Investigation into Ethnic Bias in Face Recognition Task

Wiktoria Radecka[1], *Dele Ayeni*[1]

[1]*Faculty of Engineering Technology, KU Leuven, 3000 Leuven, Belgium*

Abstract

In this paper, we explore the roots of the ethnic bias within the face recognition classification task. VGG16 neural network model is trained on two racially aware datasets created from the large-scale VMER dataset. We asses the model's performance with accuracy and learning loss plots, and furthermore illustrate the model performance across different ethnicities using confusion matrices. By employing these evaluation metrics, we measure both the overall performance of the model and any biases that may arise due to variations in ethnic representation within the training datasets.

1 Introduction

In the recent years, the use of artificial intelligence grew exponentially in every area of life and it is still growing. Computers learn by identifying different patterns and relationships within enormous datasets. With this, deep-rooted discrimination has been uncovered that stems from the dataset imbalance. This poses a big societal problem since it turns out that racism is still present in healthcare, surveillance systems, law enforcement and other areas AI is present in. Models fail to generalise their results to the population minorities. Cases of most interest to us include those regarding facial recognition. Those provoke a strong sense of injustice as they end in wrongful arrests, restricted access to services, and violations of privacy.

Literature suggests that the reason for this is that the models are trained on public datasets and those are mostly comprised of public figures and those historically have been predominantly Caucasian. This prompts a question on at which point of imbalance in the dataset, the model that is going to be trained on it becomes biased?

The aim of this project is to explore the underlying factors that contribute to the ethnic disparity caused by the machine learning models in face recognition. The input to our algorithm is a image in colour showing individual faces. We then use a neural network to output a predicted identity of an individual.

2 Related Work

Despite considerable advancements in face recognition tasks, recent research continues to reveal bias and discrimination in its algorithms, leading to inaccurate and unfair outcomes. Recent research into the causes of racial bias and approaches to mitigating it show that the addressing the ethnicity gap is complex and requires careful machine learning model consideration[9].

Recently, [3] provided most comprehensive review of bias causes. In this study, the authors discuss data-driven and scenario modelling factors underlying face identification algorithms. Specifically, image quality and decision thresholds were investigated and it was shown that the former affects not only recognition rate but also the ethnic bias. The authors also established that uniform identification decision thresholds for different ethnicities is not adequate in face recognition tasks when the underlying sub-population distributions in the dataset differ.

Comparable results were found by [6] in terms training data distribution. Considering that CNN model is based on statistical models, it is no surprise that imbalanced distributions cause the distribution differences in training and test sets and consequently cause worse model performance. To boost model performance, the authors use oversampling to mitigate negative impact of data imbalance.

The effects of dataset imbalance on the ethnicity bias has been studied extensively. This focus on bias mitigation has lead to the creation of Balanced Faces in the Wild (BFW) dataset by [8]. With this dataset, the authors emphasise the importance of balanced datasets and demonstrate that grouping subjects (faces on the images) into subgroups leads to lower number of mistakes done by the CNN model across subgroups. Following this, the authors noted that this additional step of subgroup forming requires computational resources which may become a problem with larger datasets.

Another prominent example of reducing the racial bias by mitigating the effects of dataset imbalance is the use of synthetic data. [7] notes that with this technique it is possible to control demographic distribution across the dataset and still include realistic intra-class variations. The results show that the synthetic data can be used to mitigate demographic biases in face recognition tasks.

However, a study by [1] points out that attribute manipulation, such as changing face expression or adding make up, can introduce some artifacts in the constructed images. Furthermore, the authors express their doubts that generated faces will contain appearance variations that is crucial for training the model.

Mentioned research also highlighted the challenges present in investigating racial bias in face identification task. To avoid overfitting, the state-of-art models require large-scale quality datasets which are difficult to get, not to mention data scarcity for under-represented groups. With large datasets, it is also important to note the computational cost of the model.

3 Dataset and Features

We used VGGFace2 Mivia Ethnicity Recognition (VMER)[4] large-scale dataset specifically designed for ethnicity recognition tasks. It comprises over 3 million images of almost 10 thousand individuals of different genders, ethnicities and ages (see Figure 2). The images were annotated with one of the four labels: African American, East Asian, Caucasian Latin, and Asian Indian. The dataset is approximately gender-balanced but highly imbalanced in terms of race as seen in the Figure 1.

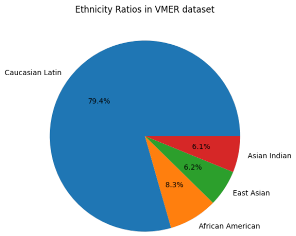


Fig. 1: Ethnicity Ratios in VMER Dataset

The VMER dataset primarily sources its images from the VGGFace2 dataset[2] which is large-scale face recognition dataset. The images are sourced from Google Image Source and vary in pose, illumination, age and ethnicity.

While the details of the labelling process are not shared, [4] notes that the annotation protocol included opinions of people belonging to different ethnicities. In this way, it is said that the bias introduced by other race effect was avoided.

4 Methods [1-1.5 pages]

In order to compare face recognition accuracies across four ethnicity classes outlined in the VMER dataset, we created racially aware subsets from it. The representation ratio of African American class

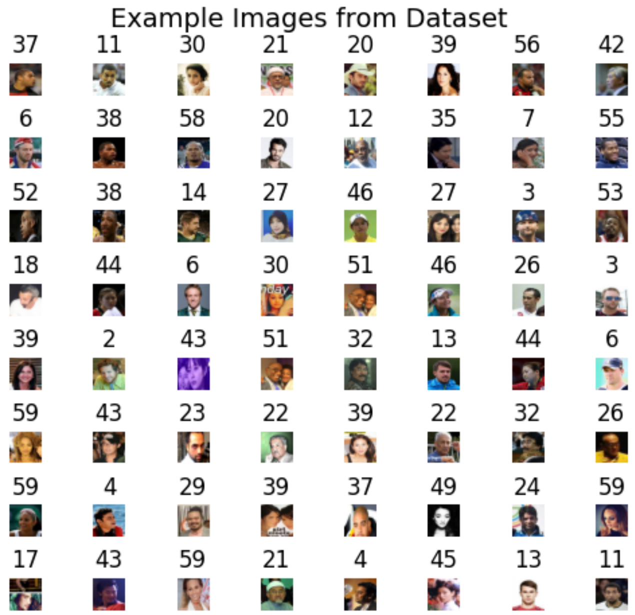


Fig. 2: Example data from VGGFace2 Image Dataset (without MIVIA labels)

was systematically varied to observe the effects of data imbalance on model performance. All created subsets contain 50 images per individual and are described in Table 1.

Tab. 1: Ethnicity ratios within each subset.

representation	Caucasian Latin	East Asian	African American	Asian Indian
original	46	4	6	4
equal	15	15	15	15
moderate	18	12	18	12
increased	11	15	17	15
dominant	9	7	37	7

In this investigation, VGG16 model trained on created image datasets. It is a part of a specially designed pipeline framework used to compare model performance in the face recognition task based on the dataset it was trained on.

4.1 Preprocessing

The images underwent a preprocessing stage before they entered VGG16 model. First, each image was resized to 224 × 224 pixels, the required input size for the model. Next, the pixel values were normalized to a range between 0 and 1. The resulting images preserved their original RGB format which allowed the model to also learn from the colour cues.

4.2 Modelling

The VGG16 model is built using Keras API and takes in input images in the form of 3D $3 \times 224 \times 224$ arrays.. The model architecture (look Figure 3) is comprised of 16 layers with learnable weights, 13 of which are convolutional layers and 3 are fully connected layers. The convolutional layers extract features from the images using smaller filters which allows the model to incorporate more non-linearities into the network and hence, boost its ability to learn complex patterns. Following max pooling layers are used to reduce spatial dimensions or feature maps and hence computational effort of the model. The final layers are fully connected, with the last one having 60 neurons for 60-class classification.

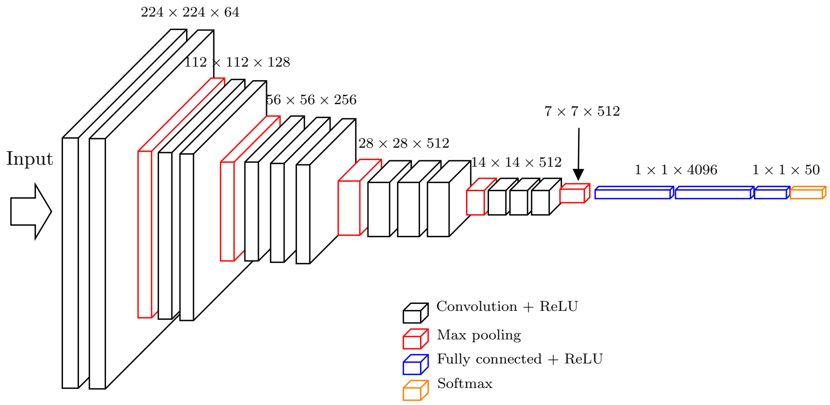


Fig. 3: VGG16 Architecture[5]

ReLU is the activation function used throughtout the model which introduces nonlinearity and allows it to learn complex patterns. However, the last layer of the network uses softmax function to create a probability score for each class. The class with the highest probability is chosen as the predicted one.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \quad i = 1, \ldots, C \tag{1}$$

4.3 Training

Once the VGG16 model architecture was set up, the training process was conducted the created racially aware subsets. However, only Equal Representation dataset was used for the testing step. Each dataset was split into three parts, for training, cross-validation and testing, according to the 3:1:1 ratio.

To train the model efficiently and effectively, Adam optimiser was user with a learning rate of 0.0001. Furthermore, the use of categorical cross-entropy loss ensured that the model would assign higher probability values to the correct class during training using one-hot encoding. The model trained for up to 60 epochs in baches of 16.

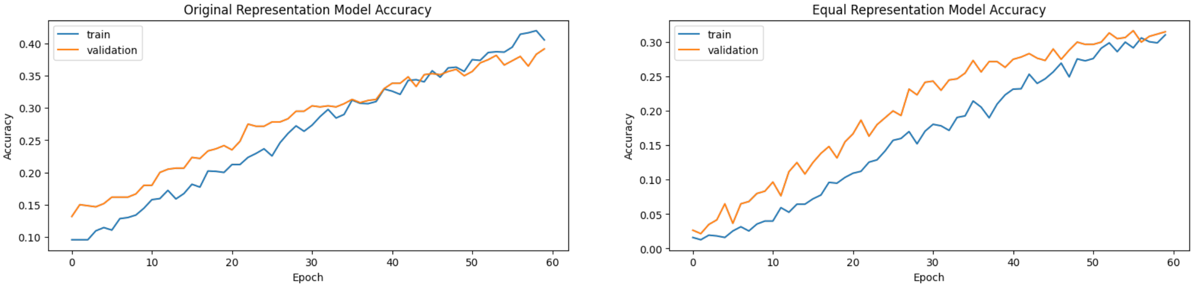
To regularize the training process and prevent overfitting, `EarlyStopping` and `ReduceLROnPlateau` callbacks were utilised. The first one ensures that the training stops if the validation loss does not improve for 10 consecutive epochs. In that case, the best-performing model is restored. The latter callback reduces the learning rate by a factor of 0.2 when the validation loss plateaus, allowing for finer updates to the model weights.

5 Results and Discussion

Both performance and accuracy of the VGG16 model is assessed with accuracy learning loss metrics. Furthermore, confusion matrices were used to provide a detailed breakdown the model's performance.

5.1 Model Training

The model trained on multiple datasets. We distinguish the cases in which VGG16 model trained on the Original Representation (OR) and Equal Representation (ER) datasets and display learning patterns in Figures 4a and 4b. Both curves are increasing which means that the model is indeed recognising and somewhat generalising patterns within the datasets. However, there are moderate fluctuations that could suggest that it has difficulties learning the patterns from both datasets, especially from ER dataset.

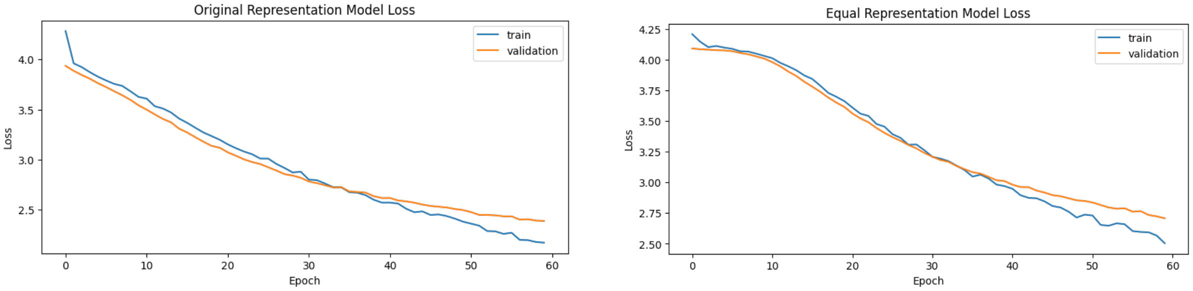


(a) Original Representation Model Training

(b) Equal Representation Model Training

Fig. 4: Model Accuracy Plots

The gap between the training and validation curves narrows which suggests that until some point model was learning generalisable patterns from the data. However, around epoch 40 for the OR, training accuracy surpasses the validation accuracy and the gap begins to widen. This might suggest that the model is trying to memorise the patterns it found rather than learning them. In contrast, training on ER dataset yielded validation accuracy larger than training accuracy practically over the entire training.



(a) Original Representation Model training

(b) Equal Representation Model training

Fig. 5: Model Learning Loss Plots

Learning loss of the model trained on the Equal Representation dataset decreases rapidly as shown in Figure 5b. Furthermore, training and validation curves come together before 30th epoch which would suggest a good generalisation.

On the other hand, in Figure 5a shows a widening gap between training and validation learning losses. This would indicate overfitting, especially since the validation learning curve could start increasing if the training was not stopped.

5.2 Model Performance

To measure the model performance across different ethnicities, we quantified how well the VGG16 model performs a face recognition task within the four ethnicity classes and illustrated that using confusion matrices.

Figure 6 shows that the model did not perform well in the face recognition task across different ethnicity groups. There is a visible bias into the Caucasian Latin ethnicity class as its individuals get identified accurately significantly more often. In addition, the lack of the diagonal character that is present in Figure 7, suggests that the OR model has comparable amounts of false identifications and true identifications.

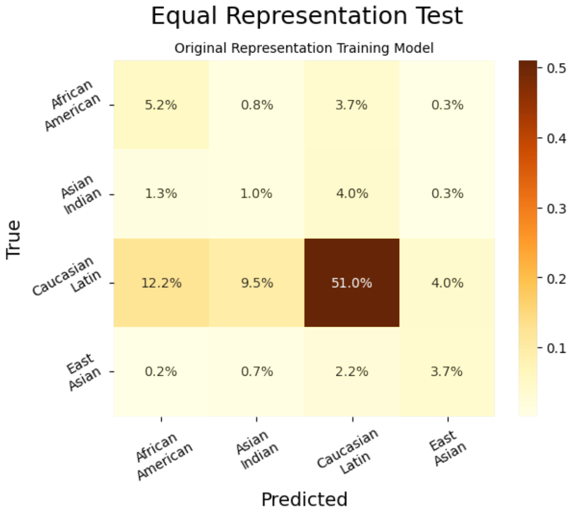


Fig. 6: Confusion Matrix For Original Representation Model on Equal Representation Dataset

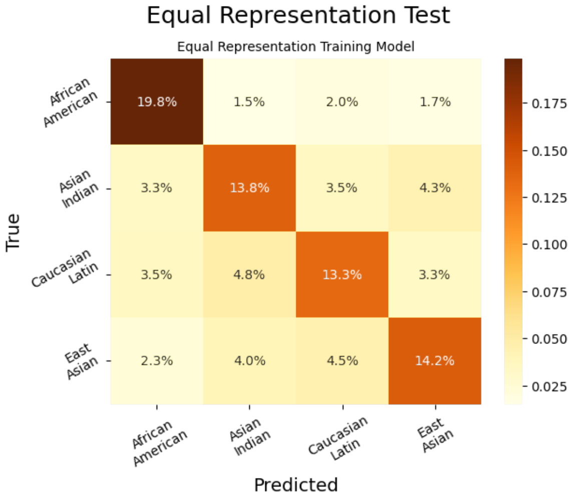


Fig. 7: Confusion Matrix For Equal Representation Model on Equal Representation Dataset

On the other hand, after the model had trained on ER dataset, it was able to accurately identify individuals across different ethnicity groups. We also notice that the wrong identifications are also less probable. We also note the sudden relative accuracy boost for African American ethnicity class.

In cases involving Moderate Representation (MR) dataset, we also note that the model trained on OR dataset even had difficulties in determining the identity in certain cases. In other words, the identity was not precisely predicted.

We further interpret the results by computing F1 scores that are presented in Table 2. Overall, the model does not perform well regardless of the dataset it trained on as the highest F1 score is below 0.5. This indicates that there is a room for improvement in the classification accuracy across all ethnic groups.

Nevertheless, the gap between ethnic groups narrowed when the model trained on ER dataset while keeping a similar trend to the Original Representation model of Caucasian Latin class having the highest score and Asian Indian the lowest. Furthermore, identification within East Asian class improved the most by changing the classification model which suggests that this class was severely under-represented in the Original Representation dataset. Also, African American class identification substantially improved.

The VGG16 model performs better when trained on ER dataset which suggests that the OR dataset caused it to favour specific ethnic groups. Nonetheless, even when trained on a more balanced dataset, the model still struggles to classify all ethnic groups accurately, particularly Asian Indian and African American. This suggests that the model should train an much larger datasets than those created for the purpose of this exploration.

Tab. 2: F1 scores for ethnicity classes within Original Representation and Equal Representation Models.

	Caucasian Latin	East Asian	African American	Asian Indian
Original Representation	0.47	0.21	0.34	0.39
Equal Representation	0.67	0.56	0.48	0.47

6 Conclusion

In this study, we investigated the impact of racial imbalances within datasets used to train facial recognition models. We evaluated the performance of a VGG16-based model across racially aware subsets. The model demonstrated how varying the representation of under-represented groups, particularly the African American category, influenced the accuracy of identity recognition. Subset with more balanced ethnic distributions led to improved performance across all groups, highlighting the importance of equitable representation in training data. Nevertheless, the model performance leaves a significant amount of room for improvement as generally in neither case the model was particularly successful in this classification task. This contributes to the notion that we need large-scale image datasets for face recognition tasks.

For future work, with access to more time and computational resources, we would explore techniques like data augmentation, synthetic data generation, or reweighting loss functions could be explored to mitigate biases further. Additionally, alternative deep learning architectures, such as EfficientNet or ResNet, to determine if they handle imbalanced data more effectively.

Contributions

Both authors designed the overall study and contributed to creating the code for racially-aware datasets. Wiktoria drafted the initial model architectures, before the decision to use VGG16 model. Dele constructed the final model architecture and pipeline. Wiktoria designed and finalised the poster, interpreted the results and prepared the initial draft of the report. Dele contributed to report revision.

References

[1] Fadi Boutros et al. “Synthetic Data for Face Recognition: Current State and Future Prospects”. In: *Image and Vision Computing* 135 (July 2023), p. 104688. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2023.104688. (Visited on 12/07/2024).

[2] Qiong Cao et al. “VGGFace2: A dataset for recognising faces across pose and age”. In: *CoRR* abs/1710.08092 (2017). arXiv: 1710.08092. URL: http://arxiv.org/abs/1710.08092.

[3] Jacqueline G. Cavazos et al. *Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?* 2020. arXiv: 1912.07398 [cs.CV]. URL: https://arxiv.org/abs/1912.07398.

[4] Antonio Greco et al. “Benchmarking deep network architectures for ethnicity recognition using a new large face dataset”. In: *Mach. Vision Appl.* 31.7–8 (Sept. 2020). ISSN: 0932-8092. DOI: 10.1007/s00138-020-01123-z. URL: https://doi.org/10.1007/s00138-020-01123-z.

[5] hongvin. *Neural Network Architectures in LaTeX*. https://github.com/hongvin/Neural-Network-Architectures-in-LaTeX/tree/main. 2020.

[6] David Masko and Paulina Hensman. “The Impact of Imbalanced Training Data for Convolutional Neural Networks”. In: 2015. URL: `https://api.semanticscholar.org/CorpusID:46063904`.

[7] Pietro Melzi et al. “Synthetic Data for the Mitigation of Demographic Biases in Face Recognition”. In: *2023 IEEE International Joint Conference on Biometrics (IJCB)*. Sept. 2023, pp. 1–9. DOI: `10.1109/IJCB57857.2023.10449034`. (Visited on 12/05/2024).

[8] Joseph P Robinson et al. *Face Recognition: Too Bias, or Not Too Bias?* 2020. arXiv: `2002.06483 [cs.CV]`. URL: `https://arxiv.org/abs/2002.06483`.

[9] Andrew Sumsion et al. “Surveying Racial Bias in Facial Recognition: Balancing Datasets and Algorithmic Enhancements”. In: *Electronics* 13.12 (Jan. 2024), p. 2317. ISSN: 2079-9292. DOI: `10.3390/electronics13122317`. (Visited on 12/05/2024).