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. I conduct a comprehensive analysis using prominent architectures, including ResNet and VGG and assess their performance on diverse demographic groups. my results indicate that strategic tuning of hyper-parameters can significantly impact error rates, particularly for underrepresented groups, such as Black women, thereby contributing to more equitable performance. These findings suggest that hyper-parameter optimization could serve as an additional layer of bias mitigation in facial recognition systems.

1 Motivation

As advancements in computer vision increase, the application of facial recognition becomes more important. Today, a multitude of use cases impact my daily life. Decisions involving employment, public security, criminal justice, law enforcement surveillance, airport passenger screening, and credit reporting [1, 11] are just a few examples. The importance of understanding the underlying drivers of racial bias contributors is imperative to ensuring racial equity is present in them. I present several examples supporting the claim that the composition of data in many datasets is a contributing factor to racial bias.

Joy Buolamwini and Timnit Gebru, in their seminal paper *Gender Shades* [2], found that "a demographic group that is underrepresented [e.g., Black Females] in benchmark datasets can nonetheless be subjected to frequent targeting." Their study of automated face recognition by law enforcement further illustrates the need for improvement. "A year-long research investigation across 100 police departments revealed that African-American individuals are more likely to be stopped by law enforcement and be subjected to face recognition searches than individuals of other ethnicities. False positives and unwarranted searches pose a threat to civil liberties."

Another example, Thomas Hellstrom, Virginia Dignum, and Suna Bensch describe COMPAS, a computer program used for bail and sentencing decisions, "has been labeled biased against black defendants [9]."

Examples of such partiality abound. Nada Hassanin quotes faculty of Law assistant professor Dr. Gideon Christian at the University of Calgary in her article [8], "In some facial recognition technology, there is over 99% accuracy rate in recognizing white male faces. But, when it comes to recognizing faces of color, especially the faces of Black women, the technology seems to manifest its highest error rate, which is about 35 percent."

In this paper, I intend to explore methods inspired by [14] and analyze to determine whether the effects of hyper-parameter tuning on neural network architectures could potentially decrease some of the large error rate gaps identified in most facial recognition systems.

Through this investigation, I seek to answer the following research questions:

1. How do specific hyper-parameters influence racial biases in facial recognition, particularly across different racial groups?
2. Can modifications in the kernel size of a convolutional neural network (CNN) improve the detection of facial features across diverse racial groups, thus reducing bias?
3. Does increasing the depth (number of convolutional layers) of a CNN by adding more layers affect the racial bias error rates?

Further, I hypothesize that strategic adjustments in hyper-parameters, such as kernel size and the number of convolutional layers, can significantly influence the performance of facial recognition systems, potentially reducing racial bias. This hypothesis is supported by studies of the overall effect of hyper-parameters on other computer vision tasks such as, [18] and [19].

2 Related Work

[25] Surveys multiple papers written until May of 2023 regarding racial bias within face recognition and efforts to mitigate it at each stage of the pipeline. Various means exist by which bias enters facial recognition models. Acquisition, the stages of initial image pre-processing and formulation from online resources which could introduce sampling bias; Face Localization, a facial alignment step to correct for positional, rotational, and scale variations of obtaining a canonical facial image representation which could be affected by pose bias; Face Representation, the process of optimizing a model to embed images into a feature embedding space and could introduce hyper-parameter bias and uncertainty bias; Face Verification, which includes one-to-one matching to determine whether two images belong to the same individual; and Face Identification, the one-to-many matching operation to identify an individual against a set of reference images, both of which may introduce evaluation bias when the dataset used to evaluate is not accurately representative of the target population [20].

Table 1: Timeline of Works and Contributions

Title	***Year***	***Contribution***
Vert Deep Convolutional Networks for Large-Scale Image Recognition [18]	2014	Demonstrated that representation depth is beneficial for classification accuracy on the ImageNet challenge using the ConvNet architectures.
Bias in Machine Learning–What is it Good for? [9]	2020	Describes how bias in facial recognition systems are connected and depend on each other. The authors propose a complex relationship between bias occurring in the machine learning pipeline and impacts on social discrimination.
Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation [10]	2021	Introduces the "FairFace" dataset. A novel face image dataset with seven racial categories defined across age groups and genders designed to mitigate race bias.
Pass: Protected Attribute Suppression System for Mitigating Bias in Face Recognition [5]	2021	Introduces a quantitative metric for measuring the trade-ff between bias reduction and drop in verification performance called the Bias Performance Coefficient (BPC).
Measuring Hidden Bias Within Face Recognition Via Racial Phenotypes [24]	2022	Introduces an alternative racial bias analysis methodology via facial phenotype attributes for face recognition rather than race categorization labels.
Distill and De-bias: Mitigating Bias in Face Verification using Knowledge Distillation [6]	2022	Proposes a novel distillation-based approach to enforce a network to attend to similar face regions, irrespective of the attribute category and illustrates it through the use of GradCAM.
Racial Bias Within Face Recognition: A survey [25]	2023	Surveys the state of facial recognition bias and methods to combat it through the review of various papers in the field.

Buolamwini and Gebru present the evaluation of three commercially available gender classification systems from Microsoft, IBM, and Face++ in [2] and find that darker-skinned females are the most misclassified group. Error rates among darker-skinned females reached a maximum of 34.7% while similar rates for lighter-skinned males was 0.8%. The authors developed a custom face image dataset called Pilot Parliamentary Benchmark (PPB), which is composed of images of members of parliament from three African countries and three European countries. They introduce the first intersectional demographic and phenotypic evaluation of face-based gender classification accuracy using the Fitzpatrick skin types [7], to their knowledge the first gender classification benchmark to do so.

Concepts that comprise good facial recognition datasets are introduced in [15] including technical standards International Organization for Standardiza-

tion/International Electrotechnical Commission (IOS/IEC) 19794-5 and International Civil Aviation Organization (ICAO) 9303. These technical guidelines discuss best practices for image quality, such as storage data types, observable characteristics in terms of gender, eye color, hair color, expression, properties (i.e., glasses), head pose (yaw, pitch, and roll), and facial landmark positions. Frequently, selected photos for facial recognition datasets from sources such as "in-the-wild" face datasets do not conform to such requirements. [23] compared ICAO compliance between African and Caucasian groups in the MORPH [17] dataset and found that slightly more than 48% of the African-American images were rated as ICAO compliant while more than 57% of Caucasian images were compliant. Balanced datasets such as FairFace [10] and annotation of datasets with facial phenotype attributes [24] have been introduced, targeting image acquisition, localization, and targeting evaluation bias. My model training will utilize the FairFace dataset, a novel face image dataset with seven racial categories defined across age groups and genders. Models trained from this dataset were found to be substantially more accurate with consistent accuracy across race and gender groups [10].

In [5] the authors introduce a new metric called Bias Performance Coefficient (BPC) that measures the trade-off between bias reduction and drop in verification performance. The PASS framework achieves better BPC values than existing baselines. Additionally, they introduce a method to quantitatively describe gender and skin tone bias in the context of face verification. They define gender and skin tone bias at a given false positive rate (FPR). I will use these metrics as my baseline evaluation criteria to assess the impact of hyperparameter tuning.

Many solutions have been introduced to help in mitigating racial bias in face recognition, such as ArcFace Loss Function [3], Distill and De-bias architecture [6], and the PASS framework [5] which claims to achieve better BPC values than existing baselines. However, machine learning algorithms, particularly within deep learning, contain a large number of hyper-parameters that are not learned during training but are chosen by the user. These hyper-parameters include the number of layers, kernel size, learning rate, and number of epochs [9]. Karen Simonyan and Andrew Zisserman [18] explore the impact depth and kernel sizes have on the accuracy of large-scale image recognition networks.

We cite a timeline of papers in Table I that have investigated, evaluated, and reviewed racial bias in facial recognition systems in table one. Of note is the absence of a study directly related to the impact of hyper-parameter variation on racial bias.

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 [14] 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+ [13] 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

Fig. 1: Sample Matching Dark

Fig. 2: Sample Non-Matching Dark

Fig. 3: Sample Matching Light

Fig. 4: Sample Non-Matching Light

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 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.

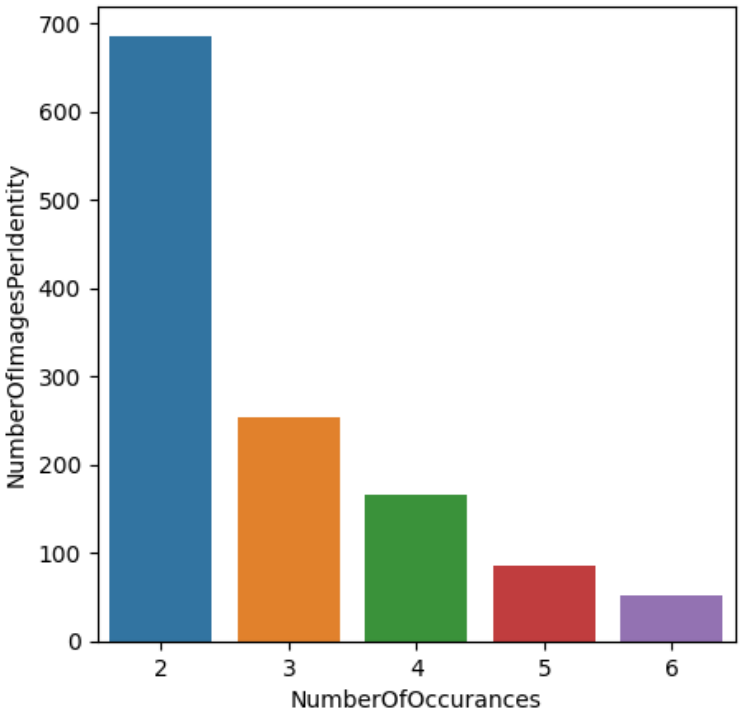

Fig. 5: Number of images per identity

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

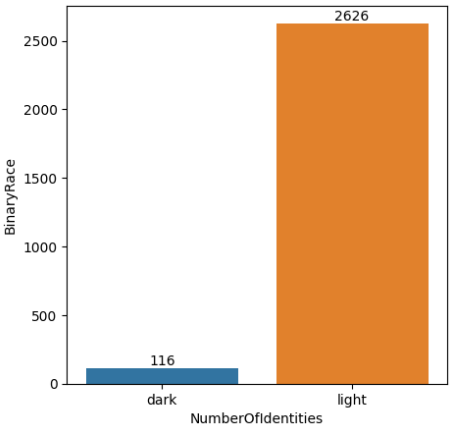

Fig. 6: Number of pairs

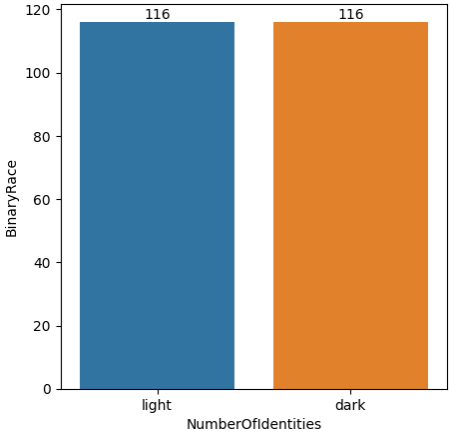

Fig. 7: Undersampled number of pairs

quality datasets and using the LFW images for test purposes only, a concept supported by [21].

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-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 [3], [16], [12], [4] report performance using ROC curve (TPR vs FPR). Per [6], [5] 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

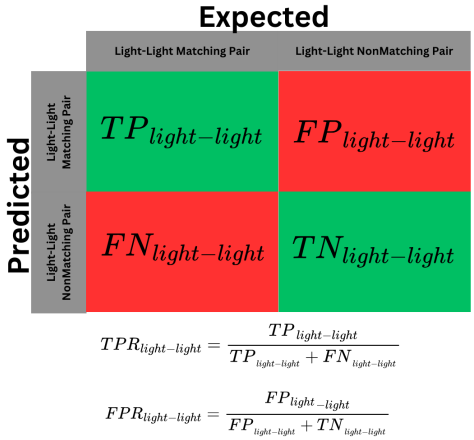

Fig. 8: Classification Report for Light-Light

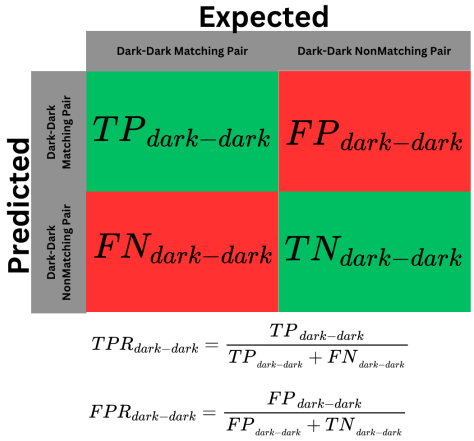

Fig. 9: Classification Report for Dark-Dark

the FPR of the best overall threshold [6], [5]. A zero bias score implies equality of odds for pairwise matching.

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.

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

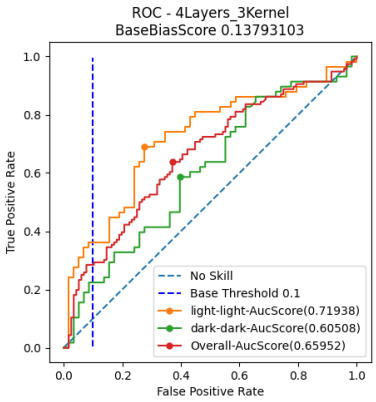

Fig. 10: 4 Layers with $3x3$ Kernel size - ROC and Bias Score.

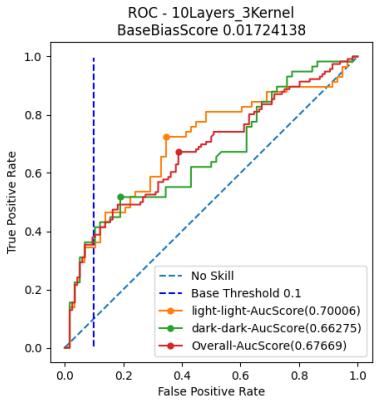

Fig. 11: 10 Layers with $3x3$ Kernel size - ROC and Bias Score.

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

$$BPC^{(F)} = \frac{Bias^{(F)} - Bias^{(F)}_{deb}}{Bias^{(F)}} \tag{2}$$

Where $Bias^{(F)}$ refers to the overall Bias obtained by original features and the corresponding bias at FPR of F. $Bias^{(F)}_{deb}$ 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 [10]. 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 [22].

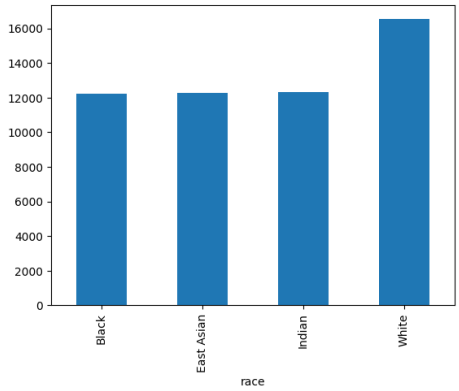

Fig. 12: Distribution of races in the FairFace training dataset.

Binarization / Regrouping of the training dataset For ease of calculations and testing, I opted to follow similar steps to [6] in which I combine and binarize the dataset—regrouping the FairFace [10] data set into two skin tone categories Light ('White' ∪ 'East Asian') and Dark ('Indian' ∪ 'Black'). As [6] mentions the skin label is not perfectly correlated with race, but it does have a high correlation nonetheless.

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 influences computational efficiency, feature resolution, and spatial hierarchies' processing. [18], 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

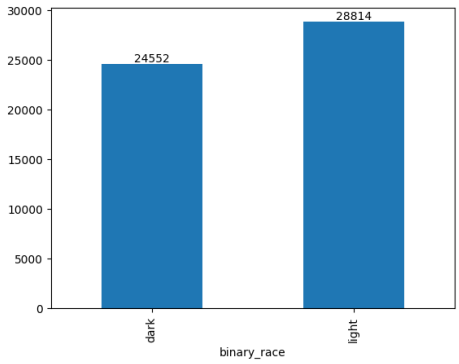

Fig. 13: Dataset distribution after data binarization.

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. [18] . 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) [18]. Models with larger 5x5 kernels displayed higher biases compared to their 3x3 counterparts, suggesting that while they capture broader spatial infor-

mation, 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.

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

Layers	BPC Score
4	0.125
8	-0.667
10	-2.000

4.2 Effect of Convolutional Layer Addition

Increasing the number of layers in a deep CNN has been shown to decrease value error percentage [18]. In my investigation, the Bias Score with a fixed kernel size of 3x3 decreased from four to eight to ten layers, respectively.

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.

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.

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.

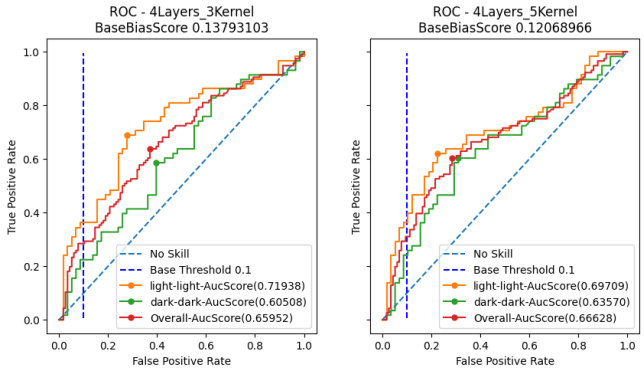

Fig. 14: ROC curves for 4 Convolution Layers and increasing Kernel Size

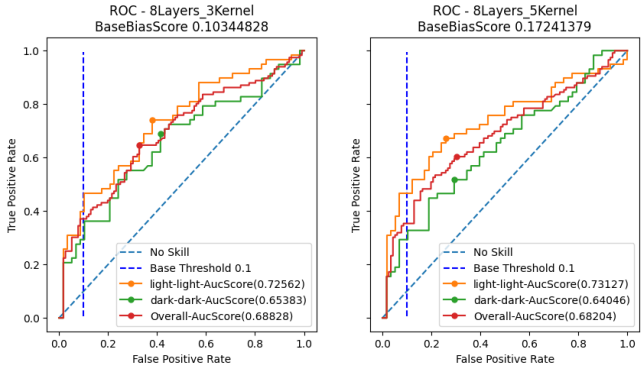

Fig. 15: ROC curves for 8 Convolution Layers and increasing Kernel Size

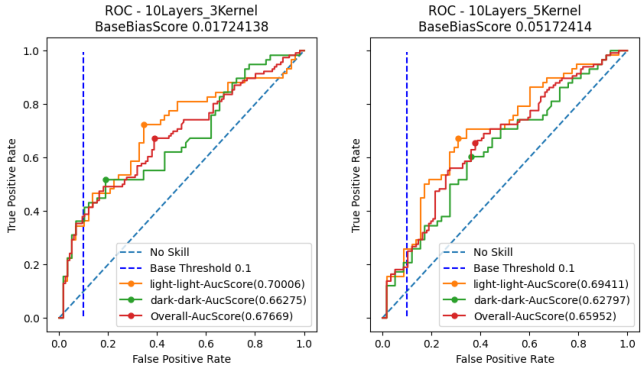

Fig. 16: ROC curves for 10 Convolution Layers and increasing Kernel Size

4.3 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 re-

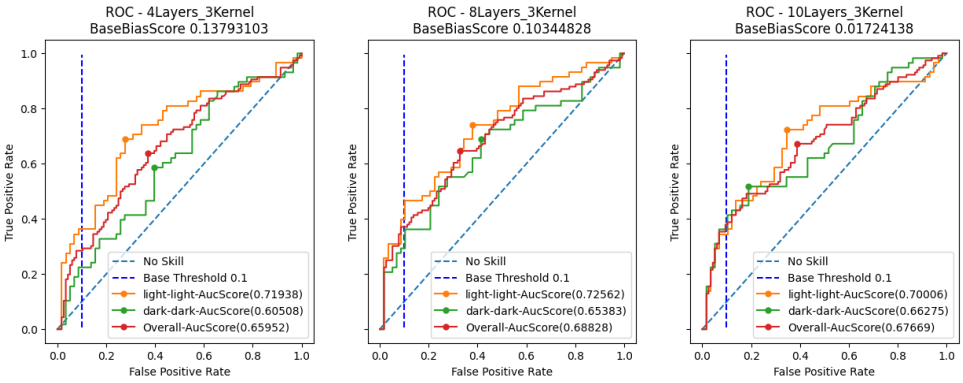

Fig. 17: ROC curves for 3x3 kernel size and increasing layers.

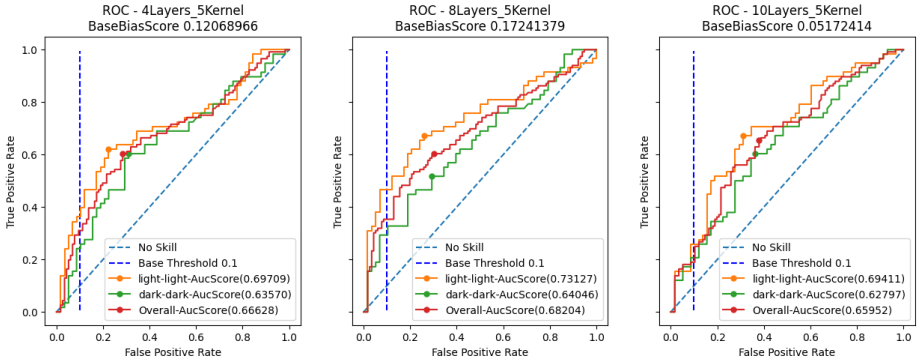

Fig. 18: ROC curves for 5x5 kernel size and increasing layers.

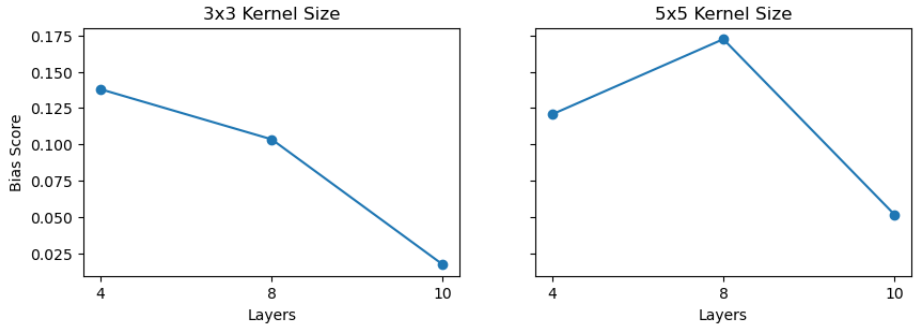

Fig. 19: Bias score change with an increase in convolutional layers.

sources required for a broader analysis. By concentrating on key hyper-parameters, I aim to provide depth rather than breadth in my investigation. This strategic

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

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.

References

1. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science **6**(2), 20 (2016)
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (23–24 Feb 2018), https://proceedings.mlr.press/v81/buolamwini18a.html

3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
4. Dhar, P., Castillo, C., Chellappa, R.: On measuring the iconicity of a face. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2137–2145. IEEE (2019)
5. Dhar, P., Gleason, J., Roy, A., Castillo, C.D., Chellappa, R.: Pass: protected attribute suppression system for mitigating bias in face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15087–15096 (2021)
6. Dhar, P., Gleason, J., Roy, A., Castillo, C.D., Phillips, P.J., Chellappa, R.: Distill and de-bias: Mitigating bias in face verification using knowledge distillation (2022)
7. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types i through vi. Archives of dermatology **124**(6), 869–871 (1988)
8. Hassanin, N.: Law professor explores racial bias implications in facial recognition technology (Aug 2023)
9. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning–what is it good for? arXiv preprint arXiv:2004.00686 (2020)
10. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
11. Labati, R.D., Genovese, A., Muñoz, E., Piuri, V., Scotti, F., Sforza, G.: Biometric recognition in automated border control: a survey. ACM Computing Surveys (CSUR) **49**(2), 1–39 (2016)
12. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
14. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). pp. 158–165. IEEE (2018)
15. Monnerat, J., Vaudenay, S., Vuagnoux, M.: About machine-readable travel documents. RFID Security 2007 (2007)
16. Ranjan, R., Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B., Nanduri, A., Chen, J.C., Castillo, C.D., Chellappa, R.: A fast and accurate system for face detection, identification, and verification. IEEE Transactions on Biometrics, Behavior, and Identity Science **1**(2), 82–96 (2019)
17. Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 341–345. IEEE (2006)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Singh, I., Goyal, G., Chandel, A.: Alexnet architecture based convolutional neural network for toxic comments classification. J. King Saud Univ. Comput. Inf. Sci. **34**(9), 7547–7558 (oct 2022). https://doi.org/10.1016/j.jksuci.2022.06.007, https://doi.org/10.1016/j.jksuci.2022.06.007

20. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–9 (2021)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
22. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
23. Vangara, K., King, M.C., Albiero, V., Bowyer, K., et al.: Characterizing the variability in face recognition accuracy relative to race. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
24. Yucer, S., Tektas, F., Al Moubayed, N., Breckon, T.P.: Measuring hidden bias within face recognition via racial phenotypes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 995–1004 (2022)
25. Yucer, S., Tektas, F., Moubayed, N.A., Breckon, T.P.: Racial bias within face recognition: A survey. arXiv preprint arXiv:2305.00817 (2023)