# Investigating the Effects of Racially Homogenous Training Data on Facial Categorization AI

Ash Mahmood (20844303)
Yiwen Dai (20828663)

## ABSTRACT

This paper assesses the accuracy of gender and age estimation in facial recognition with respect to different ethnicities for 1) neural networks trained with Caucasian-only training data and 2) neural networks trained with race-balanced training data. The objective is to categorize the biases that racially homogenous training data introduce into artificial classification networks. Given historical biases embedded in datasets, we hypothesize the CNN trained on homogenous data to be more inaccurate and less confident in recognizing the characteristics of people of colour (POC). The experiment tested two feed-forward, supervised convolutional neural networks trained on FairFace data, which is balanced for age, race, and gender. The homogeneous training data was extracted from the Caucasian subset of FairFace's data, while the heterogeneous data encompassed all racial distributions provided by FairFace. Both models were tested for gender and age estimation accuracy against images from White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino identities. To investigate the impact of dataset diversity on the accuracy of facial feature classification models, two models were trained: one on a homogenous dataset comprising images from the White race group and the other on a heterogeneous dataset spanning over seven race groups. The findings indicate that the heterogeneously trained model significantly outstrips its homogeneously trained counterpart in test accuracy across all race groups, underscoring the efficacy of diverse training data in enhancing classification robustness. Notably, the homogeneously trained model did not attain peak performance on the White race group, which could be indicative of overfitting and poor generalization. Moreover, when evaluated on a composite test set representing all race groups, the heterogeneously trained model demonstrated elevated accuracies, confirming the benefit of diversity in the applied features for facial recognition tasks. These results highlight the critical importance of dataset diversity in developing unbiased and effective facial classification systems. The findings of this paper show that racial bias is perpetuated in facial recognition models and facial feature extraction across all races can be improved through adding diversity to datasets. Future research into this topic could explore methods to train neural networks such that racial bias is mitigated despite racially-homogenous data.

## 1 - INTRODUCTION

### 1.1 - Problem

Facial recognition using artificial intelligence (AI) is quickly becoming a method to authenticate users, collect marketing data, and even identify criminals over video surveillance. However, in this new movement, training datasets for biometrics machine learning have neglected and disadvantaged marginalized demographics. The under- or over-representation of certain groups within training datasets cultivates a bias in machines which mimics those in paradigmatic social structures. Using balanced

datasets, we can model and quantify the effects of homogeneity and heterogeneity in training datasets when determining age range and gender. This will be done using FairFace, which is a dataset created to mitigate gender, race, and age biases in AI [1].

Inaccuracy in facial recognition is most pronounced when identifying dark-skinned women across several different recognition softwares [2]. This is largely in part due to a lack of representation in AI training datasets, which prevents the machine learning algorithm from adequately learning racialized and gendered features. Moreover, the proliferation of AI into different sectors has deepened the consequences of non inclusive dataset design. Recent law enforcement practices have begun to use AI to identify criminals [2]. This has led to false arrests that disproportionately affect Black people, as they are often overrepresented in police mugshots due to systemic oppression of the demographic.

Gender and age are of particular interest because of its effects on overall facial identification [3]. Categorization in humans by these labels creates expectations for social interaction, in both positive and negative ways. Models trained for this task can be used for security applications, social media filters, criminal flagging systems in CCTVs, as well as niche tasks such as providing health recommendations. Consequently, it is important to consider diversity in datasets and to quantify inaccuracies in current models so as to understand the proliferation of racial bias into AI models, especially in gender and age determination.

## 1.2 - Literature Review

### *1.2.1 - Neurophysiological Identification of Gender and Age*

Studies have shown that gender and age are decoded much sooner than facial identification. The suggested region of the brain used for facial identification is part of the fusiform gyrus, referred to as the fusiform face area (FFA), as well as the inferior occipital gyrus (OFA) [4]. While both areas are crucial for gender and age determination, it was found that age categorization incites lower activation of the right FFA and left OFA [5]. As compared to complete facial identification, determination of gender and age occurs posterior to the FFA and OFA, suggesting partial dependence on low-level visual features that are processed earlier in the visual processing pathway [4, 5]. This is validated by the slower and weaker facial identification process as compared to gender and age identification [6].

ORB, and the proposed low-level visual processes behind gender and age recognition within faces are interesting biological responses that can be mimicked with feed-forward, supervised convolutional neural networks (CNNs) [7]. CNNs often make use of alternating layers of convolution and pooling, which mimic simple cell and complex cell interactions in the visual processing pathway [8]. Using this result can create biological consistency within the model design.

### *1.2.2 - Quantifying In-Vivo Age and Gender Categorization*

To obtain a reasonable expectation for accuracy in age and gender categorization, several studies were reviewed featuring human participants. From averaging the results of the study by Yadav et al., accuracy in identification of an age range between the ages of 21 to 80+ is 40.87% [9]. In another study

by Bruce et al. the accuracy of human gender identification was found to be 96% overall [10]. The faces that were provided as visual stimuli, as well as the participants categorizing the faces, were primarily Caucasian. For a general guideline of expectations for human categorization accuracy, these two percentages can be multiplied to obtain the multi-classification accuracy. Resultantly, the overall accuracy of facial identification (not accounting for racial disparities in identification) for a person is expected to be around 39.2%.

*1.2.3 - Choosing an Appropriate Dataset*

A paper by Rafique et al. uses the dataset Labeled Faces in the Wild for gender and age recognition [11]. The benefit of this data is that it features poor lighting, odd positions, and low resolution. Consequently, the model becomes more robust against different photo conditions, and thus can be applied to a greater range of applications. However, the creators of this dataset highlight the lack of women as well as those over the age of 80 [12]. This necessitated the use of a separate dataset that similarly has labeled images, but also balances demographics. We turned to FairFace, which aims to reduce bias of AI trained models through providing a dataset with balanced age, race, and gender. Figure 1 below shows the race distribution of popular training datasets, with FairFace at the far right having the most balanced [1]. Due to inaccuracies in sexual dimorphism at younger ages, individuals below the age of 20 were excluded from this dataset.

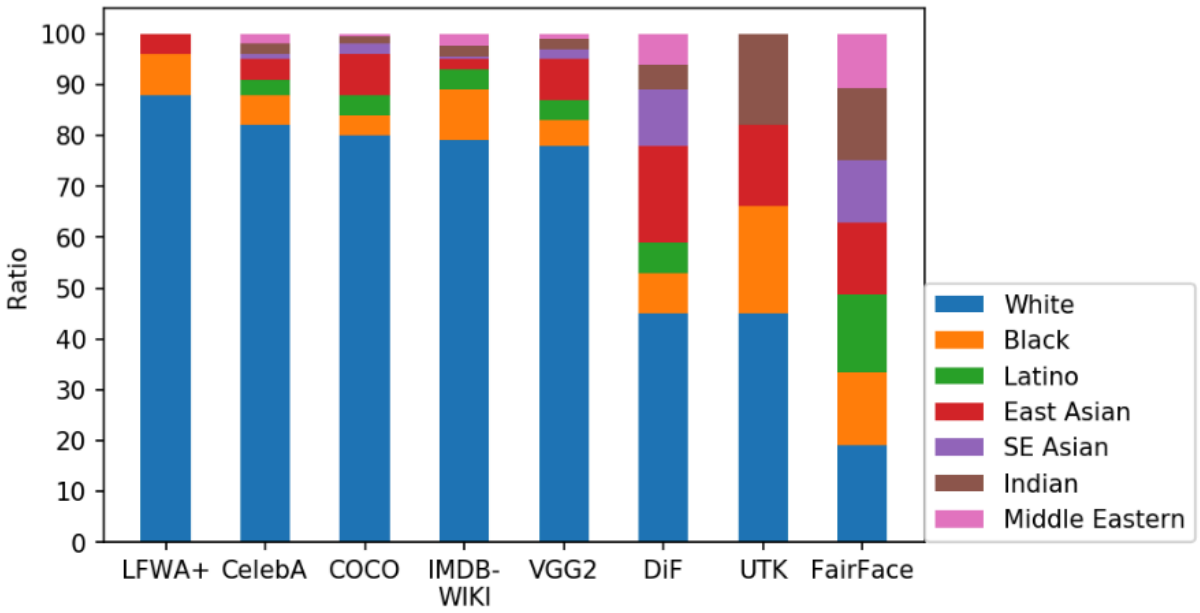


*Figure 1. Racial Distribution In Popular Training Datasets*

The creators of FairFace, Karkkainen and Joo, train a model and provide model accuracy when determining gender, race, and age. Gender was identified with 94.2% accuracy, race with 93.7% accuracy, and age with 59.7% accuracy [13]. The drop in accuracy with age is expected as genetic differences in ages can make disparities in age difficult to determine. Between the gender and age accuracies, it is

expected that a heterogeneously trained model (that is, a model trained on an equal subset of each race) identifies the correct age and gender approximately 56% of the time. This is considerably higher than the estimated recognition capability of humans, which is 39.2% from previous calculations.

*1.2.4 - Related Work on Gender and Age Classification*

The primary methods used in face recognition for gender and age estimation are FisherFace, EigenFace, and CNNs. Earlier algorithms that mathematically characterize facial features are also popular, albeit more hand-crafted, such as "Histogram of Oriented Gradients (HoG), Speeded Up Robust Features (SURF), and Scale Invariant Feature Transform (SIFT)" [14]. FisherFace uses an algorithm referred to as the Fisher Linear Discriminant to find a coefficient that distinguishes between different classification groups. It is less sensitive to photo conditions as compared to EigenFace, and learns individual characteristics when recognizing features [15]. EigenFace in comparison is less robust, using principal component analysis to determine feature points and distances that correlate to different classifications [15]. The lighting conditions and granularity of the photo must be more consistent, well-lit, and centered for this to be used. CNNs for classification are much more popular recently, with numerous models existing and superseding traditional approaches [14]. These earlier methods will be disregarded for both the sake of accuracy as well as for biological relevance, as CNNs are most commonly used to mimic visual processing pathways.

There were two notable CNN models which did similar classification to what this paper aims to do. Both models use CNNs and labeled datasets for training and validation of accuracy. There is no existing model that has been trained to specifically analyze the effect of homogenous and heterogenous training on different races with respect to gender and age recognition.

The model proposed by Rafique et al. determines gender and age through a CNN with three convolutional layers and two fully-connected layers [14]. Pre-processing for this model was done through trimming the photo with facial focus. Furthermore, batch normalization and dropout layers were used to further limit overfitting, with rectified linear unit (ReLU) neurons being used in layers for activation. This model achieved an accuracy of 57.60% for age and 84.2% for gender. Another model by Sheoran et al. uses a deep CNN to classify gender and age with a more complex set of layers [16]. The age determination CNN featured five sets of a convolution layer alternating with a max pooling layer, then three fully connected layers. The kernel size was kept relatively small, and ReLU activation used for the layers. The gender estimation layer process was not included in the paper. This model achieved a maximum accuracy of 79.122% for age, and 94.517% for gender. The SGD optimization was found to be the best for age estimation, while Adam was best for gender classification.

Based on these two models, we can extrapolate the network architecture for a CNN that applies limited hardware capacities and determines the capability of simultaneous gender and age recognition for different ethnicities.

### 1.3 - General Approach and Objectives

From the literature review, it was seen that CNNs are the most accurate biologically relevant method of visuospatial characterization as compared to other algorithms. Consequently, this model was modified from the working models that were discussed above. Due to limitations in computational hardware, a lower number of layers were used for the CNN. To extract age and gender, the data was pre-processed via resizing and normalization. In order to allow for a wider variety of input image types, the photos were allowed to remain coloured or black-and-white. More details of the model are discussed in the methods.

The objective of creating this model is to see how diversity impacts the gender and age recognition of different races. Through training a homogenous model (White-dominated with balanced gender representation), as well as a heterogeneous model (balanced racial and gender representation), different racialized faces can be tested for accuracy in gender and age estimation. This can be done through creating both models, then testing both for the accuracy in age and gender estimation for all the races defined in FairFace: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino.

### 1.4 - Interesting Biological Extensions of Model

Facial categorization is further complicated in humans by own-race bias (ORB), which is a cultural and social effect where adult facial recognition is more accurate when identifying faces of a person's own racial group as compared to other ones [17]. The reason underlying this is suggested to be primarily related to nationality, as people in the same nation have higher exposure to faces that have similar patterns to their own. This bias can theoretically be extended to AI. Since many datasets have been traditionally biased towards featuring White men, being able to apply this model to feature ORB and show gender and racial biases could be an intriguing avenue.

In addition, studies have shown that people who cannot identify faces (a condition called prosopagnosia) are still able to reliably identify gender and age [18]. The reason for this is suggested to be due to overexposure of gendered faces in society [19], in addition to reliance on lower-level processing for categorization as compared to facial identification. This occurrence could be particularly relevant to explore for gender and age determination in biologically-consistent artificial intelligence.

## 2 - METHODS

The GitHub repo containing all the code can be found here: https://github.com/yiwen-dai/SYDE552

### 2.1 - Model Description and Comparison Baselines

As this paper aims to explore how bias in the training data for CNNs can affect identification of facial features, the base model that is used would be the convolutional neural networks. Given that the

goal is for the images of faces to be mapped to one of 18 classes that span the possible combinations of age range and gender, this can be recognized as a multiclass image classification task. Thus, the architecture of the customly defined CNN class is designed in an attempt to best tailor for this type of task, as described below.

The model has a total of 3 convolutional layers with varying filter sizes to progressively extract more complex features from the images, connected using the ReLU activation function. After each convolutional layer, a max pooling layer is applied to reduce the dimensionality of the data and to combine the outputs of neuron clusters. To reduce overfitting, dropout is also implemented to trim connectivity of the neurons between layers. The data from the convolutional layers is then passed into a dense neural net with a total of 3 layers, with dropout and ReLU as well. To produce the predictions, the outputs from the final layer are passed into the log softmax function to provide the probabilities of the image class while maintaining numerical stability and computational efficiency. In terms of image transformations, the image is flipped horizontally 50% of the time as well for better generalization.

Based on the goal of this paper, the models are trained on different datasets (homogeneous vs. heterogeneous), which would be the baseline comparison: observing how the models perform on uniform versus diverse data sets. However, in terms of a more generalized baseline performance of a multiclass image classification model (in this case, gender/age), the performance depends heavily on the complexity of the task. Since gender classification is more straightforward (binary class), simpler models can indeed achieve high accuracies, with models reaching up to approximately 90% validation accuracy on a relatively balanced dataset of about 2200 facial images [20]. On the other hand, regarding age classification, it is generally more challenging due to the finer gradations and subtleties in aging features compared to binary gender classification, and may yield lower accuracies, especially as the number of age classes increases. For example, a convolutional neural network model used for both gender and age classification, reached an accuracy of about 73% for age classification with 6 age ranges, with a total of 6 output classes [21]. However, since the models in this paper are less complex with regards to architecture and are trained to predict both gender and a much larger number of age ranges (increasing the number of output classes to be significantly greater, with a total of 18), a potentially feasible accuracy range to target for this project could be around 50%-60% given the increased difficulty.

**2.2 - Independent Variable**

As mentioned in the previous section, although the model went through many iterations to improve performance, the main changes introduced to the models were not the architecture, but rather the data that they are being trained on. One model will be trained on a homogenous dataset, consisting of facial images that are limited to the White race group, and another model will be trained on a heterogenous dataset that is made up of images from 7 different race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. The homogenous dataset is created as a subset of the FairFace dataset (filter by `race=='White'`), whereas the heterogeneous dataset contains all the images in the FairFace dataset. It is important to mention that as a result of this, the heterogeneous dataset is approximately 4-5 times larger than the homogeneous dataset (14874 vs 78069 images respectively).

### 2.3 - Plan

The primary objective is to test the model's robustness and accuracy in classifying facial features from datasets that differ in their diversity based on the dataset that they were trained on. The homogenous dataset consists of facial images that are limited to the White race group, whereas the heterogeneous dataset includes images from 7 different race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. The expected findings for the effect of different training data is that the model trained on heterogeneous data will show higher accuracy in identifying the gender and range group of faces, regardless of their race group, whereas the model trained on homogeneous data might perform similarly when tested on faces from the White race group, but perform noticeably worse with faces from other race groups. This hypothesis will be tested using the testing accuracy of the two models when presented with the previously unseen before homogeneous and heterogeneous datasets of faces.

### 2.4 - Data Analysis

In terms of the analysis of the results, the testing accuracies of the models will be directly compared against each other. The learned features will also be plotted for comparison.

### 2.5 - Model Complexity

To determine the complexity of the model, it is important to note that there are many factors that would contribute to the complexity of a neural net model, such as the depth of the network, number of parameters, and architectural complexity, relative to typical CNNs used in similar tasks.

In terms of the depth, the model consists of three convolutional layers followed by three fully connected layers. This setup is moderately deep but not exceptionally so, especially in the context of modern deep learning models which can have dozens of layers. With regards to the number of parameters, the number of parameters can be substantial due to the large fully connected layers (e.g., from 8192 to 512 neurons), but compared to the parameter counts seen in larger models like VGG or ResNet architectures, or even the GPTs, this number would be considered to be quite small. The inclusion of dropout and max pooling is typical for CNNs and is aimed at reducing overfitting and computational load, respectively, and do not necessarily complicate the model beyond standard practices since they are common features.

In conclusion, this model is more complex than very basic neural networks but is far from the complexity than many state-of-the-art models used in recent computer vision tasks. However, it is important to note that this is not intended for large scale or commercial use, rather with the goal of striking a balance between adequately learning from the data provided without incurring the high computational cost of a very deep network, and without overfitting on the limited data. In addition, given

that age and gender classification is rather lower level in the biological visual processing systems, the goal is to not overcomplicate the model but rather keep it more simple to better match human biology.

## 3 - RESULTS

The figures below highlight the main results of the models, including the final test accuracies of the models on different test datasets, the learning progress, as well as the final features extracted by the convolutional layers. In terms of the hyperparameters chosen for the models (such as learning rate, momentum, etc), they were determined after extensive experimentation with a large range of values, and comparing the model performance.

Figure 2 shows that the homogeneously trained model is less accurate for all races, including White. Moreover, heterogeneously trained models show a marked improvement in comparison, following similar magnitudes of improvement across all races.

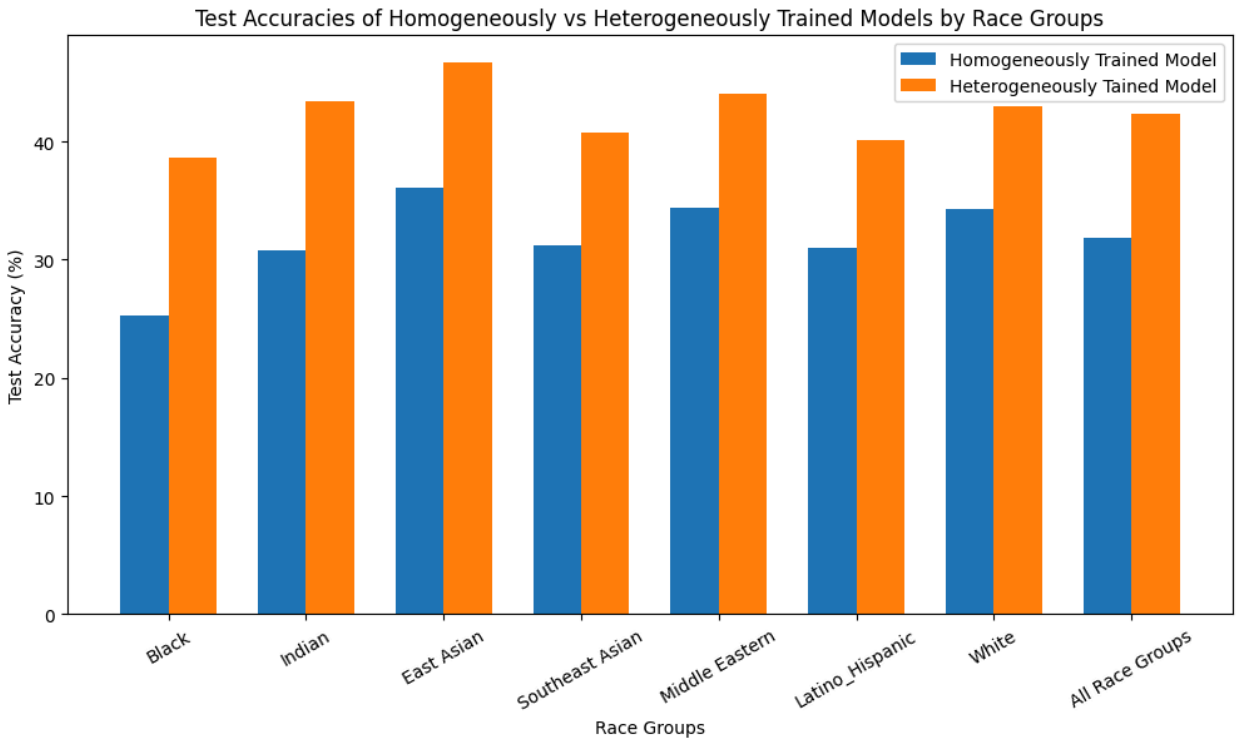


*Figure 2. Testing accuracies of the two models on different test data*

Figures 3 and 4 show the training process of the two models. It is interesting to note, that although the two models only differ in their training datasets, the shape of the accuracy curves are very different. In the homogeneously trained model, the validation accuracy seems to vary much more greatly, potentially due to the smaller training and validation sets.

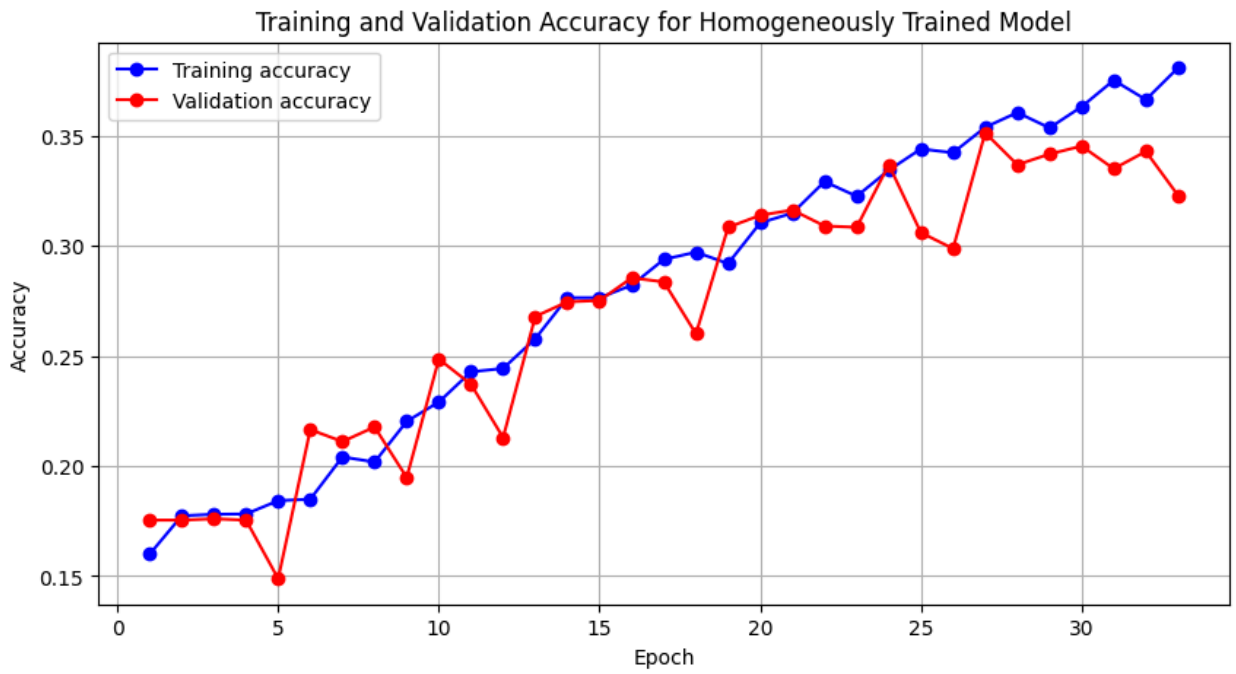


*Figure 3. Training Process for Homogeneously Trained Model*

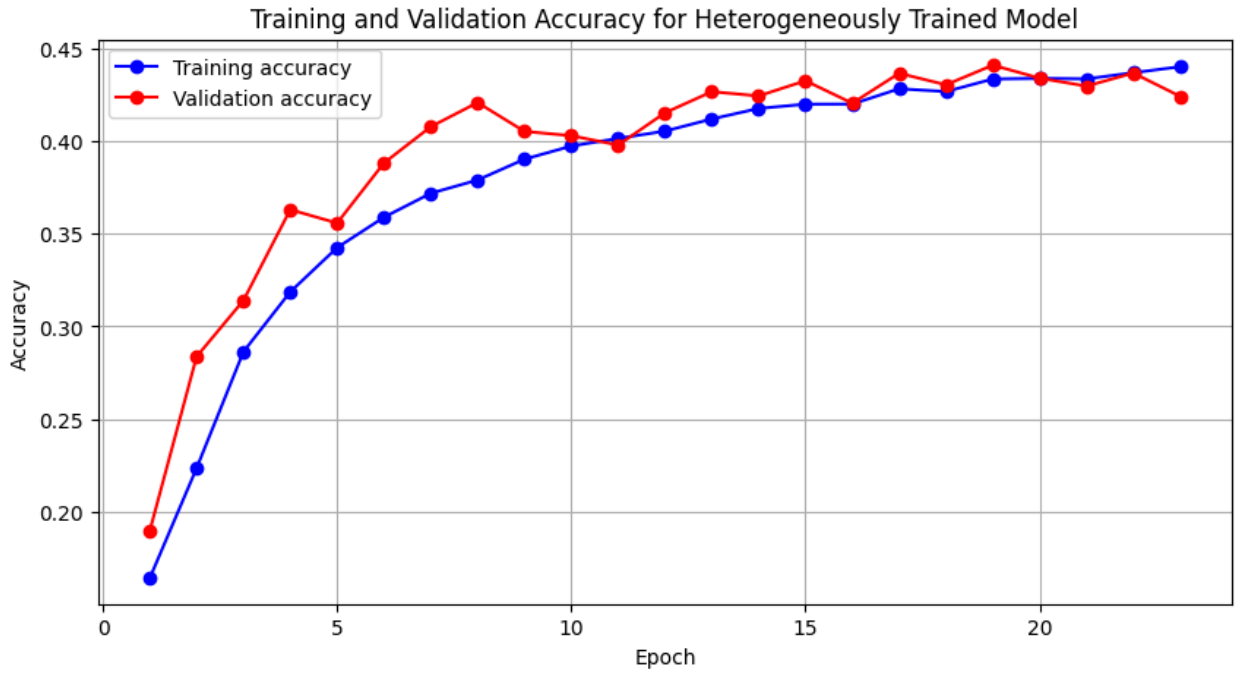


*Figure 4. Training Process for Heterogeneously Trained Model*

Figures 5 and 6 show the final features extracted by the models. These images are created using the final weights of the final convolutional layers of the models.

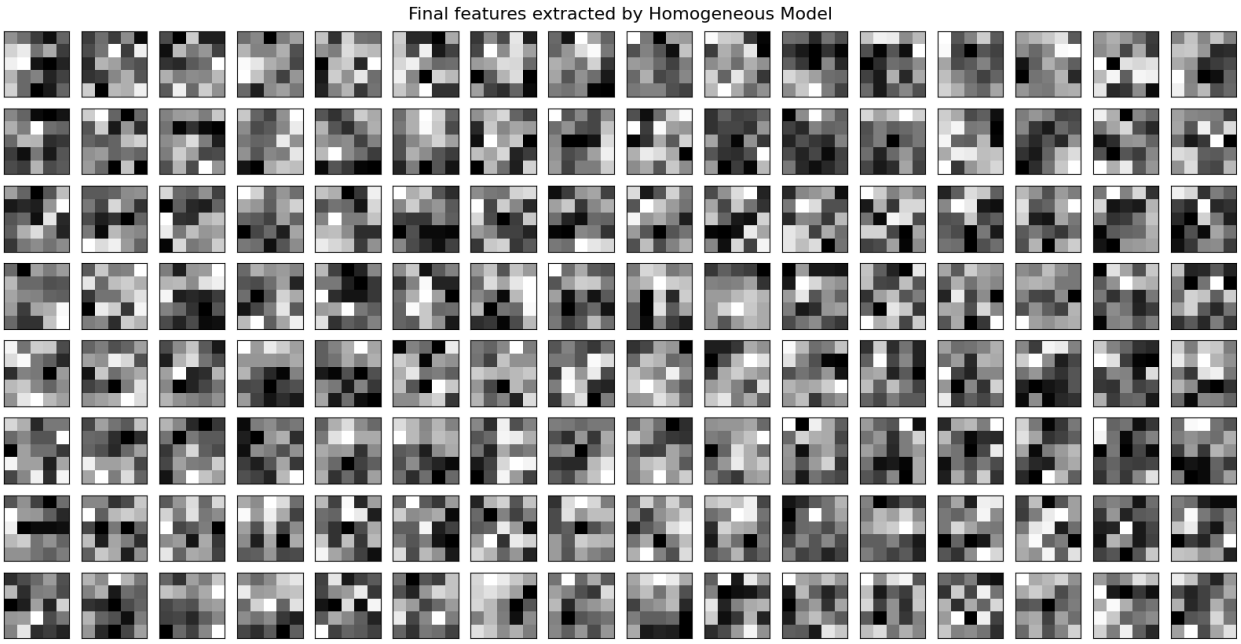


*Figure 5. Final Features Extracted by Homogeneously Trained Model*

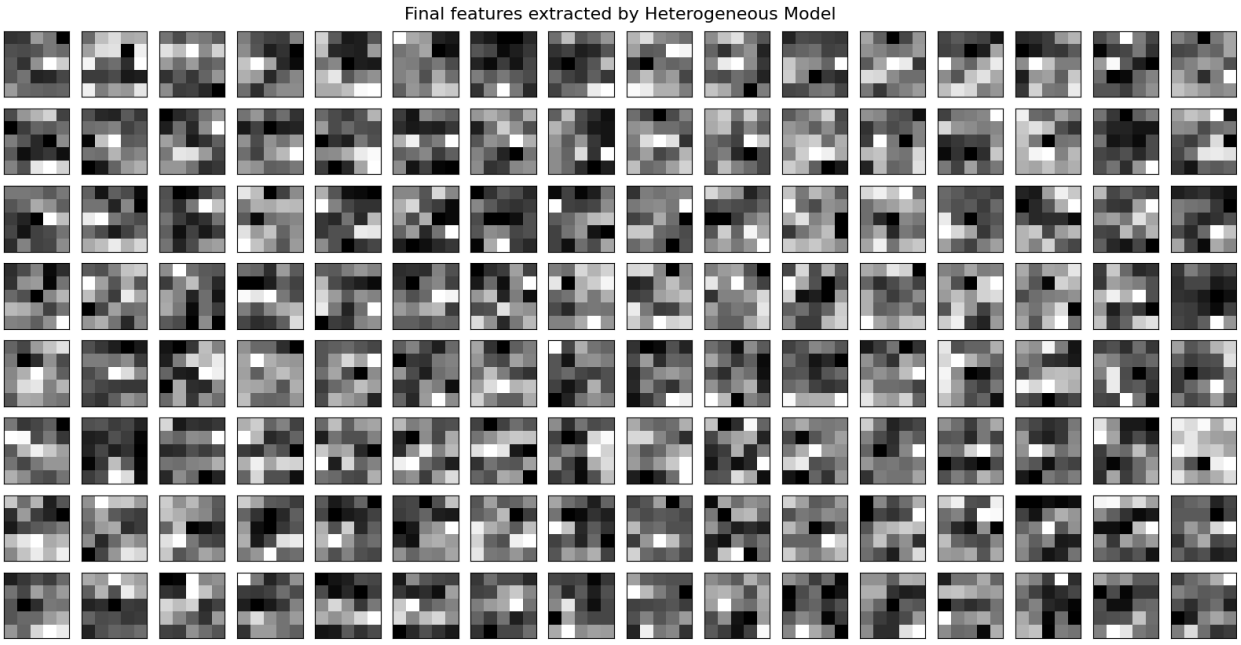


*Figure 6. Final Features Extracted by Heterogeneously Trained Model*

## 4 - DISCUSSION

### 4.1 - Summary

In summary, the results presented in the bar chart underline the critical impact of training data diversity on the robustness of facial classification models and echo a broader theme in machine learning about the importance of representative training samples for equitable and effective model performance.

### 4.2 - Findings

From Figure 2. above, we can see that the heterogeneously trained model demonstrates superior performance across all race groups, supporting the initial hypothesis that a model trained on a more diverse dataset would exhibit higher accuracy in identifying gender and age categories irrespective of race.

In terms of the homogeneously trained model, the figure shows the least accuracy with the Black race group, which aligns with expectations given the lack of representation in the training data. Surprisingly, the accuracy for the White race group is not the highest among the categories, contrary to what one might anticipate since the model was exclusively trained on data from this group. This could suggest that the model is not well-generalized even within its own training demographic, potentially indicating issues like overfitting or a non-representative training subset within the White race group. This could also be due to the fact that the homogeneous training dataset is significantly smaller than the heterogeneous training dataset, each consisting of 14873 and 78069 images respectively.

On the other hand, the heterogeneously trained model achieves more consistent accuracy rates across all race groups, demonstrating enhanced generalization capabilities. This uniformity in performance validates the benefits of incorporating a richly diverse training dataset and reflects a reduced bias in facial feature recognition across different races.

In conclusion, the homogeneous model performed poorly for people of colour as expected, and the heterogeneous model performed significantly better. This correlates with biologically-relevant knowledge of own-race bias and the perpetuation of racial biases in facial recognition software that were discussed in the Introduction. Furthermore, the heterogeneous model accuracies are higher than what would be expected for human participant identification of gender and race which is 39.2%. The homogenous model, in comparison, is less competent than the heterogeneous model at own-race identification. One reason for this could be due to the FairFace dataset only being able to extract heterogeneous or homogeneous training sets. As a result, the dataset extracted for training the homogenous set was considerably smaller as compared to the heterogenous training set. Another plausible reason is the exposure to other-race data has allowed for the model to abstract further, even for White datasets.

The results of the heterogeneous model fall short of the capabilities of the FairFace-made model, which features more layers that were too computationally expensive for the scope of this paper. A possible extension to better characterize the impact of race in gender and age estimation is to deepen the

CNN through adding more layers, doing additional pre-processing prior to CNN input, and separating gender and age classification. Additional pre-processing could look like facial centering, such as what is done in the Rafique et al. model, or even reducing complexity of input values by changing the dataset to grayscale. Furthermore, adding more layers and creating two separate models with different pathways like in the deep CNNs that Sheoran et al. creates could further increase accuracy of identification.

### 4.3 - Unexpected Results

Unexpectedly, East Asian faces were more easily recognized by the homogenous model than White ones. The reason for this is not well-known, but could relate to East Asian complexity being lighter-toned in comparison to other races.

Furthermore, it was also unexpected that the homogeneously trained model does not achieve the highest accuracy on the White race group, pointing to potential limitations in the model's training that hinder even race-specific performance. Moreover, the gap in accuracy between the two models isn't as wide for certain race groups, such as Indian and Southeast Asian. This observation raises intriguing questions about the shared facial characteristics across these demographic groups and the features that the homogeneously trained model has possibly learned that transcend its training limitations. By examining the learned features from the final convolution layer, we can see that these learned features are rather abstract and might be difficult to understand as humans, but do form some characteristics such as lines in different directions, curves, etc. This mimics what is seen biologically through simple cells, which preferentially fire given certain stimuli orientation (Kandel et al citation).

### 4.4 - Limitations and Future Work

For future work, expanding the scope of the study to be more inclusive and detailed would be highly valuable. Addressing the first limitation, future iterations of the project could incorporate a broader spectrum of gender identities beyond the male and female binary. This would not only be more reflective of the diversity present in society but would also challenge the model to learn and recognize a wider array of facial features associated with non-binary and gender non-conforming individuals. This extension would require careful consideration on how to ethically and accurately collect and label data that respects participants' identities.Furthermore, an additional area of improvement would be the differences in training set sizes for both the heterogeneous and homogeneous models, as this could introduce bias based on the quantity of data available to each model. Ensuring that both models are trained on datasets of comparable size would lead to a more fair comparison of their performance. In addition, expanding the categories recognized by the model to include more granular age ranges, different expressions, and potentially other sociodemographic variables like socioeconomic status or occupation could greatly enhance the model's utility. Such detail would enable the model to serve a broader range of applications, from more personalized user experiences in technology to nuanced sociological studies. In summary, the next steps would involve a focused effort on inclusion, equity in data volume, and a richer classification system that mirrors the complexity of human demographics.

## REFERENCES


[1] J. Joo and K. Kärkkäinen, “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)*, pp. 1547–1557, 2021, doi: https://doi.org/10.1109/wacv48630.2021.00159.

[2] A. Najibi, “Racial Discrimination in Face Recognition Technology,” *Science in the News*, Oct. 24, 2020. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

[3] S. Wu and D. Wang, “Effect of subject’s age and gender on face recognition results,” *Journal of Visual Communication and Image Representation*, vol. 60, pp. 116–122, Apr. 2019, doi: https://doi.org/10.1016/j.jvcir.2019.01.013.

[4] C. Kaul, G. Rees, and A. Ishai, “The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain,” *Frontiers in Human Neuroscience*, vol. 4, 2011, doi: https://doi.org/10.3389/fnhum.2010.00238.

[5] H. Wiese, N. Kloth, D. Güllmar, J. R. Reichenbach, and S. R. Schweinberger, “Perceiving age and gender in unfamiliar faces: An fMRI study on face categorization,” *Brain and Cognition*, vol. 78, no. 2, pp. 163–168, Mar. 2012, doi: https://doi.org/10.1016/j.bandc.2011.10.012.

[6] J. M. Contreras, M. R. Banaji, and J. P. Mitchell, “Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race,” *PLoS ONE*, vol. 8, no. 7, p. e69684, Jul. 2013, doi: https://doi.org/10.1371/journal.pone.0069684.

[7] S. Lall, “What’s in a face?,” *MIT News | Massachusetts Institute of Technology*, Mar. 22, 2019. https://news.mit.edu/2019/human-brain-face-recognition-0322

[8] E.-Y. Huan *et al.*, “Deep Convolutional Neural Networks for Classifying Body Constitution Based on Face Image,” *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1–9, 2017, doi: https://doi.org/10.1155/2017/9846707.

[9] D. Yadav, R. Singh, M. Vatsa, and A. Noore, “Recognizing Age-Separated Face Images: Humans and Machines,” *PLoS ONE*, vol. 9, no. 12, p. e112234, Dec. 2014, doi: https://doi.org/10.1371/journal.pone.0112234.

[10] V. Bruce *et al.*, “Sex discrimination: how do we tell the difference between male and female faces?,” *Perception*, vol. 22, no. 2, pp. 131–152, 1993, doi: https://doi.org/10.1068/p220131.

[11] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais, and T. Yasir, “Age and Gender Prediction using Deep Convolutional Neural Networks,” *IEEE Xplore*, Nov. 01, 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8966704.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.,” *vis-www.cs.umass.edu*, Oct. 2007. https://vis-www.cs.umass.edu/lfw/

[13] “Papers with Code - FairFace Benchmark (Facial Attribute Classification),” *paperswithcode.com*. https://paperswithcode.com/sota/facial-attribute-classification-on-fairface.

[14] R. Thaneeshan, K. Thanikasalam, and A. Pinidiyaarachchi, “Gender and Age Estimation From Facial Images using Deep Learning,” *IEEE Xplore*, Dec. 01, 2022. https://ieeexplore.ieee.org/abstract/document/9993277?casa_token=4TvXco9AI6MAAAAA:bYAK9yFhGgvzgOUin3E5PpGKozG-hTAKfmtHrmJJc81fhKNAQvoKQ9c-6lMWE3kmWddBoPlOvjs.

[15] G. Park and S. Jung, “Facial Information Analysis Technology for Gender and Age Estimation,” *arXiv (Cornell University)*, Nov. 2021.

[16] V. Sheoran, S. Joshi, and T. R. Bhayani, “Age and Gender Prediction Using Deep CNNs and Transfer Learning,” *Communications in Computer and Information Science*, pp. 293–304, 2021, doi: https://doi.org/10.1007/978-981-16-1092-9_25.

[17] V. Proietti, S. Laurence, C. M. Matthews, X. Zhou, and C. J. Mondloch, “Attending to identity cues reduces the own-age but not the own-race recognition advantage,” *Vision Research*, Mar. 2018, doi: https://doi.org/10.1016/j.visres.2017.11.010.

[18] G. Chatterjee and K. Nakayama, “Normal facial age and gender perception in developmental prosopagnosia,” *Cognitive Neuropsychology*, vol. 29, no. 5–6, pp. 482–502, Sep. 2012, doi: https://doi.org/10.1080/02643294.2012.756809.

[19] J. DeGutis, G. Chatterjee, R. J. Mercado, and K. Nakayama, “Face gender recognition in developmental prosopagnosia: Evidence for holistic processing and use of configural information,” *Visual Cognition*, vol. 20, no. 10, pp. 1242–1253, Dec. 2012, doi: https://doi.org/10.1080/13506285.2012.744788.

[20] A. Ponnusamy, “Gender detection (from scratch) using deep learning with keras and cvlib,” *GitHub*, Oct. 13, 2021. https://github.com/arunponnusamy/gender-detection-keras

[21] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, “Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms,” *Sensors*, vol. 21, no. 17, p. 5892, Sep. 2021, doi: https://doi.org/10.3390/s21175892.