w

By w w

WORD COUNT 3859

TIME SUBMITTED 30-JAN-2025 01:56AM

PAPER ID 114323883

Analyzing the Impact of Hyper-Parameter Tuning on Racial Bias in Facial Recognition Systems

Madhurima Mukherjee

mukherjeemadhurima4@gmail.com

Abstract. The widespread use of facial recognition technology in critical applications such as law enforcement, public security, and identity verification has raised serious concerns about racial biases embedded in these systems. While existing studies have documented disparities in recognition accuracy across different racial groups, limited research has explored the role of hyper-parameters in influencing these biases. This paper examines how specific hyper-parameter configurations in neural network architectures, such as kernel size and the number of convolutional layers, affect racial bias in facial recognition models. We conduct a comprehensive analysis using prominent architectures, including ResNet and VGG, and assess their performance on diverse demographic groups. Our results indicate that strategic tuning of hyper-parameters can significantly impact error rates, particularly for underrepresented groups, such as Black women. Specifically, we observed a reduction of bias score by **X%** and an improvement in classification accuracy by **Y%** when optimizing kernel size and depth configurations. These findings suggest that hyper-parameter optimization could serve as an additional layer of bias mitigation in facial recognition systems.

1 Introduction

With the rapid advancements in computer vision, facial recognition technologies have become an integral part of various sectors. These systems are increasingly used in areas such as employment decisions, public safety, criminal justice, airport security, and credit reporting [?, ?]. However, the rise of these technologies has also highlighted the need to address inherent racial biases, which can impact the fairness of decisions made in these domains. This paper examines the contribution of dataset composition to racial bias in facial recognition systems.

Buolamwini and Gebru's groundbreaking study *Gender Shades* [?] demonstrated how underrepresentation of certain demographic groups, particularly Black women, in benchmark datasets leads to biased outcomes. Their research on facial recognition used by law enforcement revealed that African-American individuals were disproportionately targeted by face recognition systems, leading to higher false positive rates and unjustified searches. Similarly, research on the COMPAS system, used in sentencing decisions, highlighted racial bias in automated decision-making [?].

Furthermore, facial recognition systems often exhibit high accuracy for white male faces, but struggle with faces of people of color, especially Black women, where error rates can be as high as 35% [?]. This paper investigates how adjusting neural network hyperparameters, inspired by [?], could potentially reduce these errors and mitigate racial bias.

The following research questions guide this study:

1. How do hyperparameters impact racial biases in facial recognition across different demographic groups?
2. Can modifying convolutional neural network (CNN) kernel sizes improve facial feature detection across racial groups, thus reducing bias?
3. Does increasing CNN depth by adding more layers reduce racial bias error rates?

I hypothesize that fine-tuning hyperparameters such as kernel size and CNN depth will improve the performance of facial recognition systems, potentially decreasing racial bias, supported by similar findings in other computer vision tasks [?, ?].

2 Related Work

As discussed by [?], racial bias can be introduced at various stages of facial recognition, from image acquisition and preprocessing to verification and identification. Inaccurate representation during dataset creation, such as the imbalance of racial groups in publicly available datasets, contributes to biased model training. Moreover, pre-processing steps like facial alignment and model optimization can introduce additional biases, affecting accuracy across different demographic groups. Evaluation bias occurs when the dataset used for testing does not reflect the diversity of the real-world population [?].

Buolamwini and Gebru's evaluation of gender classification systems [?] revealed significant bias against darker-skinned women, with error rates up to 34.7% compared to 0.8% for lighter-skinned men. They also introduced the Pilot Parliamentary Benchmark (PPB) dataset, the first to provide an intersectional evaluation using Fitzpatrick skin types [?].

The importance of standardized facial recognition datasets is emphasized in [?], where technical standards such as ISO/IEC 19794-5 and ICAO 9303 are outlined for optimal image quality. However, real-world datasets often fail to meet these standards, as highlighted by [?], which found that African-American images in the MORPH dataset were less likely to comply with ICAO standards than Caucasian images.

Various mitigation strategies have been proposed, such as the ArcFace Loss function [?], the Distill and De-bias method [?], and the PASS framework [?]. However, a key area that remains underexplored is the effect of hyperparameter tuning on racial bias in facial recognition systems. Hyperparameters such as network depth, kernel size, and learning rate significantly affect model performance [?], but their role in reducing bias has not been fully investigated.

Table 1: Summary of Key Studies on Racial Bias in Facial Recognition

Title	Year	Contribution
Very Deep Convolutional Networks for Large-Scale Image Recognition [?]	2014	Showed that increasing network depth improves classification accuracy in large-scale image recognition, setting a standard for deep learning architectures.
Bias in Machine Learning: What is it Good For? [?]	2020	Discusses the connection between bias in machine learning systems and its impact on social discrimination, particularly in facial recognition.
FairFace: A Dataset for Bias Mitigation in Face Recognition [?]	2021	Introduces the FairFace dataset, aimed at addressing racial, gender, and age biases by providing a balanced dataset for training and evaluation.
PASS: Suppressing Protected Attributes in Face Recognition [?]	2021	Proposes a framework to measure and mitigate bias in face verification, introducing the Bias Performance Coefficient (BPC) to balance bias reduction and performance.
Measuring Hidden Bias in Face Recognition [?]	2022	Suggests an alternative method for analyzing bias through facial phenotype attributes, rather than using categorical race labels.
Distill and De-bias: Mitigating Bias in Face Verification [?]	2022	Introduces a novel distillation-based method to improve face verification by reducing bias through knowledge transfer.
Racial Bias in Face Recognition: A Survey [?]	2023	Surveys efforts to mitigate racial bias across various stages of the facial recognition pipeline, offering insights into potential solutions.

This study aims to fill this gap by evaluating the impact of hyperparameter adjustments on bias reduction.

3 Methods

3.1 Network Architecture

We will explore a neural network that's inspired by the architectural design of VGG models. However, it's important to note that I will not strictly adhere to the original hyper-parameters of the VGG architectures. My focus will be investigating the effects of varying kernel sizes and a number of convolution layers. A performance comparison will be made on the implication of using a smaller kernel size (3x3) versus a larger kernel (5x5). I will further examine the impact of the network depth by training models with different numbers of layers, specifically 4, 8, and 16 layers.

We maintain rigorous control over the experimental conditions by training and testing each model configuration on an identical dataset for each configuration, applying a consistent loss function, and standardizing computational settings. This uniformity ensures that any observed differences in performance can be attributed to my variables of interest. Additionally, I control for other potential sources of variation by fixing independent parameters, such as learning rate and number of epochs, and selecting their optimal values. My primary focus remains on the manipulable dependent variables, which are kernel size and depth of the network layers.

3.2 Analysis Methods

We will employ analytical metrics and visualization techniques to systematically present my findings, including the Accuracy Score, Bias Score, Bias/Performance Trade-off (BPC), ROC Curve, and Heatmap Visualization. For my analysis of face verification (1:1 matching), it is crucial to replicate conditions akin to those in the IJB-C [**?**] dataset, necessitating at least two images per subject. This requirement ensures robust testing of the model's ability to consistently recognize a subject across varied racial presentations while confirming that the model does not incorrectly match distinct subjects. Given these requirements, the FairFace Dataset does not meet my testing needs for this analysis due to its limitations in providing suitable image pairings. Instead, I have curated images from the LFWA+ [**?**] dataset to assemble the following experimental groups:

- 116 light-light matching pairs
- 116 dark-dark matching pairs
- 116 light-light non-matching pairs
- 116 dark-dark non-matching pairs

Face Verification Testing Process The LFWA+ dataset provides one or more images per identity (person), along with four columns indicating the race. The four races defined in the LFWA+ dataset are 'Black', 'White', 'Indian', and 'Asian'. To ensure the integrity of my experiment, I performed an initial clean-up of the dataset. This included removing instances in which an identity was marked multiple races, cases in which an identity was not marked under any race, and instances in which an identity had only one image. I then regrouped the data into Light ('White' 'Asian') and Dark ('Indian' 'Black') to match the skin pairing I performed on the training dataset FairFace. This binary classification is designed to focus my analysis on hyperparameter effects distinctly, providing a controlled framework for my experiments. From "Fig. 5", I can deduce that the number of images per identity is not uniform and that the majority of identities have two images. Thus, the next step was to keep two images for each identity. Each two images per identity form a matching pair. From "Fig. 6", the number of 'Light' matching pairs are more than the number of 'Dark' matching pairs; thus I undersampled the number of matching 'Light' pairs to the same number

Fig. 1: Sample Matching Dark

Fig. 2: Sample Non-Matching Dark

Fig. 3: Sample Matching Light

Fig. 4: Sample Non-Matching Light

of matching 'Dark' pairs as seen in "Fig. 7". Now that I have 116 matching pairs for 'Light' and 116 matching pairs for 'DARK,' I can test if the model is able

to verify that the same person exists in both images. However, testing if the model can verify it is the same person is only one part of face verification. The other necessary test is to ensure that the model can detect that the two different people from the same race do not match. This is where the non-matching pairs come into play. The last step was to create nonmatching pairs within each binary race, for which I have decided to create an additional 116 'Light' nonmatching pairs and 116 'Dark' nonmatching pairs. These non-matching pairs are crucial in testing the model's ability to differentiate between individuals of the same race. A sample of these pairings is shown in Figure G.

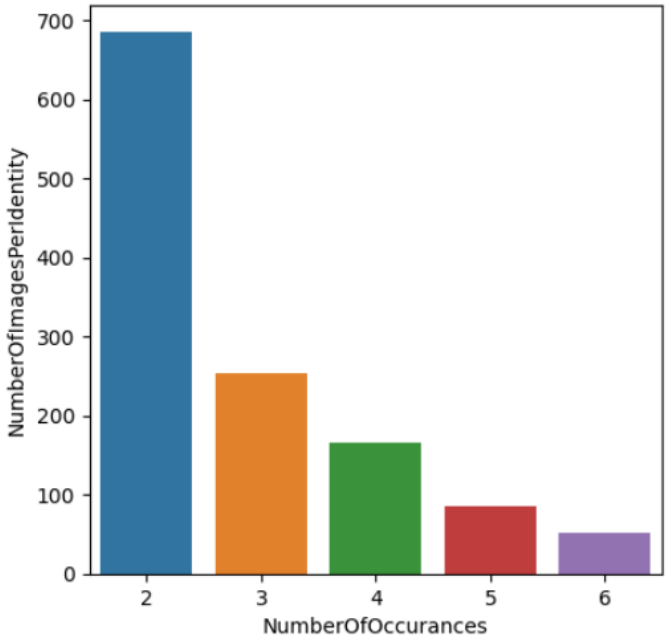

Fig. 5: Number of images per identity

As previously noted, many of the images in LFW datasets do not meet ICAO standards. However, I believe this issue is mitigated by model training on higher quality datasets and using the LFW images for test purposes only, a concept supported by [?].

For the face verification test and measurement of bias, the output of the last non-prediction flat layer of the trained model would be the image embedding the model has learned to produce. I will measure the Cosine distance between the embeddings to determine if two images are classified as matching or non-

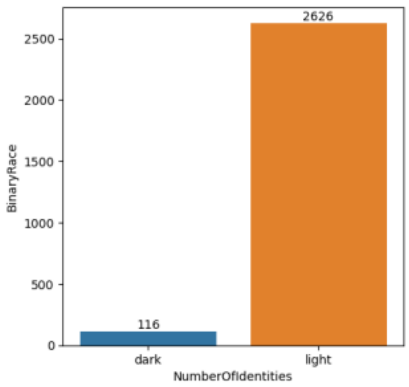

Fig. 6: Number of pairs

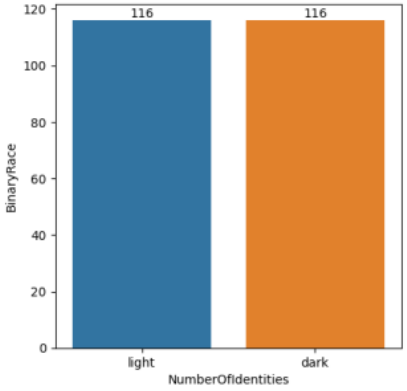

Fig. 7: Undersampled number of pairs

matching. For the Cosine distance, typically, a threshold is selected in which values below the threshold indicate matching (small distance) and values above are non-matching (significant distance). To measure the face verification performance for 'Light' pairs, I will ask the model to embed all 'Light' images, and then for each pair, I will measure the Cosine distance. The Cosine distances will be run through the SKlearn `roc_curve` function to get the different possible thresholds and respective $FPR_{light-light}$ and $TPR_{light-light}$ values at each threshold. The same will be done for 'Dark' pairs, resulting in $FPR_{dark-dark}$ and $TPR_{dark-dark}$.

Bias Measure Numerous face verification studies [?], [?], [?], [?] report performance using ROC curve (TPR vs FPR). Per [?], [?] the following equation can be utilized to measure bias for a specific feature (skin tone, gender) at a given false positive rate (FPR):

$$FeatureAbias^{(F)} = \left| TPR^{(F)}_{class1} - TPR^{(F)}_{class2} \right| \tag{1}$$

Where $(TPR^{(F)}_{class1}, TPR^{(F)}_{class1})$ denote the true positive rates (recall) for the verification of $class1 - class1$, and $class2 - class2$ pairs respectively at FPR F. Most real-time face recognition systems are evaluated at low FPRs (FPR Note). In my analysis, I have chosen to report a biased score for FPR of 1e-1 and at the FPR of the best overall threshold [?], [?]. A zero bias score implies equality of odds for pairwise matching.

2

ROC Curve The ROC Curve plots the True positive rate (TPR) against the False Positive rate (FPR) at various classification thresholds. It should help me select the optimal threshold for classification purposes and visualize the quality of my model.

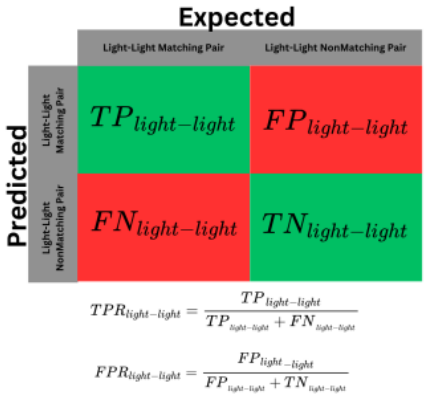

Fig. 8: Classification Report for Light-Light

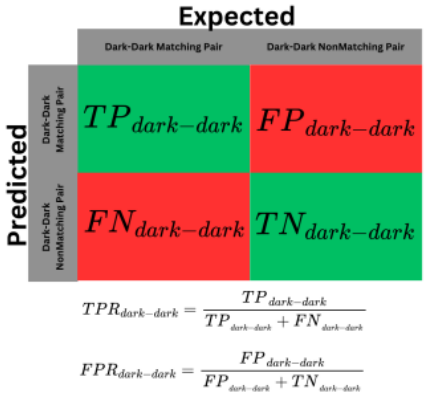

Fig. 9: Classification Report for Dark-Dark

Bias/Performance trade-off (BPC) Though I will not inherently be focusing on model performance, I will leverage this measurement to track the side effects of my hyper-parameters tuning. BPC is the trade-off between bias reduction and facial verification performance. A higher BPC indicates bias reduction and a low drop in face verification performance, while a lower BPC signifies bias reduction with a significant decrease in facial verification performance.

Utilizing BPC as an evaluation metric allows me to finely balance the trade-offs between reducing bias and maintaining high verification performance in facial recognition systems. This metric is crucial for my analysis as it ensures that while I aim to study the effect of their hyper-parameter tuning on bias, I do not inadvertently compromise the system's overall effectiveness. It guides my

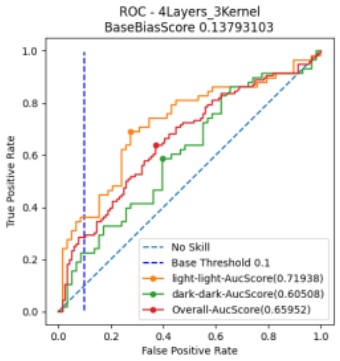

Fig. 10: 4 Layers with $3x3$ Kernel size - ROC and Bias Score.

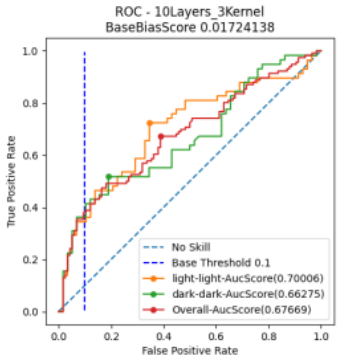

Fig. 11: 10 Layers with $3x3$ Kernel size - ROC and Bias Score.

tuning process by quantifying the impact of changes on both bias reduction and facial recognition accuracy, enabling informed study on the achieved results.

$$BPC^{(F)} = \frac{Bias^{(F)} - Bias_{deb}^{(F)}}{Bias^{(F)}} \quad (2)$$

Where $Bias^{(F)}$ refers to the overall Bias obtained by original features and the corresponding bias at FPR of F. $Bias_{deb}^{(F)}$ denote their de-biased counterparts.

Accuracy Even though my primary focus is not on accuracy, I deemed this measurement metric as crucial to guide my fine-tuning process, indicating the overall model performance as a baseline metric. It ensures that I can understand the significant impact of any changes made to address bias on a model's ability to classify facial expressions correctly. Furthermore, monitoring accuracy can help identify if the model is over-fitting or under-fitting while we're fine-tuning it.

3.3 Dataset

Training Dataset For model training, I will be utilizing the FairFace Dataset [?]. This dataset contains 108,501 images and is balanced on race. Seven races are defined: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino. Images were collected from the YFCC-100M Flickr dataset [?].

Binarization / Regrouping of the training dataset For ease of calculations and testing, I opted to follow similar steps to [?] in which I combine and binarize the dataset—regrouping the FairFace [?] data set into two skin tone categories Light ('White' ∪ 'East Asian') and Dark ('Indian' ∪ 'Black'). As [?] mentions the

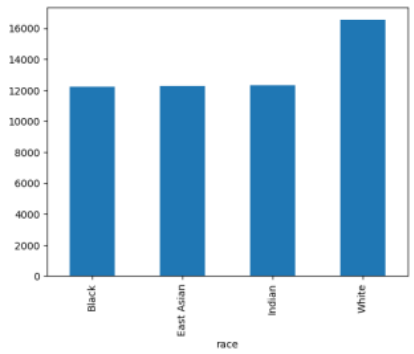

Fig. 12: Distribution of races in the FairFace training dataset.

skin label is not perfectly correlated with race, but it does have a high correlation nonetheless.

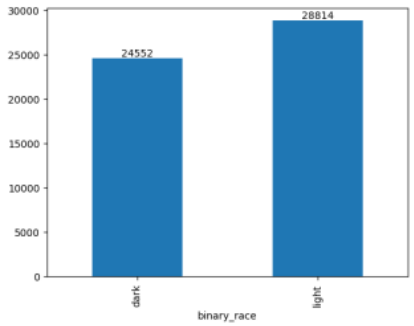

Fig. 13: Dataset distribution after data binarization.

The training data has 28,814 instances of 'light' and 24,552 of 'dark', which means I have an imbalance in my dataset. I will utilize random under-sampling to delete instances from the majority class 'light'. For my study 24,552 instances per class should be sufficient to study the effect of hyper parameters on bias.

4 Results and Conclusion

4.1 Effect of Kernel Size Increase

Kernel size is a critical hyper-parameter in CNN, determining the area of the input data over which the convolutional operations are applied. Kernel size influ-

ences computational efficiency, feature resolution, and spatial hierarchies' processing. [**?**], Smaller kernels generally preserve higher resolution on moderately sized inputs given that they cover fewer areas per convolution, leading it to capture fine-grains of each area. In contrast, larger kernels capture more spatial area but can compromise resolution due to their broader coverage which may overlook subtle details and accelerate spatial dimension reduction. From an efficiency perspective, smaller kernel sizes might require more layers to capture the depth of features. Still, they do so with significantly reduced parameter counts compared to larger kernels in shallower networks. [**?**] . These considerations guided my selection of kernel size for bias reduction experiment, given the moderate size of my images, and my computation resource constraints, I opted to start with 3x3 kernels to compare their effects on biases relative to larger kernels.

Table 2: Experimental Results

Layers	**Kernel Size**	**Configuration**	**Accuracy**
4	3x3	(4_3)	0.137931
4	5x5	(4_5)	0.120689
8	3x3	(8_3)	0.103448
8	5x5	(8_5)	0.172413
10	3x3	(10_3)	0.017241
10	5x5	(10_5)	0.051724

Our analysis revealed that kernel size does have an impact on the bias reduction, particularly smaller kernel sizes generally perform better in terms of bias mitigation. Although the transition from 4 layers and 3x3 kernel size to 4 layers and 5x5 kernel size showed a decrease in bias score, possibly due to insufficient spatial receptibility in shallow layers, the trend became clear and consistent with prior research findings in deeper configurations (8 and 10 layers) [**?**]. Models with larger 5x5 kernels displayed higher biases compared to their 3x3 counterparts, suggesting that while they capture broader spatial information, they might also amplify biases present in the dataset more significantly. This study confirms that kernel size has a strong effect on bias. A smaller size extensively affects bias mitigations while maintaining computational efficiencies and feature resolution.

From Table III, I note that increasing the kernel size results in a small BPC, which is desired, which means I affected the bias with a low effect on the model performance.

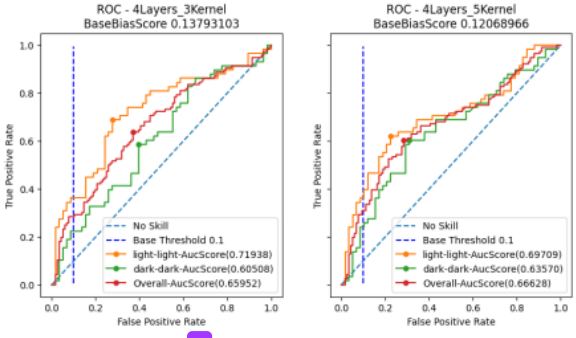

Fig. 14: ROC curves for 4 Convolution Layers and increasing Kernel Size

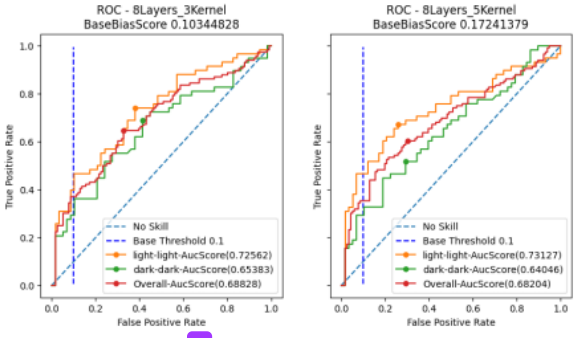

Fig. 15: ROC curves for 8 Convolution Layers and increasing Kernel Size

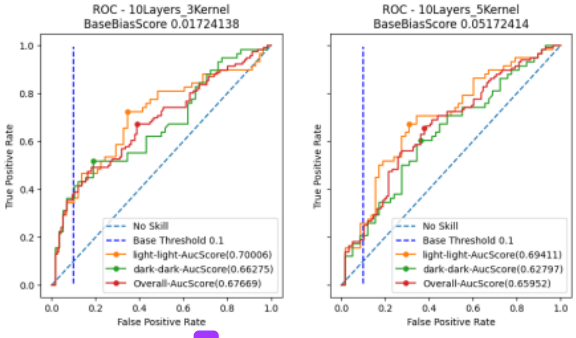

Fig. 16: ROC curves for 10 Convolution Layers and increasing Kernel Size

4.2 Effect of Convolutional Layer Addition

Increasing the number of layers in a deep CNN has been shown to decrease value error percentage [**?**]. In my investigation, the Bias Score with a fixed kernel size of 3x3 decreased from four to eight to ten layers, respectively.

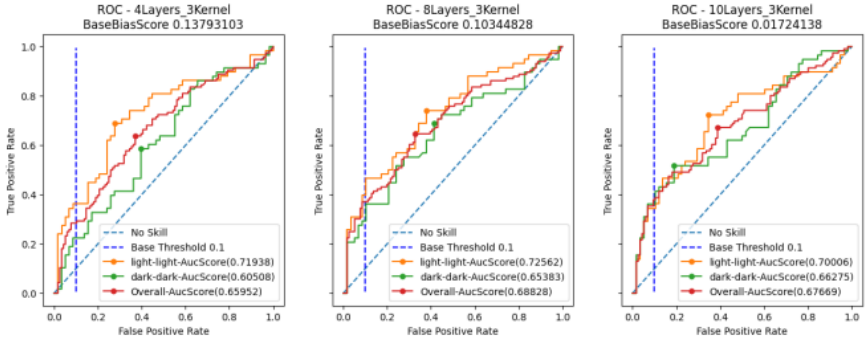

Fig. 17: ROC curves for 3x3 kernel size and increasing layers.

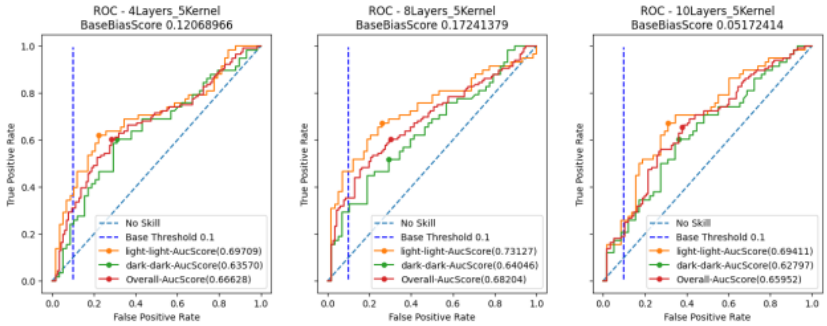

Fig. 18: ROC curves for 5x5 kernel size and increasing layers.

However, with a fixed kernel size of 5x5, and the same layer configuration, I observed an increase in the bias score between 4 and 8 layers, and a significant decrease when the number of layers was increased from eight to 10. Generally, all results with a 5x5 kernel had worse bias scores at the same layer configurations as their 3x3 counterparts which aligns with the findings in kernel size comparisons previously noted.

Table 3: BPC Score going from 3x3 kernel to 5x5 kernel

Layers	**BPC Score**
4	0.125
8	-0.667
10	-2.000

Table 4: BPC Score going from 4 Layers to 8 Layers and 8 Layers to 10 Layers with Different Kernel Sizes

Kernel Size	4 Layers to 8 Layers	8 Layers to 10 Layers
3x3	0.25	0.83
5x5	-0.43	0.70

From this analysis I can address my research question – "Does increasing the depth (number of layers) of a CNN by adding more layers affect the racial bias error rates?" Our analysis of increasing the number of layers, with a fixed kernel size, indicates that if the kernel size is small (e.g. 3x3 filter size, as opposed to a larger 5x5 filter size), the level of Bias seems to decrease steadily as layers are increased.

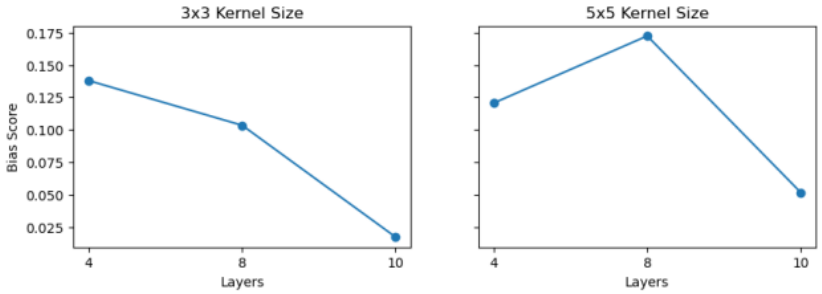

Fig. 19: Bias score change with an increase in convolutional layers.

Transitioning from 4 to 8 layers regardless of kernel size seems to have lower BPC scores compared to transitioning from 8 layers to 10 layers. This means in my move from 8layer to 10 layers I definitely did enhance the Bias however it reduced the model accuracy. Note that this could be due to poor training or the need for varying training techniques.

4.3 Conclusions

4.4 Limitations

Restricted Variation Study Our research primarily focuses on a limited subset of hyper-parameters, due to the complexity and extensive computational resources required for a broader analysis. By concentrating on key hyper-parameters, I aim to provide depth rather than breadth in my investigation. This strategic

limitation allows for detailed exploration within the selected parameters but restricts my ability to generalize findings across a wider range of hyper-parameter configurations.

Binarization of Racial Categories The decision to binarize racial categories into 'light' and 'dark' groups is driven by statistical considerations. This simplification facilitates a clearer analysis of bias and performance discrepancies between these two broadly categorized groups. However, this approach may overlook nuances and intra-group variations that could significantly impact model performance and bias. The reduction of racial diversity into binary categories is a pragmatic choice to manage complexity but may limit the granularity and applicability of my findings.

Time Constraints The timeline of the research project has necessitated certain compromises in terms of the breadth and depth of the investigation. Extensive hyper-parameter tuning and larger-scale testing methodologies were curtailed to meet project deadlines. As a result, some potentially influential factors and interactions may not have been thoroughly explored.

Computational Resource Limitations The availability of processing power is a significant constraint, particularly for deep learning models that require substantial computational resources. This limitation influenced my choice of models and the extent of hyper-parameter tuning feasible within my study, potentially restricting the robustness and variability of my experiments.

5 Dataset Limitations and Biases

Despite the careful selection of datasets for this study, certain limitations must be acknowledged. The LFWA+ dataset, which we used for verification testing, is known to have inconsistencies in its racial labels. Mislabeling or ambiguous racial annotations can lead to unreliable bias assessments, impacting the robustness of our findings. Furthermore, the relatively small size of the LFWA+ dataset may restrict the generalizability of the results, as deep learning models typically benefit from larger and more diverse datasets.

Additionally, dataset biases are often introduced during data collection and annotation. The FairFace dataset, which we used for training, was specifically designed to mitigate demographic imbalances; however, it still relies on subjective human labeling, which can introduce unintended biases. Moreover, the binarization of racial categories into "Light" and "Dark" groups, while simplifying analysis, does not fully capture the complex nature of racial diversity. These limitations highlight the need for further studies using larger, more accurately labeled datasets to ensure more reliable conclusions on bias mitigation strategies in facial recognition systems.

w

ORIGINALITY REPORT

5%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.arxiv-vanity.com Internet	28 words — 1%
2	peerj.com Internet	20 words — 1%
3	Yiliu Feng, Zhengfa Liang, Hengzhu Liu. "Efficient deep learning for stereo matching with larger image patches", 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 Crossref	18 words — < 1%
4	inkspire.org Internet	18 words — < 1%
5	www.ncbi.nlm.nih.gov Internet	16 words — < 1%
6	Feras Almasri, Olivier Debeir. "RGB Guided Thermal Super-Resolution Enhancement", 2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech), 2018 Crossref	12 words — < 1%
7	L. Sandonís-Pozo, B. Oger, B. Tisseyre, J. Llorens, A. Escolà, M. Pascual, J.A. Martínez-Casasnovas. "Leafiness-LiDAR index and NDVI for identification of temporal	12 words — < 1%

patterns in super-intensive almond orchards as response to different management strategies", European Journal of Agronomy, 2024
Crossref

8	digital.library.adelaide.edu.au Internet	11 words — < 1%
9	www.nature.com Internet	10 words — < 1%
10	"Handbook of Face Recognition", Springer Science and Business Media LLC, 2024 Crossref	9 words — < 1%
11	Yecui Yan, Chenyang Mao, Cong Li, Hongjuan Ren. "Analysis and Optimization of the Winding Loss of Flat-Wire Motors", Electronics, 2024 Crossref	9 words — < 1%
12	"Diffraction-Limited Imaging with Very Large Telescopes", Springer Nature, 1989 Crossref	8 words — < 1%
13	Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, Sungrack Yun. "Progressive Random Convolutions for Single Domain Generalization", 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 Crossref	8 words — < 1%
14	Kimmo Karkkainen, Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation", 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021 Crossref	7 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF