# Exploring Dataset Bias in Facial Gender Classification

Rebecca Du
rebecca.du@duke.edu
Duke University
Durham, North Carolina, USA

Anish Parmar
anish.parmar@duke.edu
Duke University
Durham, North Carolina, USA

Alex Whitehead
alex.whitehead@duke.edu
Duke University
Durham, North Carolina, USA

Ethan Cheng
ethan.cheng@duke.edu
Duke University
Durham, North Carolina, USA

## Abstract

This study investigates how dataset composition and model architecture jointly influences fairness and accuracy in face-based gender classification. Using convolutional neural networks (AlexNet and ResNet18) pre-trained on ImageNet and IMDB, we fine-tuned models on racially balanced ("Equal") and biased ("US") datasets to evaluate the effects of varying demographic representations in training. Across most experiments, balanced datasets produced higher overall accuracy and fairer performance across racial and gender subgroups, with the exception of ResNet-18 pre-trained on IMDB due to pre-existing feature bias. Ensemble models combining race-specific classifiers through majority voting further improved accuracy, reduced misclassification confidence, and trained more efficiently than unified models, with AlexNet showing the largest gains. A Mixture of Experts (MoE) AlexNet and ResNet18 models were investigated, which routes each image to their corresponding race-specific classifier; The AlexNet MoE produced the fairest and highest classification accuracy across all models. GradCAM visualizations revealed consistent misclassification patterns linked to certain characteristics, such as side profiles, age extremes, and occluded faces, highlighting the sensitivity of models to dataset composition. These findings underscore the importance of diverse and balanced datasets, the role of model depth in bias amplification, understanding model misclassifications and confidence, and the potential of ensemble training methods to improve robustness and fairness in facial classification systems.

## Keywords

Artificial Intelligence, Dataset creation, Face classification, Gender classification, Diversity, GradCAM, AlexNet, ResNet, Ensemble Models

**ACM Reference Format:**
Rebecca Du, Anish Parmar, Alex Whitehead, and Ethan Cheng. 2026. Exploring Dataset Bias in Facial Gender Classification. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

The rapid development of Artificial Intelligence (AI) in recent years has led to its application in a variety of fields. One such application is face-based gender classification, which has seen prominent usage in security systems and user analytics. However, research into the field has brought forth concerns regarding fairness and consistency for different demographic groups. These troubles raise questions about how dataset design and model architecture jointly influence model performance.

The interwoven nature of dataset composition and model architecture make it difficult to completely extricate bias from a model setup. Factors such as race and gender distribution the dataset impact the learned feature space, while architectural components like model depth may amplify or suppress imbalances.

Our project investigates how dataset curation and model architecture affect gender classification across racial groups. We designed experiments that test dataset composition and architectural design to determine their isolated and combined effects. Specifically, we compared models with the same architecture but trained on different datasets (racially balanced or skewed datasets) and evaluated the performance across intersectional subgroups. We also employ interpretability techniques to visualize trends in model bias and misclassification. In doing so, we are able to construct a set of guiding principles for dataset curation and model selection to improve face-based gender classification tasks.

Our findings provide further insight into improving fairness in facial analysis by identifying practices to improve generalization. Applying these findings can lead to the development of more transparent benchmarks for future gender classification systems.

## 2 Related Works

The issue of bias in face recognition and gender classification systems has been extensively researched. The topic first garnered widespread attention when [Buolamwini and Gebru 2018] conducted an audit of commercial gender classifiers. Their findings indicated that commercial IBM, Microsoft, and Face++ models consistently performed worse on individuals with darker skintones, with dark-skinned women having the worst classification rate. Their work established the foundation for closer investigation into intersectional bias in facial classification models.

To address the issue of imbalance in datasets, [Kärkkäinen and Joo 2019] created the FairFace dataset which is balanced across race, gender, and age. Their models trained on FairFace were consistently more accurate across demographic groups than competitors trained on biased datasets. We use FairFace as one of the primary datasets in our experiments to examine how dataset-level variations can influence model bias.

In contrast to what the FairFace researchers proposed, [Gwilliam et al. 2021] challenge the need for demographic balance. Their experiments demonstrated that training exclusively on one racial group (e.g. African faces) paradoxically led to less biased results

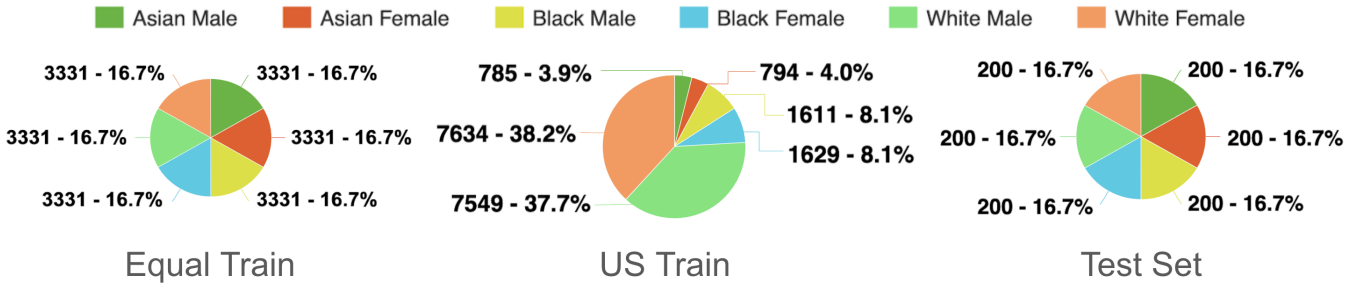



**Figure 1: Distribution of race and gender across datasets**

than training on diverse datasets. This discovery motivates our comparison of different training dataset splits, as well as our ensemble models.

More recent work by [Cherepanova et al. 2022] examined how dataset imbalance conceals bias measurement. By varying the ratio of certain demographic groups in the training and test datasets, the researchers demonstrated that imbalanced splits could mask performance disparities between genders. This motivates our investigation into controlling dataset balance across dimensions (e.g. race and gender), as well as using interpretability tools like GradCAM [Selvaraju et al. 2016] to visualize bias.

These studies demonstrate the shift from facial bias awareness to identifying axes like dataset design and model architecture to mitigate the issue. Our research is situated at the intersection of these axes in order to analyze how dataset curation strategies and network structure jointly influence gender classification.

## 3 Methodology

### 3.1 Overview

We conducted a set of experiments that varied both the training data distribution of different racial groups and the trained model architecture. These models were evaluated under two dataset splits and cross architectures, including pretrained, standard convolutional networks, and ensemble setups.

To isolate how individual factors impacted performance, we independently varied **dataset composition, pretraining sources** and **model architectures**. The results of these experiments are discussed in more detail in Section 4.

### 3.2 Datasets and Preprocessing

The training and testing datasets were created with a custom mix of UTKFace [Jangedoo 2019] and FairFace [HuggingFaceM4 2024] images. The datasets were selected for the variety of facial orientations and the quantity of data.

UTKFace consists of frontal face images with race labels covering 4 racial groups: White, Black, Asian, and Indian. Because of a skew towards White faces in UTKFace (42.5%), we incorporated FairFace for a larger and more comprehensive dataset. FairFace consists of images with a wide variety of facial orientations, quality, and clutter with race labels covering 7 racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. To create more interpretable experiments, we decided to only compare three racial subgroups: White, Black, and Asian. Labels were extracted from the metadata. For the Asian label, we combined UTKFace's Asian and Indian labels and FairFace's Indian, East Asian, and Southeast Asian labels.

Each image was resized to 224 × 224 pixels, normalized, and randomly transformed. We created two training splits to examine the impact of demographic imbalance. The training data consisted of 3 racial subgroups and 2 gender groups for a total of 6 subgroups. The balanced training set, called "Equal Train", consisted of a near equal proportion of all 6 subgroups (19,986 total). The imbalanced training set, called "US Train", was modeled after the 2022 US census [Facts 2024] (20,002 total). The percentages of the 6 subgroups were normalized to 100% to account for the races not included in our experiment. To assess generalization, we created testing data consisting of 200 images from each subgroup with half of the images coming from UTKFace and the other half coming from FairFace (1,200 total). These dataset splits can be seen in Figure 1.

### 3.3 Model Architecture and Training

We used different convolutional, MoE, and ensemble models to analyze how network architecture and pretraining dataset influenced gender classification. We chose ResNet18 [He et al. 2015] and AlexNet [Krizhevsky et al. 2012] as they are well-known models with varying levels of depth. Table 1 summarizes the architecture, pretraining configuration, and key traits of the models.

To note, ImageNet is a massive labeled visual database containing over 14 million images organized according to the WordNet hierarchy and widely used for image classification benchmarking and pretraining. IMDB-WIKI [Yuulind 2023] is a facial image dataset consisting of over 500,000 images of celebrities collected from IMDb and Wikipedia with age and gender labels for facial analysis tasks.

All models were trained for binary gender classification using cross-entropy loss and stochastic gradient descent (SGD) with early stopping based on validation accuracy. Each model was fine-tuned on either the Equal or US training splits described earlier.

### 3.4 Evaluation and Metrics

Model performance was assessed on test splits using **overall accuracy** and **subgroup accuracy**. We also reported overall confidence scores, subgroup confidence scores, and misclassification confidence scores. Misclassifications were numerically and visually compared across all models.

**Table 1: Evaluated models and their pretraining sources. All models were fine-tuned for binary gender classification.**

| Model (Architecture) | Pretraining Source |
|---|---|
| ResNet18 (Baseline) | ImageNet |
| ResNet18 (from scratch) | IMDB-WIKI |
| AlexNet (Baseline) | ImageNet |
| AlexNet (from scratch) | IMDB-WIKI |
| ResNet18 Ensemble | ImageNet |
| AlexNet Ensemble | ImageNet |
| ResNet18 MoE | ImageNet |
| AlexNet MoE | ImageNet |

Additionally, we performed **GradCAM** [Selvaraju et al. 2016] on the models for further interpretability. GradCAM (Gradient-weighted Class Activation Mapping) is a technique that uses the gradients of a target class flowing into the final convolutional layers of a CNN to produce a localization heatmap showing which regions of an input most informed a model's prediction.

## 3.5 Implementation Details

All experiments were implemented in PyTorch using the torchvision model zoo for pretrained architectures. Data pre-processing and augmentation were implemented using torchvision's transforms library. Models were trained and evaluated on a local CPU, cloud TPU, or local GPU. To ensure reproducibility, random seeds were fixed, and all dataset splits were pre-defined.

# 4 Results

## 4.1 Balanced vs. Imbalanced Datasets

We found that for most instances, having a more diverse unified training dataset improved overall accuracy.

**Table 2: Overall Accuracy Comparison between Equal and US train datasets.**

| Model | Equal Train | US Train |
|---|---|---|
| ResNet18 (ImageNet) | **89.42** | 89.33 |
| AlexNet (ImageNet) | **86.75** | 86.25 |
| AlexNet (IMDB) | **88.50** | 87.83 |
| ResNet18 (IMDB) | 87.83 | **88.33** |

Fine-tuning classification models on a dataset containing an equal distribution of racial groups (Equal Train) generally yielded higher overall accuracy compared to fine-tuning on a dataset with a biased distribution (US Train). This pattern held across all four models except ResNet18 when pretrained on the IMDB dataset as shown in Table 2.

This exception is likely due to a combination of ResNet's depth and the inherent feature bias in the IMDB dataset which primarily contains White faces. Since ResNet18 has a high number of layers, it is more likely to overfit on its learned bias features from pre-training. Fine-tuning with Equal Train on top of that, goes against the ingrained biased patterns learned from IMDB which results in lower accuracy. However, AlexNet did not share the same trend when pre-trained on IMDB. Since ResNet18 is deeper than AlexNet, it benefits more from diverse and balanced pretraining sources such as ImageNet, whereas AlexNet's limited depth constrains its ability to generalize under imbalanced conditions. This enables AlexNet to benefit more from fine-tuning on Equal Train.

Furthermore, fine-tuning on an equally distributed dataset affected accuracy differently across demographic subgroups, despite a consistent higher overall accuracy. Interestingly, across all splits, White males and females consistently achieved high accuracy, showing only a slight decrease in accuracy under Equal Train, where their representation in the dataset was substantially lower. This overall accuracy increase, combined with the observed White accuracy decrease, indicates improved accuracy for underrepresented groups (Black and Asian) and therefore a fairer performance distribution.

## 4.2 Unified vs. Ensemble

In this section, we test how different model architectures and training strategies affect accuracy, bias, and confidence in facial gender classification across racial subgroups, finding that the accuracy of a classification approach depends on the underlying model architecture.

**Table 3: Ensemble and MOE Performance**

| Model | White | Black | Asian | Overall |
|---|---|---|---|---|
| ResNet18 MoE | 93.00 | 84.50 | 83.25 | 86.92 |
| ResNet18 Ensemble | 93.00 | 87.25 | 89.25 | 89.83 |
| AlexNet Ensemble | 91.50 | 87.25 | 91.00 | 89.92 |
| AlexNet MoE | **98.75** | **98.00** | **96.00** | **97.58** |

*4.2.1 Ensemble Models.* To further explore the effect of dataset composition on gender classification, we created ensemble models consisting of three separate gender classification models, each fine-tuned exclusively on Asian, Black, or White racial subgroups. The training data consisted of a combination of our US Train and Equal Train datasets. Each ensemble model is based on AlexNet pretrained on ImageNet or ResNet-18 pretrained on ImageNet. The three fine-tuned models for each architecture were then combined using a majority-vote ensemble, where each model produced a prediction and the final output was determined by consensus.

We found that these ensemble models required less training time and achieved a higher overall accuracy, along with lower misclassification confidence, compared to models trained on the unified datasets (US Train or Equal Train). This improvement was consistent across both AlexNet and ResNet-18 architectures. Notably, AlexNet's overall accuracy increased by more than 3%, whereas ResNet-18 improved by only 0.41%. This difference likely reflects the impact of model depth: because AlexNet is shallower, combining multiple models effectively increases its representational capacity, while ResNet-18's greater inherent complexity yields smaller marginal gains.

We also hypothesize that the ensemble approach balances biases across the race-specific models, mitigating overfitting that may

occur when training on a single race model and improving upon the limited generalization of the unified model. The unified model, which relies on a single decision boundary, appears to capture fewer subgroup-specific patterns, whereas the ensemble leverages race-specific features to inform more nuanced and balanced predictions.

*4.2.2 Mixture of Experts (MoE) Models.* Building upon the insights gained from our ensemble analysis, we developed a Mixture of Experts (MoE) model designed to leverage the strengths of each race-specific classifier. Through prior evaluation of our AlexNet individual-race classifiers, we observed that each racially fine-tuned model achieved its highest performance on faces corresponding to its own subgroup (e.g., Black faces worked best on the Black-trained model). This observation motivated the construction of a MoE architecture, which first determines an image's race through manual input, and subsequently sends the image to the corresponding expert model (Asian, Black, or White) for gender prediction.

Unlike the ensemble model, which combines predictions from all three racial experts through majority voting, the MoE relies on expert selection rather than majority vote. By directing each image exclusively to the most relevant model, the MoE preserves the race-specific optimization achieved during fine-tuning while avoiding potential noise from less suitable models. This design more closely mirrors the ideal use of specialized networks within demographic subdomains.

The MoE architecture demonstrated strong improvements in classification performance over our initial ensemble model. For AlexNet, the MoE achieved 97.58% accuracy, substantially exceeding both the unified and ensemble approaches. These results indicate that when race is input, the MoE can perform at the peak accuracy of the specialized model for that subgroup, effectively maximizing the benefit of racial specialization. However, for ResNet, we found that due to our individual models not necessarily benefitting from the race-specific training (e.g., Black faces had the best accuracy through the White model), the MoE model performs poorly. This finding highlights the importance of model selection based on model architecture, as differing ensemble approaches will deliver distinct results.

**Table 4: Misclassification Confidence Comparison between ImageNet Models.**

| Model | Misclassification Confidence Score |
|---|---|
| ResNet18 - Equal Train | 75.6 |
| ResNet18 - US Train | 73.6 |
| ResNet18 - MoE | 81.0 |
| ResNet18 - Ensemble | **67.1** |
| AlexNet - Equal Train | 73.3 |
| AlexNet - US Train | 75.0 |
| AlexNet - MoE | 71.8 |
| AlexNet - Ensemble | **70.5** |

Interestingly, when evaluating our ensemble model, we found that in addition to achieving higher overall accuracy, the ensemble also produced lower average confidence scores for overall and misclassified images due to aggregating confidences across three individual race models. Since our ensemble models show these larger differences (approximately 8% in ResNet18 and 3% in AlexNet) compared to their unified counterparts, they could better serve as a useful indicator for flagging potential misclassifications, mitigating downstream consequences of errors arising from biased model behavior.

## 4.3 Individual Race Models

This section describes the results from the race-specific models used to produce our ensemble ResNet and AlexNet models (pre-trained on ImageNet). Our general finding is that Black and Asian models performed very well on their race, but had an accuracy trade-off when evaluated on images of each other. However, White accuracy was either first or second regardless of what race models are trained on. Models trained solely on White faces had the highest overall accuracy for ResNet and AlexNet.

**Table 5: Overall and Individual Race Accuracy for Individual Race ResNet18 Models**

| Model | White | Black | Asian | Overall |
|---|---|---|---|---|
| White | **93.00** | **86.25** | **86.75** | **88.67** |
| Black | 87.50 | 84.50 | 83.25 | 85.08 |
| Asian | 88.75 | 80.00 | 83.25 | 86.50 |

*4.3.1 ResNet18 Single Race.* Table 5 shows that across all evaluations, the model trained on White faces achieved the highest overall accuracy, demonstrating strong generalization across racial groups. In contrast, the Asian and Black-trained models exhibited clear tradeoffs: each performed well on its own racial group but showed reduced accuracy when tested on the other. Notably, the White model consistently maintained the top accuracy regardless of the model's trained group. Our GradCam revealed that the White model primarily focused on the mouth and general facial shape, while the Asian model emphasized the nose and eyes, and the Black model concentrated most strongly on the nose region.

**Table 6: Overall and Individual Race Accuracy for Individual Race AlexNet Models**

| Model | White | Black | Asian | Overall |
|---|---|---|---|---|
| White | **98.75** | 81.25 | 86.50 | **88.83** |
| Black | 83.75 | **98.00** | 78.25 | 87.00 |
| Asian | 83.50 | 80.00 | **96.00** | 86.17 |

*4.3.2 AlexNet Single Race.* For AlexNet, each race-specific model achieved its highest accuracy on the race it was trained on, with the White model typically ranking second, except in the White-trained case, where White faces performed best, followed by Asian faces. The tradeoff between the Asian and Black-trained models was less pronounced than in ResNet18, yet the White-trained model again demonstrated the highest overall accuracy across all groups. GradCam analysis indicated that the White model attended most

strongly to the eyes, nose, and mouth; the Asian model emphasized the eyes and facial contour; and the Black model focused primarily on overall facial shape.

*4.3.3 Cross-Architecture Generalizabilities.* In both architectures, the White-trained model consistently achieved the highest total accuracy, possibly reflecting an overrepresentation of White faces in our large-scale pretraining datasets, ImageNet, which includes images from North America and Europe. Both architectures' individual models prioritized differing features, yet they both showed a common trend that the White model had the best overall accuracy. Another notable cross-architecture pattern is that each White model prioritized different features, yet GradCam analysis showed an overlap in the mouth region. This encourages additional exploration into why models trained solely on White images generalize best to other races.

## 4.4 Misclassification Identification

While evaluating model misclassifications across our unified and ensemble models, we observed consistent patterns of image-related and age underrepresentation issues that contributed to higher misclassification probabilities. These are not images that the model will always misclassify, but are images that the model is more likely to struggle with (most of the misclassifications across our models included images with these traits). Problematic images frequently involved children, elderly individuals, individuals wearing glasses, faces that were partially obstructed and/or covered, eyes that were closed or nearly closed, and side-view angles. Examples of each of these types of images are shown in Figure 2.

Using Grad-CAM visualizations, we found that the models primarily focused on facial regions visible in frontal views, such as facial curves and the areas between the eyes, nose and mouth. The problematic images, as described above, often disrupted or obscured these key regions through glasses, squinted eyes, or partial occlusion. Side-view angles prevented the models from effectively outlining features best seen from a frontal view. This is especially visible in Figure 2 where the model is able to focus on facial features in the side view but cannot extract enough data to correctly inform its prediction. Additionally, images of children and the elderly were misclassified more frequently, likely because their facial features deviate from those of the adult faces that primarily make up the training dataset, illustrating the effects of underrepresentation.

We also observed misclassifications linked to behavioral differences across groups that, when reflected in another group, led to biased predictions. For instance, a large proportion of females (especially White) in the dataset were smiling, which caused the models to heavily associate smiling with the female label. This resulted in some male faces that displayed broad smiles being misclassified as female with high confidence as shown in Figure 3.

These misclassification patterns highlight important limitations regarding when and how such models should be applied, emphasizing the need for continued work toward robustness and fairness in facial classification systems. They also provide valuable guidance for dataset curation, as skewed or inconsistent image characteristics can reduce both accuracy and demographic consistency across models. Analyzing the types of images a model misclassifies helps identify cases where human oversight is needed, ensuring that errors with potentially serious consequences are detected and mitigated during deployment.

## 4.5 Filtering Misclassification Types from the Test Set

To improve evaluation consistency, we removed test images containing children, elderly individuals, subjects wearing glasses, side profiles, closed eyes, or other facial obstructions. The dataset size was normalized across demographics to maintain balance. Notably, FairFace contained a higher proportion of removed samples compared to UTKFace, reflecting UTKFace's stricter criteria for frontal headshots. We used this new cleaned test set to evaluate the ResNet18 and AlexNet model variants that were initially pre-trained on ImageNet.

**Table 7: ResNet18 Models Performance on Cleaned Test Set**

| Model | White | Black | Asian | Overall |
|---|---|---|---|---|
| Asian | 98.18 | 85.45 | 97.27 | 93.64 |
| Black | 97.27 | 95.45 | 90.91 | 94.55 |
| White | 98.18 | 92.73 | 96.36 | 95.76 |
| Ensemble | 99.09 | 90.91 | 97.27 | 95.76 |
| U.S. Train | 99.09 | 96.36 | 93.64 | 96.36 |
| MoE | 98.18 | 95.45 | 97.27 | 96.97 |
| Equal Train | 98.18 | 97.27 | 96.36 | **97.27** |

*4.5.1 ResNet18 Results.* As seen in Table 7, we found that overall and individual race accuracy increased across all race-specific models, unified models, the MoE, and the ensemble model after refining the test data. Compared to our original test set, the ensemble model did not perform the best and was surpassed by the MoE and Equal Unified model in overall accuracy. However, the ensemble maintained its characteristic trend of lower misclassification confidence, with a difference of approximately 12% compared to the unified models. Interestingly, White classification accuracy remained the highest for all model variants.

**Table 8: AlexNet Models Performance on Cleaned Test Set**

| Model | White | Black | Asian | Overall |
|---|---|---|---|---|
| White | 99.09 | 85.45 | 93.64 | 92.73 |
| Asian | 96.36 | 84.55 | 100.00 | 93.64 |
| Black | 95.45 | 99.09 | 89.09 | 94.55 |
| U.S. Train | 97.27 | 92.73 | 93.64 | 94.55 |
| Equal Train | 97.27 | 90.91 | 96.36 | 94.85 |
| Ensemble | 99.09 | 89.09 | 97.27 | 95.15 |
| MoE | 99.09 | 99.09 | 100.00 | **99.39** |

*4.5.2 AlexNet Results.* Similarly for AlexNet, we found that overall and individual race accuracy increased across all race-specific models, unified models, the MoE, and the ensemble model following test set refinement. However, unlike ResNet18, the ensemble and MoE models continued to achieve the highest overall performance

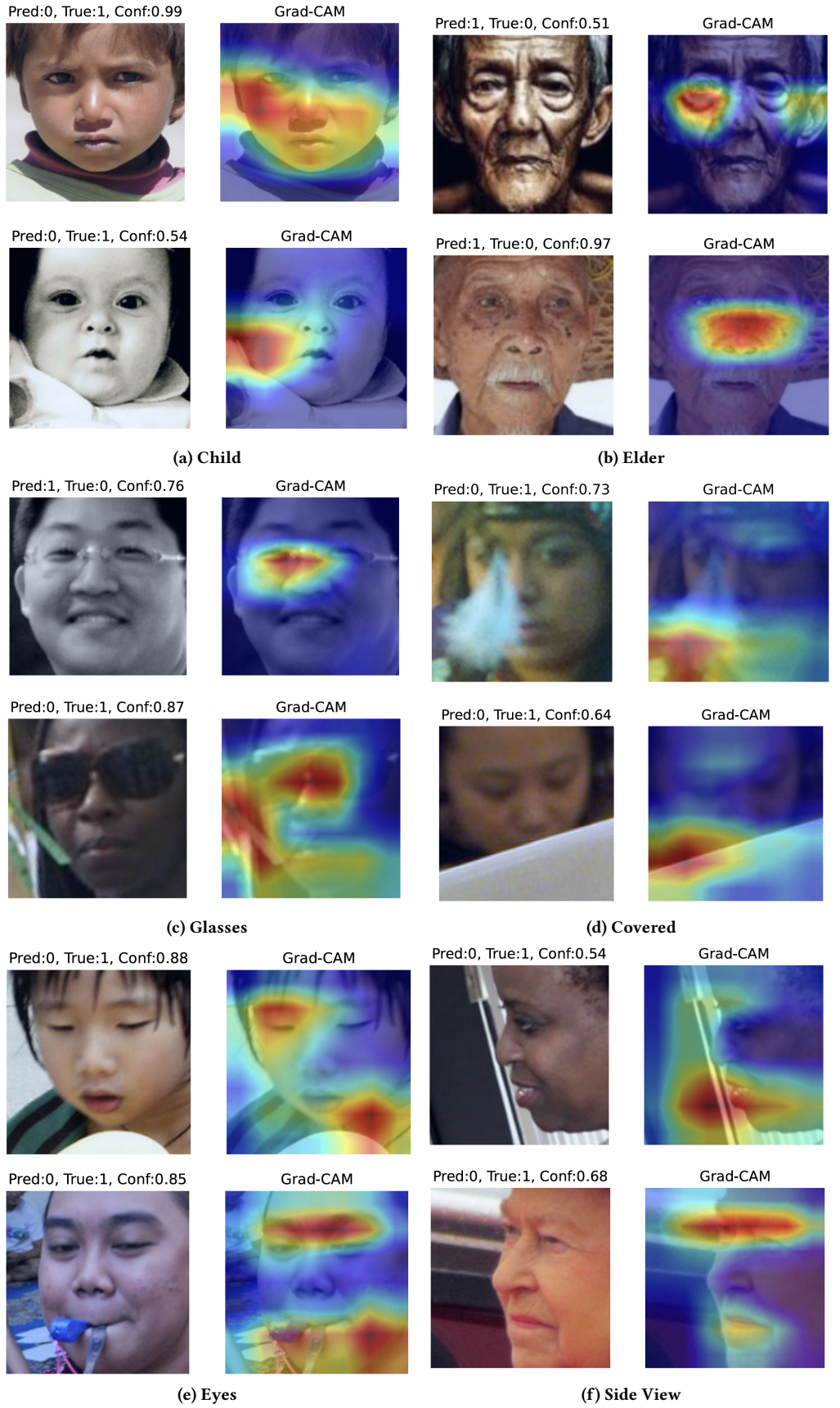
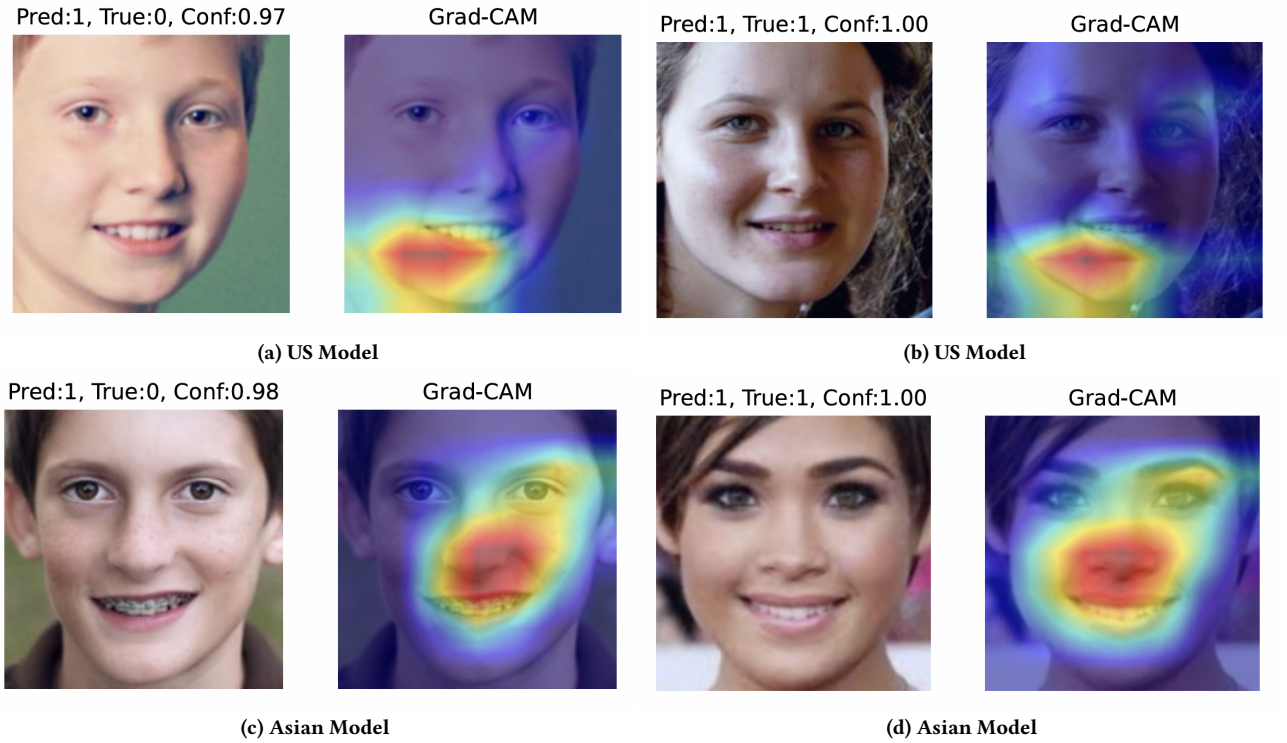



**Figure 2: Examples of misclassifications on certain characteristics. [0: male] [1: female]**

**Figure 3: Focuses on smile across different models.**




**(a) US Model** **(b) US Model**

**(c) Asian Model** **(d) Asian Model**

as compared to our unified or race-specific models. The MoE in particular continued its exceptional performance, rising to an accuracy of over 99%. Similar to ResNet18, our misclassification confidence trend persisted for the ensemble, with the ensemble's confidence remaining approximately 5% lower than that of unified models. Interestingly, the MoE misclassification confidence was also around 3% lower than that of the unified models.

*4.5.3 Big Picture Takeaways.* These results illustrate that low-quality or ambiguous data negatively affects performance across all groups, rather than disproportionately impacting any single subgroup. Both the ResNet and AlexNet results also confirm our expectation that these misclassification types are harmful to classification. A perplexing finding is how the models trained solely on Asian or Black faces, still have the highest classification accuracy for White faces on ResNet. Simply training on only White faces achieves an overall accuracy equivalent to that of the ensemble model for ResNet as well. This opens the door for further exploration into why White images are easier to classify, since for both models, the White accuracy is very high. Perhaps it is a data quality discrepancy issue between races or the fact that models are noticing key White facial features that hold higher impact in classification for any race.

It is worth noting that the size of our test set shrunk significantly (1,200 images to 330) after removing images with misclassification types from our test set. In the future, it is worth explore what results would be on a more expansive set of data.

# 5 Conclusion

## 5.1 Key Takeaways

Our results show that dataset composition and model depth play critical roles in gender classification performance and fairness. Fine-tuning on an equally distributed dataset improved overall accuracy and reduced bias across demographic groups compared to training on a racially imbalanced dataset. The only exception was ResNet-18 pretrained on the IMDB dataset, likely due to its depth and the IMDB dataset's bias toward White faces. In contrast, AlexNet benefited more from balanced fine-tuning, suggesting that shallower architectures generalize better under equitable data conditions.

Ensemble models combining race-specific classifiers achieved higher accuracy, lower misclassification confidence, and faster training times than unified models. The accuracy improvement was greater for AlexNet, potentially reflecting its sensitivity to increased representational diversity, while deeper networks like ResNet-18 saw smaller improvements. These ensemble models also helped mitigate subgroup bias by leveraging race-specific patterns to inform more balanced predictions.

A Mixture of Experts (MoE) model, which routes each image to its corresponding race-specific classifier based on identified race, achieved the highest overall accuracy for AlexNet (99.39% on clean set). The AlexNet MoE also had lower misclassification confidence, comparable to that of the ensemble model. This makes the AlexNet MoE the most effective approach for facial gender classification in our experiments, provided that race can be accurately identified prior to inference. It also is the most fair model due to achieving

the highest individual accuracy for each race. In contrast, the MoE did not work as well for ResNet-18. These results show that the success of MoE architectures is highly dependent on model depth and design, reinforcing the need to align classification strategies with the underlying model architecture. The difference in accuracy pre and post-data cleaning also points to the importance of good data.

Additional analyses revealed consistent misclassification trends tied to image quality and representation, particularly among children, the elderly, and faces with obstructions or side views. Additionally, behavioral patterns, such as frequent smiling among White females, contributed to biased predictions across groups. Overall, our findings highlight the importance of balanced datasets, careful curation of image diversity and quality, model selection and layer depth, and the use of ensemble methods to improve fairness and reliability in facial classification systems.

## 5.2 Future Work

Future work should further examine the sources of misclassification identified in this study. While we observed that certain image characteristics, such as side profiles, facial occlusions, and age-related facial features, were more likely to be misclassified, we cannot conclusively determine whether these factors directly cause the observed errors. Systematic testing across controlled datasets could help isolate these effects and clarify their impact on model performance. The higher accuracy observed for the White subgroup compared to Asian and Black groups, even after the removal of misclassified images, warrants further investigation. Additionally, cleaning the training data, rather than only the testing data, could provide deeper insight into how data quality influences overall model accuracy and GradCAM feature localization during evaluation.

The benefits of ensemble and MoE models composed of race-specific models in our gender classification experiments encourage exploring their impact for other classification tasks. It would be worthwhile to see if the increasing gap in misclassification confidence persists in similar tasks.

# 6 References

# References

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*. PMLR, 77–91.

Valeriia Cherepanova et al. 2022. A Deep Dive into Dataset Imbalance and Bias in Facial Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4651–4660.

USA Facts. 2024. Our Changing Population. https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/. Accessed: 2025-10-05.

Max Gwilliam, Sam Dobson, Victor Sanchez, and Nicolas Pugeault. 2021. Rethinking Common Assumptions to Mitigate Racial Bias in Face Recognition Datasets. In *British Machine Vision Conference (BMVC)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

HuggingFaceM4. 2024. *FairFace Dataset.* https://huggingface.co/datasets/HuggingFaceM4/FairFace Accessed: 2025-10-10.

Jangedoo. 2019. *UTKFace - New Dataset.* https://www.kaggle.com/datasets/jangedoo/utkface-new Accessed: 2025-10-10.

Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913* (2019).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NeurIPS)*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv preprint arXiv:1610.02391* (2016).

Yuulind. 2023. *IMDb Clean Dataset.* https://www.kaggle.com/datasets/yuulind/imdb-clean Accessed: 2025-10-28.