclassification patterns.

Table 1: Per-class precision, recall, and F1-score on FER2013 test set

Emotion	Precision	Recall	F1-score	Support
Angry	0.662	0.654	0.658	491
Disgust	0.906	0.527	0.667	55
Fear	0.631	0.519	0.570	528
Happy	0.894	0.906	0.900	879
Sad	0.569	0.643	0.604	594
Surprise	0.840	0.832	0.836	416
Neutral	0.689	0.732	0.710	626
Accuracy		0.7261		
Macro Avg	0.742	0.687	0.706	–
Weighted Avg	0.728	0.726	0.725	–

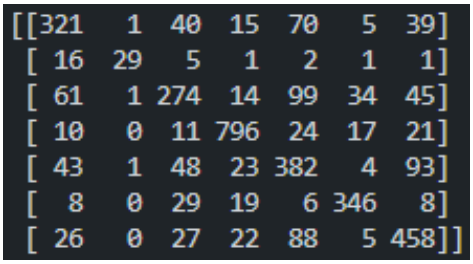

Figure 4: Confusion matrix for FER2013 test set predictions

The model achieves its highest performance on the "happy" class, with recall exceeding 90%, followed by "surprise" at over 83%. These expressions exhibit strong and consistent facial cues (e.g., smiles, open mouth, raised eyebrows), making them easier to distinguish.

The most challenging emotions are "fear" and "disgust", with recall near 52%. In the case of "disgust", this is largely attributable to severe class imbalance (only 55 examples in the test set). "Fear" is frequently confused with "surprise" and "sad", reflecting overlapping facial features such as widened eyes and raised brows. Similarly, "sad" is often confused with "neutral", which is a common issue in FER models due to subtle expression differences.

Overall, the confusion matrix shows that errors are not uniformly distributed across classes. Instead, misclassifications are concentrated among visually similar emotions, which indicates that the model learns meaningful structure rather than making arbitrary mistakes.

These trends are consistent with prior work on FER2013 and the baseline reported by Khaireddin and Chen (Khaireddin and Chen, 2021), confirming that the trained model exhibits realistic and interpretable error behavior.

8.2 Overall FER Performance

The VGG-based model trained in this work achieves an overall test accuracy of 72.61% on the FER2013 private test split. This is within 0.7 percentage points of the 73.28% reported by Khaireddin and Chen (Khaireddin and Chen, 2021), indicating that our reimplementation successfully reproduces a strong baseline while enabling additional fairness analyses.

The per-class performance follows common patterns observed in prior work. "Happy" and "surprise" attain the highest recall, whereas "disgust" and "fear" remain the most challenging classes. Negative emotions such as "fear" and "sad" are frequently confused with each other, while "happy" is rarely misclassified.

8.3 Performance by Apparent Gender

Table 2 reports the number of test examples and corresponding accuracy for each group.

Table 2: Accuracy by apparent gender on the FER2013 test set.

Group	# Samples	Accuracy
Man	2192	69.89%
Woman	1397	76.88%

Overall accuracy is approximately 7 percentage points higher for women than for men. However, recall averaged over emotion classes is very similar (67.61% for men vs. 68.67% for women), and the average per-class recall disparity between genders is 6.8 percentage points. This suggests that both gender groups are "seen" by the model to a comparable degree, but that the distribution of errors across classes is not identical. From a fairness perspective, the model is not gender-neutral: women benefit from consistently higher accuracy, even though men are more prevalent in the dataset.

8.4 Performance by Apparent Race

Table 3 shows the corresponding breakdown by apparent race. We consider six racial categories predicted by DeepFace. It is important to note that the Indian group is reported for completeness but contains relatively few examples.

Table 3: Accuracy by apparent race on the FER2013 test set.

Group	# Samples	Accuracy
Asian	745	75.57%
Black	219	73.97%
Indian	36	66.67%
Latino/Hispanic	238	68.49%
Middle Eastern	197	73.10%
White	2154	71.96%

The maximum difference in accuracy between racial groups is roughly 8.9 percentage points, with Asian-presenting faces achieving the highest accuracy (75.57%) and Indian-presenting faces the lowest (66.67%). As stated previously, since the Indian group is small ($n = 36$), these numbers should be interpreted with caution. Even when focusing on the better-represented racial groups, however, we observe several percentage points of variation: Black and Middle Eastern subjects achieve higher accuracy than White subjects, while Latino/Hispanic faces perform somewhat worse.

These results indicate that the model's behavior is sensitive to apparent race. Performance is not uniformly higher for majority groups. Instead, several minority groups outperform whites, while others underperform. This pattern suggests that demographic advantage is not determined solely by representation frequency, but by more complex interactions between data distribution and learned representations.

8.5 Summary of Disparities

From a fairness perspective, FER2013 and the VGG-based model exhibit both dataset-level and model-level disparities. At the dataset level, male

and white-presenting faces are overrepresented. At the model level, accuracy differs by approximately 7 percentage points between apparent genders and by up to 8.9 percentage points between racial groups. Although recall is nearly identical for men and women, indicating similar detection rates, the residual accuracy and class-wise recall gaps show that error patterns remain demographically structured. These results show that high overall accuracy can hide meaningful differences in how the model performs across demographic groups.

9 Conclusion

In this work, we reproduced a high-performing VGG-based facial emotion recognition model on the FER2013 benchmark and extended its evaluation beyond aggregate accuracy to include a demographic fairness analysis. Our implementation achieved an overall test accuracy of 72.61%, closely matching the state-of-the-art single-network performance reported by Khaireddin and Chen (Khaireddin and Chen, 2021). This confirms that the model architecture and training procedure generalize well and provides a reliable foundation for subsequent fairness evaluation.

To enable demographic analysis on a dataset that lacks ground-truth demographic labels, we augmented FER2013 with proxy labels for apparent race and gender using DeepFace. Although these labels do not represent true identities, they allow for a practical audit of model behavior under apparent demographic groupings. In addition to that, we conducted a dataset-level audit and found substantial representational imbalance: male-presenting and white-presenting faces are overrepresented, while non-white women are particularly underrepresented. This confirms that demographic skew is already present at the dataset level, independently of any model training.

The model evaluation revealed disparities in performance across demographic groups. Accuracy differed by approximately 7 percentage points between apparent genders and by up to 8.9 percentage points across racial groups. Interestingly, these disparities did not correspond to a simple majority advantage: in several cases, minority groups achieved higher accuracy than the majority group, indicating that bias does not arise solely from data frequency but from more complex interactions between representation quality and learned features. While recall for men and women was nearly identical, suggesting comparable detection rates, the residual differences in accuracy and classwise recall indicate that errors remain systematically structured.

These findings demonstrate that high overall performance does not guarantee equitable behavior. A model that appears successful in aggregate can still exhibit uneven performance across demographic groups, raising concerns for real-world deployment in social, educational, or clinical contexts. Our results reinforce the argument that fairness analysis should be treated as a standard component of model evaluation rather than an optional add-on.

This study has several limitations. Demographic labels were inferred using a secondary model and therefore reflect algorithmic categorization rather than self-identified attributes. Additionally, intersectional evaluations were constrained by small subgroup sizes, impacting statistical reliability for underrepresented populations. Future work should prioritize datasets with explicit demographic annotation, and incorporate uncertainty-aware labelling.

In conclusion, this work demonstrates that state-of-the-art FER performance can coexist with demographic disparity. By combining model replication, dataset auditing, and fairness evaluation within a single framework, this project highlights the necessity of broadening evaluation standards beyond accuracy alone, and contributes to a more responsible and transparent understanding of facial emotion recognition systems.

References

F. Abdat, C. Maaoui, and A. Pruski. 2011. Human-computer interaction using emotion recognition from facial expression. In *Proceedings of the UKSim 5th European Modelling Symposium on Computer Modelling and Simulation (EMS)*.

I. Dominguez-Catena, D. Paternain, and M. Galar. 2022. Assessing demographic bias transfer from dataset to model: A case study in facial expression recognition. In *Proceedings of the IJCAI–ECAI 2022 Workshop on Artificial Intelligence Safety (AISafety 2022)*, Vienna, Austria.

Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. 2023. Metrics for dataset demographic bias: A case study on facial expression recognition. *Preprint*, arXiv:2303.15889. ArXiv preprint.

B. Fasel and J. Luettin. 2003. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1).

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. 2015. Challenges in

representation learning: A report on three machine learning contests. *Neural Networks*, 64.

D. K. Jain, P. Shamsolmoali, and P. Sehdev. 2019. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120.

Y. Khaireddin and Z. Chen. 2021. Facial emotion recognition: State of the art performance on fer2013. Technical report, Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6).

N. Mehendale. 2020. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(3).

C. Pramerdorfer and M. Kampel. 2016. Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*.

E. Sariyanidi, H. Gunes, and A. Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6).