Investigation into Ethnicity Bias in Face Recognition Task

Wiktoria Radecka [1] Dele Ayeni[1]

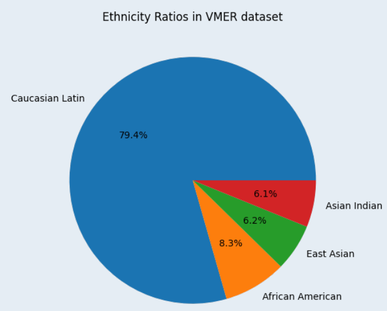
[1]Faculty of Engineering Technology

Introduction

In recent years, the exponential growth of artificial intelligence (AI) across various sectors has revealed significant discrimination, particularly affecting minority populations. Given the prominence of racial bias, the related research has focused on its causes as well as mitigation methods.

The project aims to investigate how data imbalance contributes to ethnic disparities in machine learning models used for face recognition, with the goal of addressing and mitigating these biases.

Dataset and Features



We used VGGFace2 Mivia Ethnicity Recognition (VMER) large-scale dataset designed specifically for ethnicity recognition task. It includes the large-scale VGG-Face2 dataset with over 3 million images of almost 10 000 individuals labeled with one of four dedicated ethnicity labels (see Figure 1). The images vary in pose, age, illumination, gender and ethnicity which exposes the model to a vast range of visual features.

Figure 1. Ethnicity Ratios in VMER Dataset

Methods

The effect of the data imbalance is evaluated using the pipeline illustrated in Figure 2. After creating datasets with varying ethnicity ratios, we standardise the images by resizing them to 224 by 224 pixels. The data is then split into training, cross-validation and testing sets. The images in the form of 3D matrices are inputted into VGG16 model shown in Figure 3.

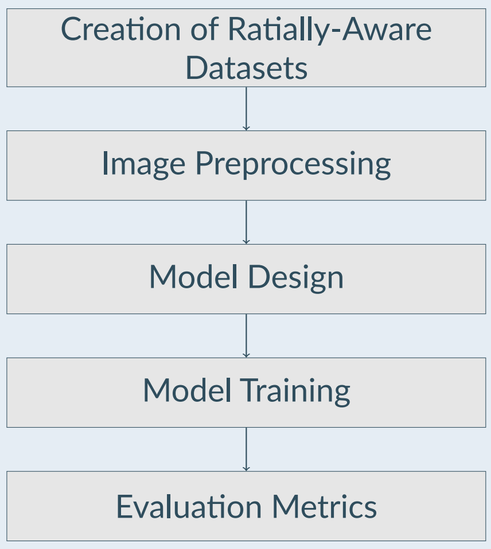


Figure 2. Face Recognition Pipeline

For training and validation only one subset with specific ethnicity representation is used. For testing, we utilise testing sets from different subsets. Finally, we plot confusion matrices to show how accurately the model predicts person's identity within each ethnicity class.

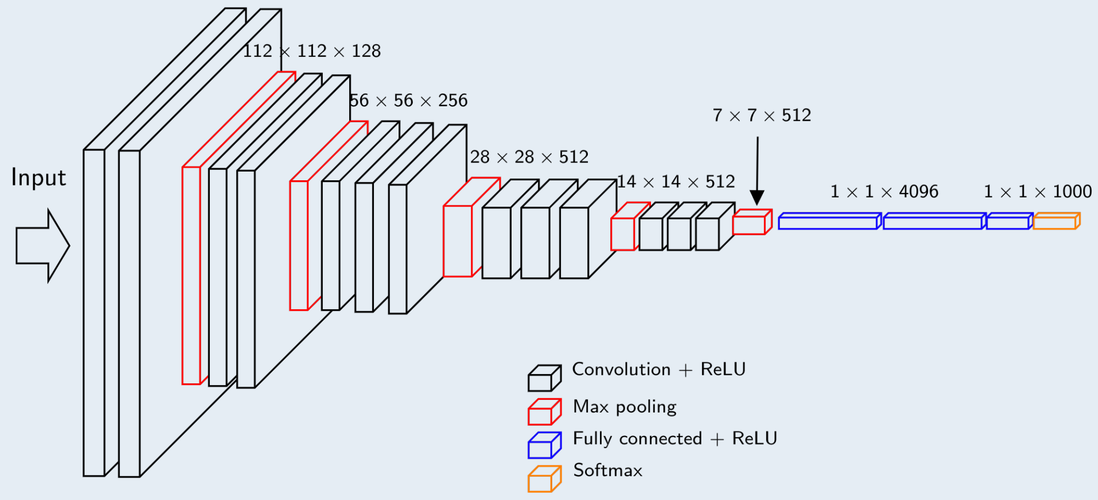


Figure 3. VGG16 Model Architecture

Results/Discussion

In the Figure 4, we plot VGG16 model's **accuracy** and **learning loss** metrics. From the training and validation values in the accuracy plot, we can infer that the model initially learns general patterns in the data. However, later on as the training accuracy surpasses validation accuracy, the model begins to memorise the training data.

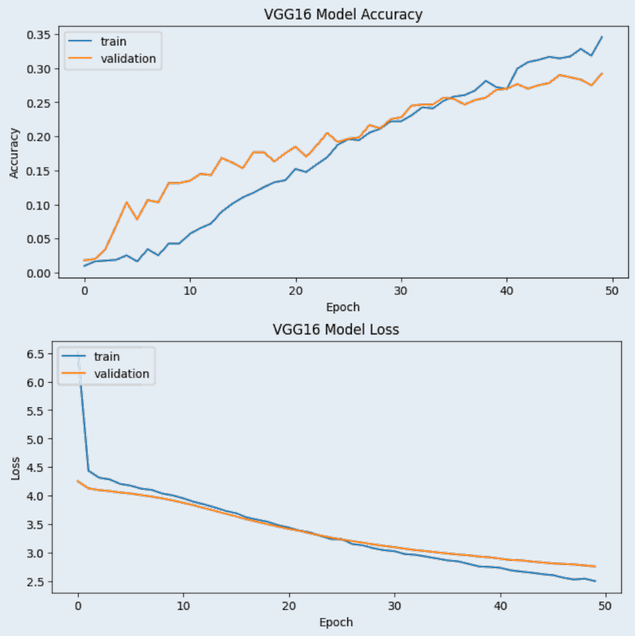


Figure 4. Model Training Performance

We present our results in form of **confusion matrices**. From Figure 5, it is evident that the dataset on which the model trained influences the accuracy with which it predicts later labels.

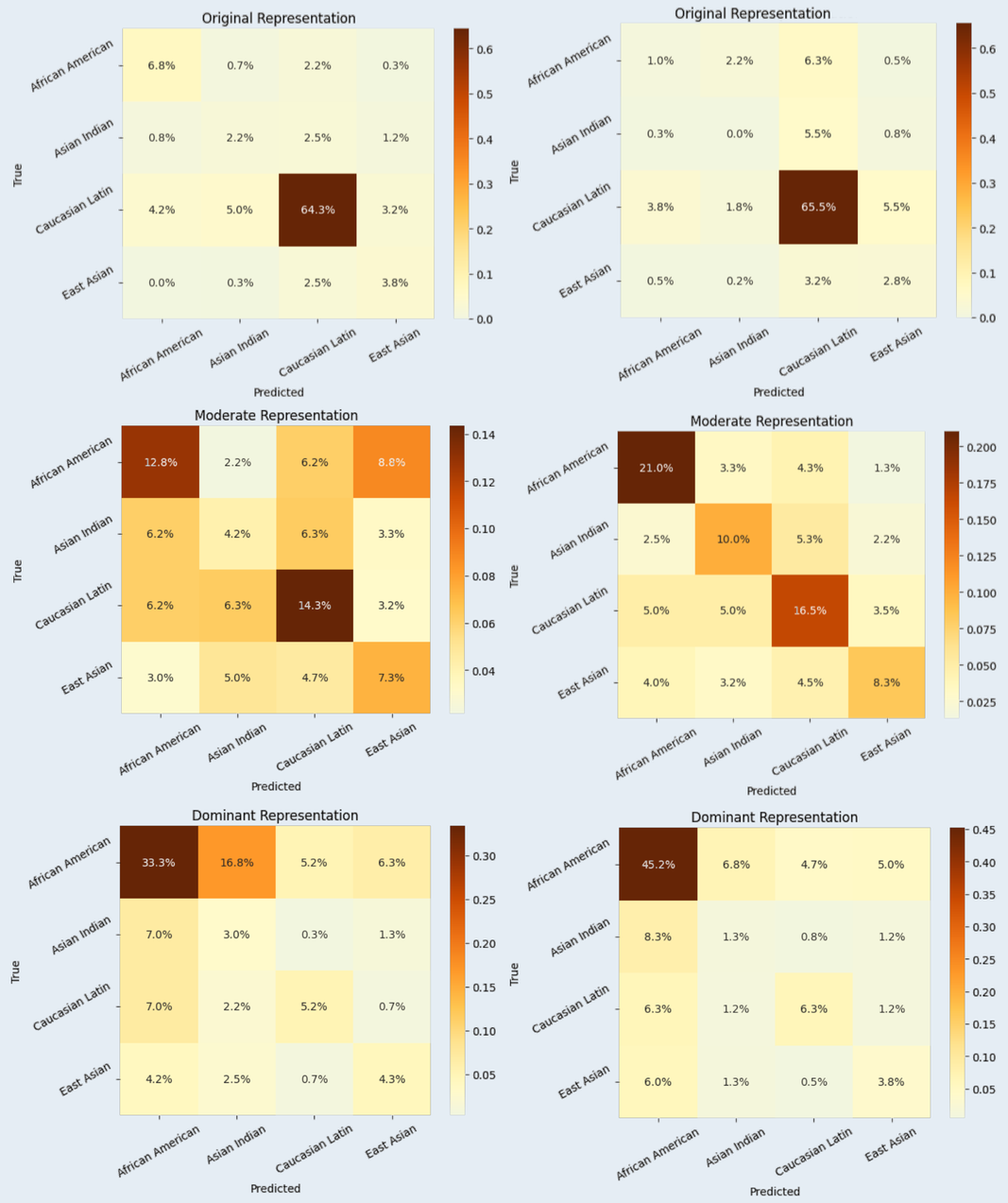


Figure 5. Confusion Matrices for Face Recognition on Diverse Datasets (left original representation, right moderate representation)

1. **It is difficult to pinpoint what causes the model to overfit**. It can be that it is caused by the reduced dataset size as well as training over a large number of epochs.
2. **Class imbalance significantly influences accuracy scores within each class**. The model seems to always favour the majority class.
3. **Representation within the test dataset affects the number of true positives**. This is because there is a limited number of possible true positives in each class.

Conclusion/Future Work

Dataset imbalance has a significant effect on the accuracy scores for each ethnicity group. The dataset on which face recognition model trains influences how the model performs post-testing. Possible solutions inlcude data augmentation.