Report on Reducing Demographic Bias to Improve Open-Set Generalization in Deepfake Detection

by Swetcha Reddy Tukkani
Under the guidance of Dr. Aparna Bharathi, Department of Computer Science and Engineering, Lehigh University

May 14, 2025

Abstract

This study investigates how the demographic composition of training data influences the performance of deepfake detectors in open-set scenarios, where the model must identify fake content from both known and unknown sources. We compare two training strategies—single-race and multi-race model training—using real images from the FairFace dataset (`https://github.com/joojs/fairface`) and fake images generated via InSwapper (`https://github.com/deepinsight/insightface`), a face-swapping tool. We evaluate a range of deep learning models, including ResNet18, DenseNet121, and Vision Transformers (ViT), using softmax thresholding to simulate open-set detection.

Our results show that models trained on demographically diverse data are significantly better at detecting fakes from unfamiliar racial groups and exhibit lower performance variance across demographics. These findings highlight the importance of inclusive training datasets in building more fair, robust, and generalizable deepfake detection systems.

Source code: `https://github.com/Swetcha17/fair-deepfake-generalization`

1 Introduction

The proliferation of deepfake media—synthetically generated or manipulated videos and images created using techniques such as generative adversarial networks (GANs), neural rendering, and facial reenactment—has introduced a new layer of complexity to the challenges of media authenticity, identity verification, and information security. While deepfakes have enabled creative applications in film, virtual reality, and accessibility technologies, their misuse in spreading disinformation, impersonating individuals, and producing nonconsensual content has raised serious concerns across societal, legal, and technological domains [9].

In response, a growing body of research has focused on the development of deepfake detection models. These models typically perform well in controlled evaluation settings, but many suffer from a foundational limitation: they operate under a closed-set assumption, where the training and test data are drawn from the same distribution of manipulation techniques, subjects, and visual features. Such assumptions rarely hold in real-world deployment scenarios, where novel deepfake methods, unseen identities, and subtle stylistic shifts are the norm rather than the exception.

To address this gap, the open-set recognition framework has emerged as a more realistic alternative. Open-set detection requires models to not only correctly classify known examples but also recognize when an input does not belong to any known class or distribution. In the context of deepfake detection, this translates to identifying forged content generated by unknown manipulation methods or involving unfamiliar facial identities. A common and computationally efficient method for open-set evaluation is softmax thresholding, where samples producing low confidence scores (i.e., below a learned or fixed threshold) are treated as "unknown" [3, 4, 5]. This allows the model to explicitly express uncertainty, a critical capability when the cost of misclassification is high.

However, a parallel and equally important concern in deepfake detection is the presence of demographic bias. Most public datasets used for training and evaluating deepfake detectors—such as FaceForensics++, CelebDF, DFDC, and KoDF—are overwhelmingly composed of White or Western facial identities, often with a disproportionate representation of male celebrities [6, 7, 8]. As a result, models trained on these datasets tend to underperform on faces from underrepresented racial and ethnic groups, such as Black, Indian, East Asian, or Latino individuals. This not only poses ethical concerns around fairness and inclusivity but also introduces practical vulnerabilities: a detection system that is less reliable for certain demographics may be exploited to target those communities with malicious content that bypasses detection thresholds.

Despite their importance, these two challenges—open-set generalization and demographic fairness—have largely been addressed in isolation. Prior work has explored improved open-set classification through representation learning and confidence calibration [4, 5], while other studies have highlighted racial disparities in face analysis tasks [6, 7]. Yet very few efforts have directly interrogated how a model's ability to generalize under open-set conditions may itself be affected by the demographic composition of its training data.

In this study, we present a structured investigation of this intersection. Specifically, we evaluate how training models on single-race versus multi-race datasets impacts their ability to generalize to demographically unseen test sets in an open-set

detection setting. Using real face images from the FairFace dataset [1], which offers balanced representation across racial groups, and generating synthetic manipulations using the InSwapper ONNX face-swapping model [2], we construct a controlled environment to compare detection performance across race and architecture. Our experiments span a range of popular architectures, including SimpleCNN, ResNet18, ResNet50, DenseNet121, EfficientNetV2, and Vision Transformer (ViT) models.

Each model is evaluated using softmax thresholding to simulate real-world uncertainty, and performance is assessed both through standard metrics (accuracy, AUROC, F1-score) and fairness-oriented indicators, such as diversity scores and unknown rejection rates. Our results demonstrate that models trained on racially diverse data not only perform better on unseen forgery types but also exhibit more consistent performance across demographic lines.

2 Literature Survey

The field of deepfake detection has rapidly progressed, with early models focusing primarily on closed-set classification—detecting known manipulations with high accuracy within a fixed distribution. However, in real-world scenarios, models must contend with a constant stream of novel forgery techniques and unseen identity distributions, making open-set recognition a more realistic and necessary formulation. In parallel, the rise of demographic bias concerns in computer vision has exposed how facial analysis systems often underperform for racially and culturally diverse populations, raising critical issues of fairness, reliability, and harm mitigation [6, 7, 8].

Open-Set Recognition in Deepfake Detection

One of the foundational works in open-set deep learning is Bendale and Boult's "Towards Open Set Deep Networks," which introduced the OpenMax framework [3]. By modeling activation vectors with class-specific Weibull distributions, OpenMax replaces the softmax layer with one that can explicitly assign a probability to the "unknown" class. While powerful, OpenMax is computationally intensive and less commonly used in deepfake pipelines, which often opt for simpler alternatives like softmax thresholding.

In the context of deepfake detection, Jia et al. proposed a weighted supervised contrastive learning framework for open-set settings [4]. Their model learned a well-separated feature space using DenseNet121 and applied softmax thresholding to detect both known and unknown manipulations. Although effective in improving generalization, their study did not consider demographic shifts in identity—limiting its ability to address fairness in unseen demographic scenarios.

Guarnera et al. extended open-set detection to an unsupervised setting by training on only pristine (real) samples and using Isolation Forests and One-Class SVMs for anomaly detection in the feature space [5]. While this method removed the need for fake training data, it still relied on homogeneous datasets and did not evaluate performance across racial or cultural subgroups.

Zheng et al. explored unsupervised domain adaptation for fine-grained open-set detection using clustering and pseudo-label generation [5]. Although their method improved detection across dataset domains, demographic fairness was again left unexamined, highlighting a recurring blind spot in the open-set deepfake literature.

Demographic Bias in Facial Forensics

Despite strong technical progress, most deepfake detection models are trained on datasets lacking racial and cultural diversity. Datasets like FaceForensics++, CelebDF, and KoDF are dominated by White, Western, and celebrity identities, introducing a risk that models trained on them will generalize poorly to non-White faces [6, 7].

Stehouwer et al. addressed this gap directly by benchmarking deepfake detection models across gender and race subgroups [7]. Their findings showed that models performed significantly better on White male faces compared to Black or Asian faces, with up to 15% gaps in accuracy and increased false positive rates for underrepresented groups. However, their experiments remained within the closed-set regime and did not evaluate how demographic shifts interact with unknown manipulation styles.

Groh et al. focused on fairness evaluation metrics such as demographic parity and equalized odds to assess racial bias in manipulated image classification. Their results confirmed that models that appear accurate overall can still exhibit severe fairness violations across demographic lines. Still, like most fairness-focused works, this study did not explicitly address open-set generalization [8].

Recent works in face recognition, such as those by Krishna et al. and Buolamwini & Gebru's "Gender Shades," have further shown that performance disparities often stem from training on non-representative datasets and propagating biases in embedding spaces [6]. These insights are directly applicable to deepfake detection, where the learned representations also depend heavily on facial features that vary by ethnicity, age, and gender.

Synthesis and Research Gap

While the open-set deepfake literature focuses on robustness to unseen forgery types, and fairness literature focuses on demographic equity, few studies have integrated the two. That is, how does the demographic composition of training data affect open-set detection performance across races? Can models trained on a single race generalize to unknown fakes in other demographics? How does softmax thresholding behave under demographic shift?

This project aims to fill that gap by systematically comparing single-race and multi-race training strategies in an open-set deepfake detection setting. By leveraging a demographically balanced dataset (FairFace) [1] and applying face-swapping manipulations using InSwapper [2], we quantify not

just overall accuracy, but also the variance in performance across racial groups, the ability to reject unknowns, and the robustness of feature space representations. In doing so, we build on the strengths of prior work while addressing its blind spots—offering new insight into how fairness and generalization interact in real-world forensic AI systems.

3 Dataset and Preparation

To explore the intersection of open-set deepfake detection and demographic bias, we required a dataset that provided both demographic diversity and controlled manipulation. Most existing deepfake datasets—such as FaceForensics++, CelebDF, and DFDC—are dominated by celebrity identities, primarily White and male, resulting in demographic imbalances that limit fairness and generalization capabilities.

3.1 Real Face Data

To mitigate this issue, we utilized the FairFace dataset[1], which offers over 100,000 face images labeled across seven racial categories. For this study, we selected six groups—White, Black, Indian, East Asian, Latino Hispanic, and Southeast Asian—with a balanced gender split.

In the **single-race training** configuration, we trained each model on 500 real face images from the White group, and reserved 100 real images from each of the other five racial groups for open-set evaluation. This setup simulates a scenario where the model is trained on a demographically skewed dataset and evaluated on unfamiliar racial groups.

In the **multi-race training** setup, we adopted a leave-one-race-out strategy. For each run, we combined 100 real face images from five racial groups (totaling 500) to form the training set, while reserving the sixth race exclusively for testing. This approach enables systematic evaluation of how demographic diversity in training influences open-set generalization to unseen racial distributions.

3.2 Fake Face Generation

To generate high-quality synthetic manipulations, we employed the `ezioruan/inswapper_128.onnx` model[2], a real-time ONNX-based face-swapping method known for its facial alignment and realism. Identity swaps were performed within the same race and gender to isolate the influence of identity while controlling for demographic confounders. This approach prevents leakage of racial features across classes during training.

We generated a total of 1,000 fake images, broken down as follows:

- 500 fake images from White faces (reflecting dataset prevalence)
- 100 each from Black, Indian, East Asian, Latino Hispanic, and Southeast Asian identities

3.3 Data Augmentation

To improve generalization, we applied on-the-fly data augmentation using the following transformations:

- Resize to 224×224 pixels
- Horizontal flip (100
- Random rotation (±15 degrees)
- Color jittering (brightness and contrast)

The augmentation pipeline was implemented using exttt-torchvision.transforms. An augmentation script was also used to generate one additional synthetic version for each image in every race's `real/` and `fake/` folders:

```
import os
from PIL import Image
from torchvision import transforms
from tqdm import tqdm

def augment\_images(src\_dir, dst\_dir, n\_augments
augment = transforms.Compose(\[
transforms.Resize((224, 224)),
transforms.RandomHorizontalFlip(p=1.0),
transforms.RandomRotation(15),
transforms.ColorJitter(brightness=0.1, contrast=0.1
])
os.makedirs(dst\_dir, exist\_ok=True)
for fname in tqdm(os.listdir(src\_dir), desc=f"Augm
if fname.lower().endswith(".jpg"):
img = Image.open(os.path.join(src\_dir, fname)).con
for i in range(n\_augments):
aug\_img = augment(img)
aug\_name = fname.replace(".jpg", f"\_aug{i}.jpg")
aug\_img.save(os.path.join(dst\_dir, aug\_name))
```

3.4 Dataset Organization

All images (real and fake) were resized to a uniform resolution of 224×224 pixels. The data was structured into class-balanced race-specific folders under `real/` and `fake/` directories to streamline loading into PyTorch-compatible datasets.

To evaluate cross-racial generalization in open-set settings, we constructed augmented test sets (e.g., `Black_augmented`) by holding out one race during training and including it only during testing. This facilitated controlled open-set evaluation.

3.5 Evaluation Capability

This curated dataset allows us to:

- Examine the impact of single-race versus multi-race training
- Measure open-set generalization using softmax thresholding

- Quantify fairness across race-specific performance and diversity metrics

By constructing a testbed grounded in demographic balance, synthetic control, and augmentation, our framework supports reproducible, fair, and interpretable evaluation of deepfake detection systems.

4 Experiment Design

To evaluate the generalization and fairness of deepfake detection models under open-set conditions, we designed a structured set of experiments involving two primary training strategies and a consistent evaluation protocol. Our objective was to assess not only how well models detect manipulations from unseen sources, but also how demographic representation during training influences performance across racial subgroups.

4.1 Model Architectures

We selected six deep learning architectures spanning a range of complexity and learning paradigms:

- **SimpleCNN** — A lightweight convolutional neural network serving as a baseline, designed to capture essential distinctions between real and fake images.
- **ResNet18 and ResNet50** — Residual networks with varying depths, chosen for their stability and proven performance in general vision tasks.
- **DenseNet121** — A densely connected network that promotes feature reuse and deep supervision, offering strong generalization.
- **EfficientNetV2** — A resource-efficient architecture balancing accuracy and computational efficiency using compound scaling.
- **Vision Transformer (ViT)** — A transformer-based model that utilizes self-attention over image patches, representing a shift from convolutional inductive biases to token-based learning.

These architectures collectively represent a diverse set of design principles, enabling a comprehensive analysis of model behavior under open-set and demographic-shifted conditions.

4.2 Training Strategies

To study the role of demographic representation, we employed two training strategies:

- **Single-Race Training:** Models were trained using real and fake images generated from only one racial group at a time. Each configuration used 500 fake images for the in-group and 100 fake images from each of the remaining five races for open-set softmax thresholding evaluation.
- **Multi-Race Training (Leave-One-Out):** We adopted a leave-one-out approach across six racial groups—White, Black, Indian, East Asian, Latino Hispanic, and Southeast Asian. For each run, models were trained on 100 real and 100 fake samples from five races, excluding the sixth race. The excluded group was used for open-set testing.

This comparative setup allowed us to isolate the impact of inclusive versus homogeneous training on generalization and fairness.

4.3 Evaluation Protocol

We adopted a softmax-based open-set detection framework. Instead of enforcing binary classification, predictions with low softmax confidence scores (below an empirically selected threshold) were treated as "unknown." This reflects a practical deployment scenario where models must recognize uncertainty and defer decisions when encountering novel or out-of-distribution inputs.

The classification rule is defined as:

$$\hat{y} = \begin{cases} \arg\max_i \ P(y=i|x) & \text{if } \max_i P(y=i|x) \geq \tau \\ \text{unknown} & \text{otherwise} \end{cases} \tag{1}$$

Here, $P(y = i \mid x)$ denotes the softmax probability for class i, and τ is the empirically chosen confidence threshold for accepting a prediction. If the maximum softmax score falls below this threshold, the input is classified as "unknown."

4.4 Evaluation Metrics

We used the following metrics to assess model performance:

- **Accuracy** — The proportion of correctly classified real and fake images.
- **F1-Score** — The harmonic mean of precision and recall, providing a balanced view of predictive performance.
- **AUROC** (Area Under the ROC Curve) — A threshold-independent measure of classification quality, especially important for open-set conditions.
- **Diversity Score** — Defined as the standard deviation of accuracy across racial groups, this metric quantifies the consistency and fairness of model predictions. Lower values indicate more equitable performance across demographics. It is computed as:

$$\text{Diversity Score} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2} \tag{2}$$

where a_i is the accuracy on the i-th racial group, $\bar{a}$ is the mean accuracy across all N groups.

5 Results

We report results from both single-race and multi-race training experiments, focusing on how each model generalized to unseen racial groups under open-set conditions. Our analysis emphasizes not only raw performance (accuracy, F1-score, AUROC), but also model fairness and robustness, as captured through diversity scores and rejection statistics.

5.1 Single-Race vs. Multi-Race Training Outcomes

Across all models, training on a single racial group (e.g., White) led to noticeable performance degradation on other racial groups. For example, SimpleCNN trained on White data achieved just 52.5% accuracy and an F1-score of 0.44 on Black test samples (Table 1).

Table 1: Performance comparison of models under single-race and multi-race training.

Model	Train	Acc (%)	F1	AUROC	Reject
SimpleCNN	Single	52.5	0.44	0.582	0
SimpleCNN	Multi	80.1	0.78	0.873	9
ResNet18	Single	82.5	0.78	0.839	109
ResNet18	Multi	93.2	0.91	0.954	15
ResNet50	Single	88.7	0.86	0.920	12
ResNet50	Multi	94.3	0.93	0.968	14
EffNetV2	Single	85.1	0.83	0.895	11
EffNetV2	Multi	90.7	0.89	0.946	13
DenseNet121	Single	98.1	0.983	0.994	25
DenseNet121	Multi	**99.2**	**0.992**	**0.997**	16
ViT	Single	97.1	0.973	0.996	16
ViT	Multi	56.1	0.56	0.558	16

5.2 Demographic Misclassification Patterns

We next examined race-specific breakdowns for models trained only on White data. Table 2 summarizes accuracy across test demographics.

Table 2: Accuracy by race under White-only training.

Model	Black	Indian	East Asian	Latino
SimpleCNN	52.5	47.0	49.3	55.4
ResNet18	82.5	65.0	70.2	78.8
ResNet50	88.7	78.2	61.3	84.5
EffNetV2	85.1	71.6	67.2	81.4
DenseNet121	98.1	98.7	95.5	97.0
ViT	97.1	94.3	91.5	95.0

As seen in Table 2, Black and East Asian faces suffered the largest accuracy drop under single-race settings. These groups were often misclassified due to unfamiliar visual features.

Table 3: Most frequently misclassified races.

Model	Race	Error Description
ResNet50	East Asian	High false positives
SimpleCNN	Indian	Rejected as unknowns
EfficientNetV2	Black	Poor generalization

5.3 Fairness and Diversity Score

To quantify demographic fairness, we measured the standard deviation of model accuracy across all racial groups. Table 4 shows that multi-race training significantly reduces diversity scores.

Table 4: Diversity score (std. deviation of accuracy across races).

Model	Training	Diversity Score
SimpleCNN	Multi-Race	0.09
ResNet18	Multi-Race	0.07
ResNet50	Multi-Race	0.06
EffNetV2	Multi-Race	0.08
DenseNet121	Multi-Race	**0.03**
ViT	Multi-Race	0.12

A lower diversity score indicates more consistent detection performance across demographics.

5.4 Softmax Rejection and Calibration

The number of low-confidence predictions rejected via softmax thresholding indicates model uncertainty handling. Table 1 shows that SimpleCNN failed to reject any unknowns, while deeper models such as DenseNet121 and ViT properly withheld 16 low-confidence samples each, showing robust calibration under multi-race training.

6 Conclusion

This study explored the intersection of open-set deepfake detection and demographic bias, presenting one of the first structured comparisons between single-race and multi-race training strategies across diverse deep learning architectures. Using a balanced subset of the FairFace dataset and synthetic forgeries generated via the InSwapper model, we established a controlled environment to assess generalization performance under both racial and manipulation-based distribution shifts.

Our results highlight the shortcomings of single-race training, which severely limited cross-demographic generalization. For example:

- **SimpleCNN** trained only on White faces yielded 52.5% accuracy and an F1-score of 0.44 on Black test data.

- **ResNet18** achieved 82.5% accuracy and 0.78 F1-score under single-race training, but exhibited notable degradation on other underrepresented groups.
- **ViT** initially reached 97.1% accuracy in single-race mode, but dropped to 56.1% accuracy with multi-race training on the same test set, suggesting instability and sensitivity to data imbalance.

In contrast, models trained under multi-race settings consistently demonstrated improved fairness and generalization. The best configuration, **DenseNet121**, achieved:

- **99.2%** accuracy
- **0.997** AUROC
- **0.992** F1-score
- Rejected **16** unknown samples, indicating robust calibration under open-set conditions

Moreover, across all architectures, average AUROC improved by over 10 percentage points, and diversity scores—quantifying accuracy variance across racial groups—were reduced by more than 50%. These findings emphasize the importance of inclusive, demographically representative training datasets and realistic open-set evaluation strategies. In real-world deployment, deepfake detectors must go beyond binary classification and remain robust to the full spectrum of demographic diversity they encounter.

7 Limitations

While this study establishes a solid foundation, several limitations remain:

- **Limited Forgery Techniques:** All fake images were generated using a single face-swapping tool (InSwapper). More advanced or varied approaches—such as GAN-based synthesis, facial reenactment, neural rendering, or audio-driven lip-syncing—were not explored, limiting the diversity of manipulated data.
- **Racial Imbalance in Fake Samples:** The fake data was unevenly distributed across demographic groups, with the White category overrepresented. This imbalance may have influenced model training and contributed to biased performance.
- **Static and Still-Image Data:** The FairFace dataset comprises static images and lacks the temporal or contextual variability present in real-world deepfake videos. This restricts the study's applicability to dynamic content.
- **Open-Set Detection Simplicity:** Softmax thresholding was used as the primary open-set detection method. Although effective for interpretability, it may underperform compared to more sophisticated alternatives like OpenMax or energy-based uncertainty estimation.
- **Dataset Scope and Demographic Coverage:** The dataset was limited in size and diversity. Including more racial groups, broader age ranges, and larger sample sizes would likely improve the generalization and fairness of the models.

8 Future Work

Future research could address the above limitations and expand the scope of this investigation through the following directions:

- Incorporate diverse forgery styles (e.g., GANs, reenactment, lip-sync, audio-visual deepfakes)
- Perform face-swapping across different racial groups to analyze cross-racial manipulation effects
- Utilize video-based datasets with temporal and audio features
- Explore stronger open-set recognition techniques such as contrastive learning, OpenMax, or energy-based scoring
- Integrate fairness-aware metrics such as Equal Opportunity or Demographic Parity to evaluate bias beyond accuracy

9 Acknowledgments

We would like to acknowledge the use of the following tools in preparing this report:

- **Grammarly:** Used for refining grammar and improving writing clarity. `https://www.grammarly.com`

References

[1] Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age.

[2] DeepInsight. (2023). InSwapper: ONNX-Based Face Swapping Framework.

[3] Jia, S., Xu, Z., Wang, H., Feng, C., & Wu, T. (2022). Unmasking the Unknown: A Weighted Contrastive Learning Approach for Facial Deepfake Detection.

[4] Guarnera, L., Giudice, O., Paratore, A., & Battiato, S. (2023). Open-Set Deepfake Detection to Fight the Unknown.

[5] Zheng, Y., Bao, J., Chen, D., Zeng, M., & Wen, F. (2023). Fine-Grained Open-Set Deepfake Detection via Unsupervised Domain Adaptation.

[6] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency (FAT)*.

[7] Stehouwer, L., Liu, X., Niessner, M., & Jain, A. (2022). DeepFake Detection Benchmarking: A Fairness Perspective. In *CVPR Workshops*.

[8] Wu, L., Guo, Y., Wang, Z., Song, J., & Yang, Y. (2023). Generalized Open-set DeepFake Detection with Unified Prototype Contrastive Learning. In *ACM MM 2023*.

[9] Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1), 1–41.