(Automatic) Speech Recognition

Race Bias on Emotion Recognition with ASR

Marta Español López
s1103330

June 2023

Radboud Universiteit

1 Introduction

Problem: Recently it has been noticed that Artificial Intelligence (AI) models may not perform equally good though all the classes present in a dataset. Not generalizing properly raises many ethical concerns that propitiate discrimination and inequality. AI models can be used in various decision-making processes, such as hiring or criminal justice. If this models exhibit any bias, it can result in unfair outcomes for individuals pertaining to a specific class. The bias can also come from the data used, not from the model. Datasets may have a limited representation, this is, lacking diversity, presenting an over-representation of a certain class and the under-representation of another one, or they can also be indequately labelled. This can result in reduced accuracy or increased error rates for some subgroups, marginalizing and/or excluding even more those groups. For this research study we will be focusing on whether there is a race-dependent bias when predicting emotion from speech.

Background: With the increase of popularity of AI, there have been appearing some concerns regarding bias on the performance of the model, where it does not work equally well for all the subgroups of the population [1]. Bias in emotion recognition can be observed due to different features e.g. gender, language, race or other demographic categories [2]. There have been recent reports of racial bias in the field of Automatic Speech Recognition (ASR), where models perform poorly on Black people, when processing African American Language (AAL) [3]. There are several approaches to reduce the existing bias in an AI model [4]. For instance, by balancing out the classes in the dataset, or training the model in a weighted manner, where the over-represented class is given less weight.

Research Question: Is an ASR model equally robust when predicting emotion from different race data?

2 Method

For this study the general workflow is the following:

1. Pre-process of the audio samples. Including the creation of a dataframe with the corresponding labels for further analysis.
2. Perform data augmentation and re-do Step 1.
3. Iterate over the audio files and extract the mel-frequency cepstral coefficients (MFCCs) for modelling. Add them to the existing dataframe.
4. Split the data into training (train) and testing (test) subsets.
5. Z-normalize the data into zero mean and unit variance $N(0, 1)$.
6. On hot encode the 6 different classes.
7. Define a baseline (CNN) model.
8. Train the model on the train split of the data, and save the best checkpoint, while monitoring the accuracy and loss.
9. Obtain the emotion predictions on the test split and visualize the overall results and the per-race results.
10. (Optional) Hyper-parameter tuning with grid-search.

Since our object of study is whether there is any kind of bias towards predicting emotion for a certain race we will focus more on the post-modelling analysis, rather than in obtaining a good overall performance by using a complex model.

As for the different experiments performed, the model was first trained on the baseline model only with the existing data (biased). Then a series of three experiments followed after doing data augmentation. For each of these experiments a distinct noise value was chosen when generating the new data, those being $0.005, 0.05$ and 0.1. This four experiments were then repeated a minimum of three times to ensure that the results were not due to some stochastic effect. In the next section (3) you can find information about how the method is applied and it is described more in depth.

3 Set-up

3.1 Exploratory Data Analysis

To carry on this task we chose the CREMA-D dataset [5] which stands for Crowd-sourced Emotional Multimodal Actors Dataset. It has a collection of 7,442 clips from 91 actors, of which 48 are male and 43 female actors between the ages of 20 to 74 years old, coming from different races, namely African American, Asian, Caucasian or Unknown for the actors with an Unspecified race. And from different ethnicity, those being Hispanic or non-Hispanic. Each one of the actors spoke a selection of 12 sentences with different emotions (anger, disgust, gear, happy, sad and neutral) and different emotion levels (low, medium, high or unspecified). For this task we are only focusing on the emotion, and not on the emotion level. The dataset includes a video clip of each given sentence, as well as just the audio files in MP3 and WAV format. For this study we are using the WAV files only. The labels for the data were obtained by crowd-sourcing, where a total of 2443 participants rated the emotion and emotion levels based on the combination of the video clip and audio, the video alone, and the audio alone. So, each sample has more than 7 ratings.
There was an actor in the dataset for which the race was not specified ("Unknown"), so for the purpose of this study we decided to discard the samples belonging to this actor (Actor 1047), since the race of each actor is of our interest. Luckily there was only one actor for which the race was unspecified.

This dataset seemed to be a good choice for this task due to the specification of each actor's race, as well as providing a fair number of different emotions to predict, and a good sample size.

When doing an Exploratory Data Analysis (EDA) on this data we can see that the samples per emotion are nicely balanced as it can be seen in Figure 1, with a slight decrease on the neutral emotion, but the difference does not appear to be significant for a correct prediction. Opposed to that, when inspecting the distribution of races we obtain a much different result, where the distribution is clearly skewed and biased towards the Caucasian race, which has a much higher sample count than African American or Asian, as it can be seen in Figure 2. To handle this bias, data augmentation is performed so new samples are generated for the underrepresented classes. In that way we can check whether the bias on the predictions for each given race is due to some specific speech feature, and not only because the dataset is not balanced. In the next sub-section this procedure is explained in more detail.

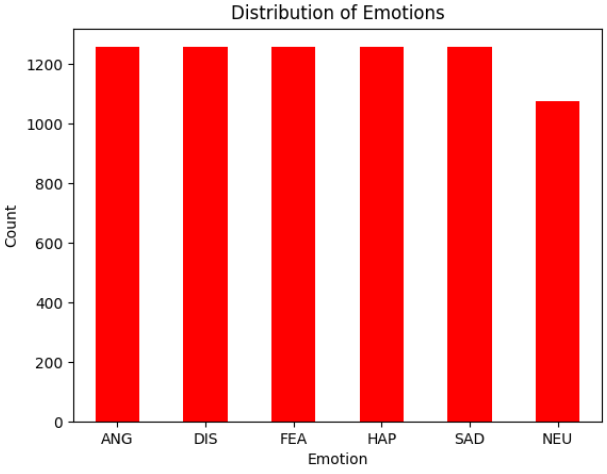


Figure 1: **Distribution of emotions.** Across all the samples in the dataset. On the x-axis each given emotion: angry (ANG), disgust (DIS), fearful (FEA), happy (HAP), sad (SAD) and neutral (NEU). On the y-axis, the count of samples per emotion.

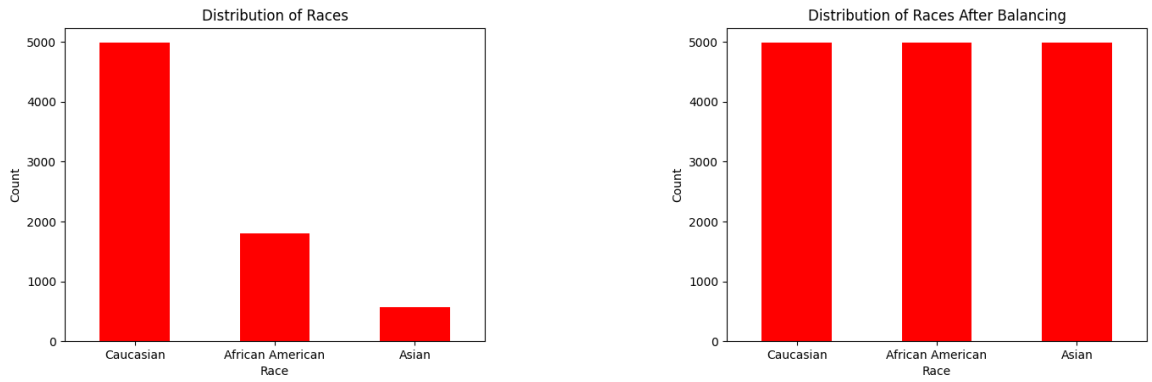


Figure 2: **Distribution of races.** On the right, the distribution before performing data augmentation, on the left, the distribution after performing data augmentation. On the x-axis each given race. On the y-axis the count of samples per race. For both plots, the samples with race: "Unknown" was discarded.

3.2 Data Augmentation

From our understanding there are two reasons why there could be a bias on the model's predictions depending on a certain race. Firstly, the bias could be due an unbalanced dataset where there is one clearly over-represented class (Caucasian), and two under-represented classes (African American and Asian). Or, secondly, there may be an actual bias of the model that yields a better performance for the Caucasian class, than for the other two, not dependant on the number of samples per class. To assess this problem, data augmentation is performed in order to balance the dataset by generating new samples for the underrepresented classes. This is done by randomly selecting samples from the underrepresented class and adding a specified amount of white noise at random time points of the audio. White noise is characterized by having an equal amount of energy at every frequency within the specified range. By doing this we end up with the same amount of samples per class, as it can be seen again on Figure 2.

3.3 Model Architecture

The model used is a simple Convolutional Neural Network (CNN) composed of three 1D convolutional layers followed by two dense layers after flattening, and combined with three dropout and two 1D max pooling layers, as it can be seen in Figure 3. Next to each type of layer there is the output size in parenthesis, and underneath, when required, the appropriate activation function. The last dense layer has output size 6 since we are predicting 6 different emotions (classes). The total number of parameters for the whole model is 314,054.

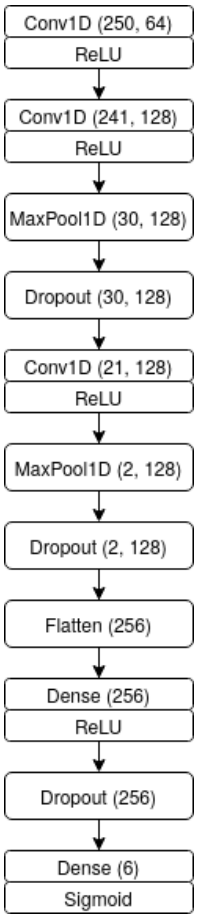


Figure 3: **CNN model architecture.**

The reasoning behind using a simple CNN model is because they are computationally efficient during the inference phase, and are relatively easier to interpret compared to more complex models, by examining the filters we can gain insight on which speech characteristics are relevant for emotion prediction across different races. CNNs also usually have less parameters than other complex models. And they are known to be good at capturing local and global patterns regardless of their position in the input. Although a more complex model such as an RNN or a model that uses some kind of attention mechanism could have yielded better results, our main goal was to study whether the model got biased by some race dependent speech feature, so, while still important, the overall accuracy of properly predicting the emotions was left in a second plane.

3.4 Hyper-parameter Tuning

Since hyper-parameter tuning is computationally expensive, taking over three hours to complete, it was only done for the best performing model (Baseline with data augmentation and $\eta = 0.005$). To find the best values for each parameter a grid search was performed. This brute-force approach provided us with the optimal combination of hyper-parameter values for our model. The manually specified subset of different hyper-parameter values to combine and evaluate include the following:

- Batch size: 30, 36, 42
- Number of epochs: 25, 50, 75, 100
- Optimizer: Adam, Stochastic Gradient Descent (SGD)

The hyper-parameters were previously set to: batch size = 32, number of epochs = 40 and SGD optimizer. After performing the grid search it yields that the best hyper-parameters for this model are: batch size = 36, number of epochs = 75 and still SGD optimizer. Re-training the model with the best hyper-parameters did not bring a significant improvement on the model's performance.

3.5 General Experimental Set-up

Environment and implementation: We chose a baseline notebook that performed an analysis on the same dataset of our choice. We then added the corresponding modifications to adapt it to our research questions. The whole implementation [1] is in Python and it is embedded in an Interactive Python Notebook (.ipynb file). The main additions to the notebook are marked as "newly added" next to the corresponding cell title. All the experiments were run in Google Colaboratory which offers a total of 12.7GB of RAM and 107.7GB of disk storage. The whole notebook takes around 1 hour and 30 minutes to run, excluding the hyper-parameter tuning, which is the most costly step.

Experimental decisions: For this study we divide the data into a training split (train) and a testing split (test) in a 80% and 20% proportion respectively. This is done to check if the model is able to generalize properly with unseen data after training it. So, the ability to handle real-world data is put into test. It also helps to detect if the model over-fitted with the training data, and thus, does not generalize adequately. Four different experiments are performed using the same CNN model described above, focusing both in the overall performance of the model, and the how the model performs for each specific race. The experiments include: a baseline run without data augmentation, and three other experiments with different values of noise for the newly generated data when balancing the dataset. All four experiments were repeated at least three times to ensure that the results are not because of stochasticity. The experiments will be explained in a bit more detail in Section 4. As for the evaluation of the results, the model's overall accuracy is computed, as well as the per-race accuracy. Moreover, for each of the four experiments, the precision, recall and f1-score are computed for each emotion, for the whole model predictions, and for each race individually.

3.6 Evaluation Metrics

To obtain a comprehensive assessment of the model's performance for each experiment, and to investigate the research question we have to compute several statistics so we can determine whether there is bias or not race-wise.

Overall accuracy: Accuracy of the emotion prediction model across all races together. It measures the percentage of correct predictions made by the model. It provides an overall view of the performance of the model in general, not focusing on each emotion separately.

Per-race accuracy: Accuracy of the emotion prediction model for each individual racial group separately. Provides a view of the model's performance only on a specific race, and same as the previous metric, it focuses on the genral performance, not on each emotion separately.

[1] All the code and results for each run is available at: https://github.com/Martaesplo/Emotion_Rrecognition_ASR.git

Confusion matrix: Provides a representation of the classification performance of the model, by comparing the predicted labels to the actual ones (ground truth). Performed for both, the overall model predictions and the per-race predictions in order to check for biases or miss-classifications in the predictions.

Precision and Recall: Precision measures the proportion of correctly recognized emotions out of all the positively predicted instances. It focuses on minimizing the false positives, while recall focuses on minimizing the false negatives, this is, the tendency of the model to fail to recognize the correct emotion. These two metrics offer an overview of the error types, namely, false positives and false negatives. A low precision would indicate that there are many false positives, and low recall would suggest that the model is missing a high number of positive samples. This is computed per-emotion, not for the whole model.

F1-score: It combines the precision and recall into a single value, by computing the harmonic mean precision and recall for each emotion. It provides a balanced measure of the model's accuracy considering both false positives and false negatives. This is computed per-emotion, not for the whole model.

4 Experiments

For all the experiments the same hyper-parameters were used as stated in 3.4; in order to keep a controlled environment as much as possible. And the same evaluation procedures were followed. All four experiments were repeated a minimum number of three times.

4.1 Before Data Augmentation

A first experiment was performed with the data available in the dataset, which, as mentioned in section 3.1 is biased towards the Caucasian class. This counts as a control experiment, on which we can later compare the performance of the experiments using a balanced dataset. After training the model the analysis performed included the creation of a confusion matrix comparing the predictions made with the actual labels, as well as a confusion matrix for each race that contains only the predictions belonging to the given race. The reasoning behind doing three extra matrices per-race is to evaluate whether there is a bias not only on the overall accuracy of correct predictions, but also to check whether there is a bias for a certain emotion to be predicted better for one race than for another.

4.2 After Data Augmentation

In this part of the study we perform three different experiments. New data is generated by adding white noise to randomly selected samples from the underrepresented classes, in this case, the African American and Asian classes. The three different amounts of noise added are: 0.005, 0.05 and 0.1. Different values of noise were chosen to investigate its effect on the model's performance, and make sure that we don't loose too much overall accuracy when generating new noisy data. After this procedure the pipeline followed is the same as for the experiment without data augmentation. So, for each different noise level, the overall model performance and the per-race performance evaluation metrics are computed.

5 Analysis & Results

In this section the results from the experiments are reported, they are commented and discussed in the next Section 6. Further results: monitoring of accuracy and loss over training time, confusion matrices for the remaining two runs and the corresponding metrics can be found in the GitHub repository under the folder named "results".

Experiment	Overall model	Caucasian	African American	Asian
Baseline	40.35	68.74	23	8.25
Baseline $\eta = 0.005$	44.33	32.58	30.87	36.54
Baseline $\eta = 0.05$	38.3	36.29	30.62	33.09
Baseline $\eta = 0.1$	32.63	42.11	29.57	28.3

Table 1: **Mean accuracy across runs.** Contains the mean accuracy of the different (3) runs of experiments, for both, the overall model accuracy, and the per-race accuracy, in percentages (%). The "Baseline" refers to the experiment previous to performing data augmentation. "Baseline $\eta = x$" refers to the experiment after doing data augmentation where η is the x amount of noise added to the newly generated noisy samples.

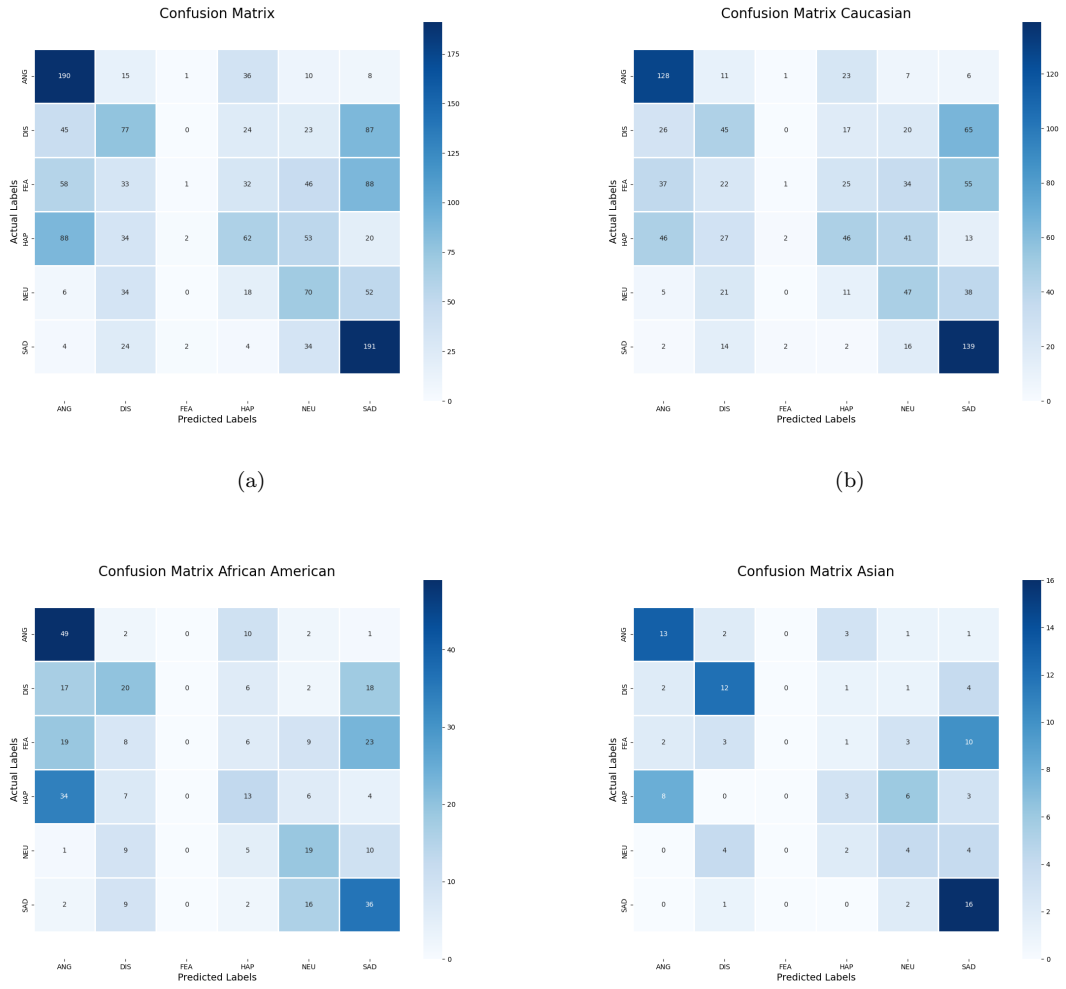


Figure 4: **Confusion matrices for the baseline experiment WITHOUT data augmentation.** On the x-axis there are the predicted labels, on the y-axis the actual ones. (a) Overall model (b) Caucasian only (c) African American only (d) Asian only. Belong to run number 1.

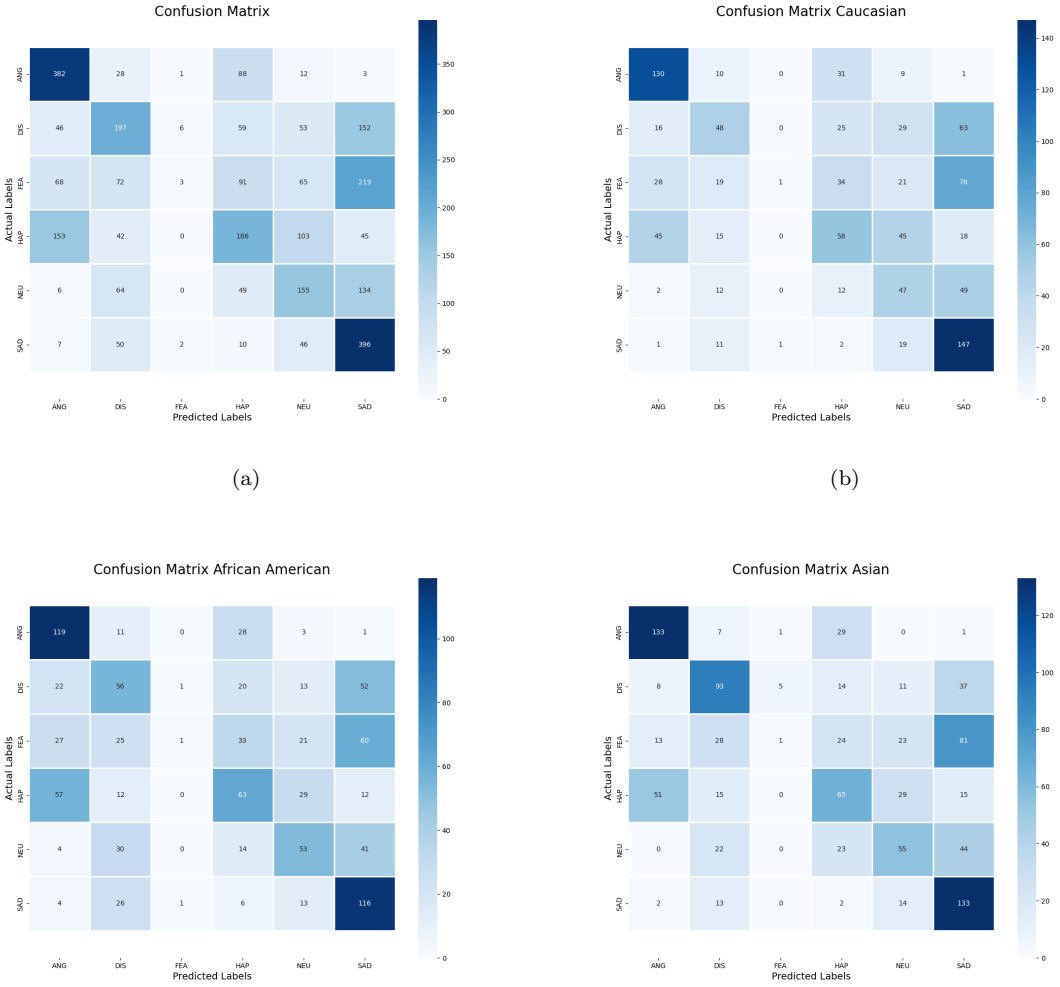
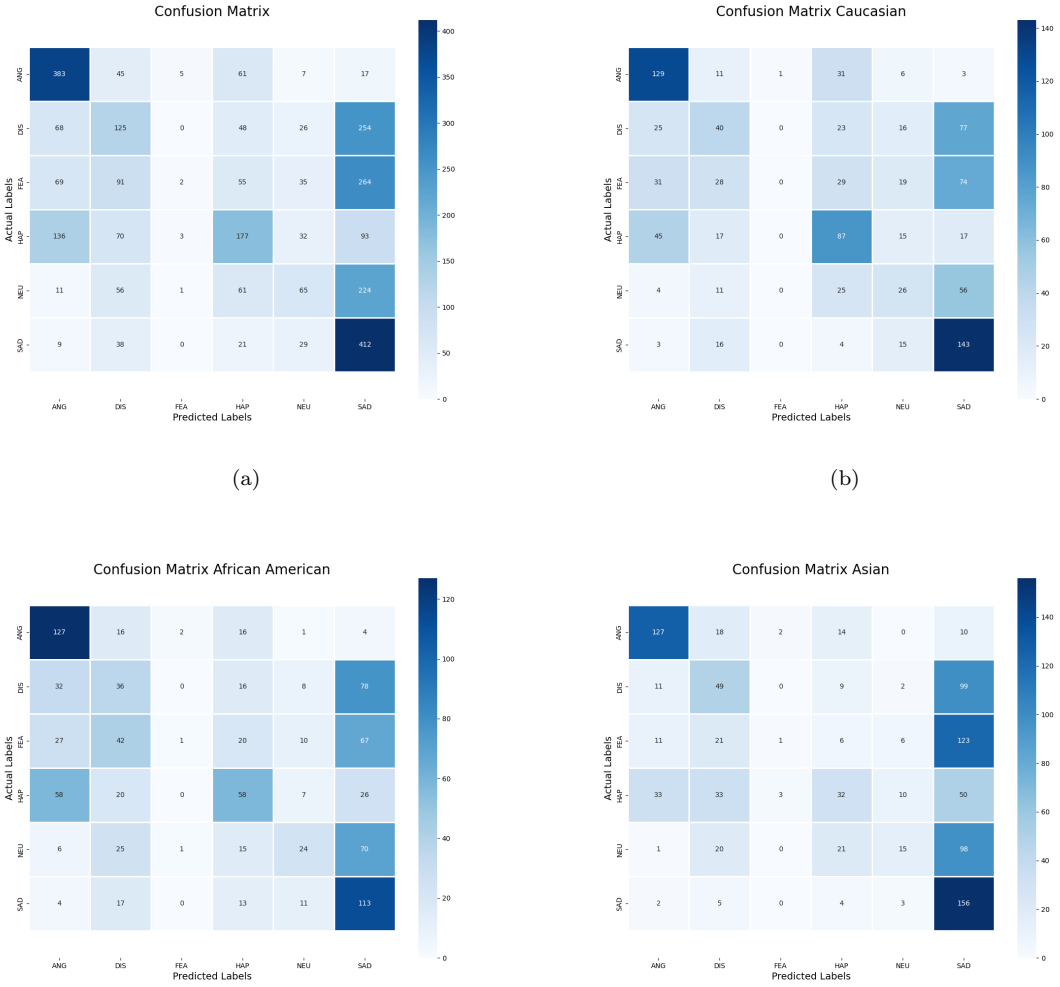


Figure 5: **Confusion matrices for the baseline experiment WITH data augmentation and noise $\eta = 0.005$.** On the x-axis there are the predicted labels, on the y-axis the actual ones. (a) Overall model (b) Caucasian only (c) African American only (d) Asian only. Belong to run number 1.



(a) (b)

(c) (d)

Figure 6: **Confusion matrices for the baseline experiment WITH data augmentation and noise $\eta = 0.05$.** On the x-axis there are the predicted labels, on the y-axis the actual ones. (a) Overall model (b) Caucasian only (c) African American only (d) Asian only. Belong to run number 1.

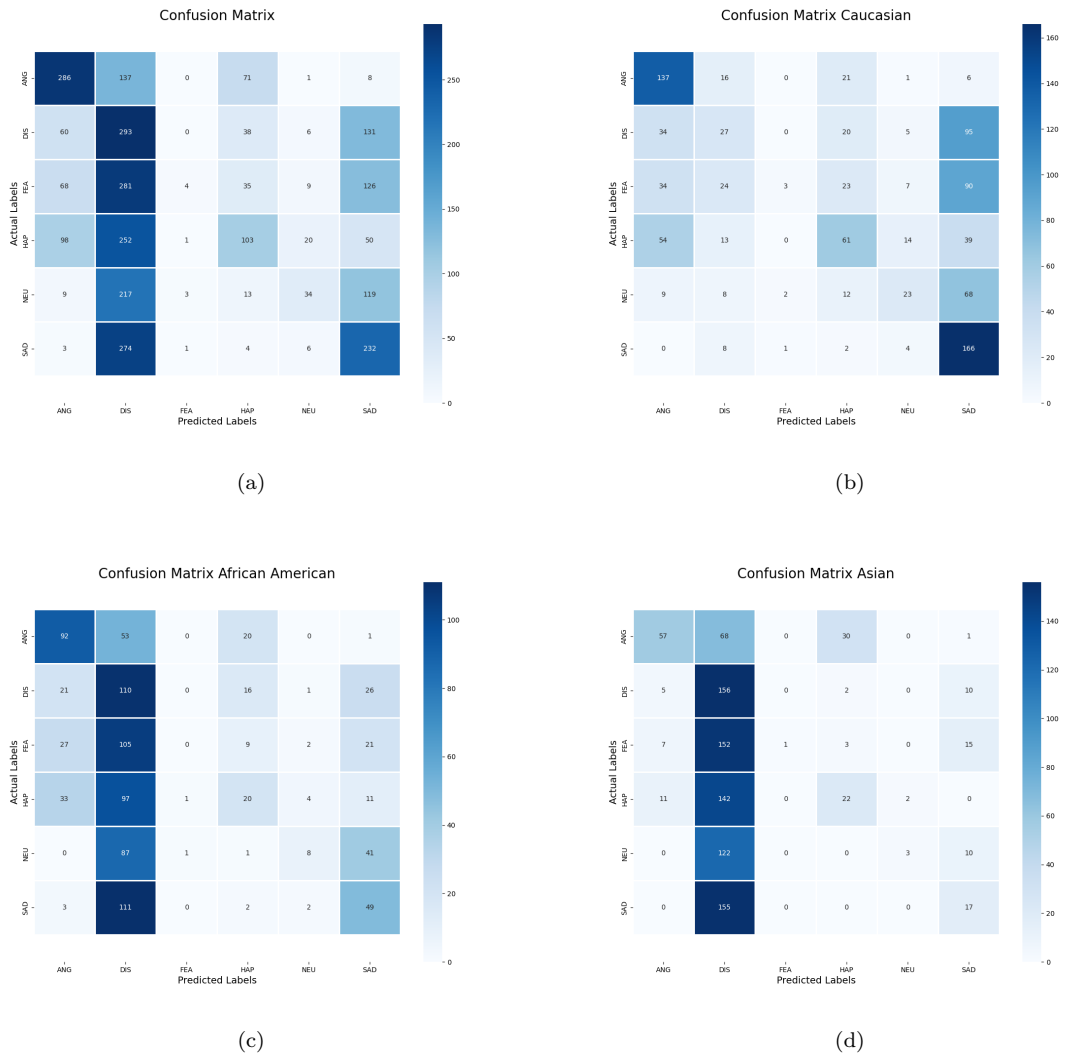


Figure 7: **Confusion matrices for the baseline experiment WITH data augmentation and noise $\eta = 0.1$.** On the x-axis there are the predicted labels, on the y-axis the actual ones. (a) Overall model (b) Caucasian only (c) African American only (d) Asian only. Belong to run number 1.

6 Discussion

Overall performance: Table 1 contains the mean accuracy computed across the three independent runs of the experiments. In general we can say that this model struggles to correctly predict the corresponding emotion, as the maximum accuracy reached is 44.33%. Which belongs to the experiment done with data augmentation and using a the lowest amount of noise ($\eta = 0.005$). It can be appreciated how the performance drops when increasing the amount of noise added to the newly generated samples, the worst performing model being the one with the highest noise level ($\eta = 0.1$) which yields a 32.63% accuracy. This can be expected since when adding noise we are introducing additional uncertainty and variability from the regular patterns present in the training data. And some features may be distorted or masked, so the model may be led to make incorrect predictions. We could attribute the improvement of overall performance of the model with $\eta = 0.005$ compared to the baseline, to having a balanced dataset. So, balancing the dataset introduces a $\sim 4\%$ improvement to its classification ability.

Per-race performance: When looking at the per-race accuracy form Table 1, we can see that there is a clear imbalance when correctly predicting emotions for different races, when we are only using the data given in the dataset. As we would expect, since the Caucasian class is over-represented, it obtains

a much higher accuracy compared to the other two classes. At this point we wondered whether that bias was due to an unbalanced dataset or because there was an actual bias in the model. This motivated the decision of balancing the dataset and performing several experiments with different noise values. After doing so, it can be seen that the imbalance of the correctly predicted emotions per race is almost banished, all the classes obtain a similar accuracy. And the differences in accuracy could most likely be attributed the stochastic effect when training the model.

Per-emotion performance: Comparing the confusion matrices on Figures 4, 5, 6 and 7 we could say that in general the model performs in a similar way when predicting the different emotions for each race. There does not seem to be an emotion that is predicted more accurately in one race than in another. In all the experiments there seems to be a predisposition of the model to correctly predict "anger" and "sad", as the count of correctly predicted instances is quite high compared to the other emotions. The "sad" emotion has a high count of miss-classified predictions, while the "fear" emotion is barely predicted in all the experiments. When listening to the audio files, sometimes the emotion expressed seems a bit ambiguous, which may have led the annotators to miss-label some of the samples, resulting in a bias of the model to classify the "fear" emotion as "sad". The similarity between the audio files belonging to these two classes may have made the model not able to properly distinguish between "fear" and "sad". This miss-classification its most likely not due to the dataset, since the number of samples per emotion is almost the same, all the classes are almost perfectly balanced, thus, this is most likely because of an algorithmic bias.
When looking at Figure 7 it can be observed a huge miss-classification for the "disgust" emotion. For the other two runs, the emotion with the high count of predictions is different, being "fear" and "happiness" for run 2 and 3 respectively. This may be due to the data augmentation step, in which there may be a certain emotion that is randomly sampled more often than the others, ending up being over-represented, and with the high addition of noise the model fails on learning the correct features.

7 Conclusion

To sum up, the best performing experiment is the one with a balanced dataset and the lowest level of noise ($\eta = 0.005$), with a top performance of 44.74%. When the classes are equally distributed, none of the experiments presents bias towards a certain race. As for the emotions predicted and their correctness rate, "anger" and "sad" seem to be the better classified, while "fear" is very poorly predicted.
The results suggest that this specific CNN model is robust when predicting emotion from data belonging to different races. If we consider that the first observable imbalance on the per-race accuracy was solved by doing data augmentation on the dataset, we can think that the model does not generate a bias for any race. Contrary to some research done in this field, as mentioned in the Background section, for this specific dataset Caucasian, African American and Asian speech can be recognized and classified with equal accuracy.

As future work, some improvements on the experimental design of the study could include the comparison of this simple CNN model to other models such as the ones mentioned in Section 3.3. This would provide more insight on the factors that contribute to the bias, and how they interact with other categories that may be also contributing to the bias, such as gender or ethnicity. It would also help on assessing the robustness and generalization capabilities of the model.

References

[1] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[2] Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. Gender De-Biasing in Speech Emotion Recognition. In *Proc. Interspeech 2019*, pages 2823–2827, 2019.

[3] Joshua L Martin and Kelly Elizabeth Wright. Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, 12 2022. amac066.

[4] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, 2023.

[5] Keutmann MK Gur RC Nenkova A Verma R. Cao H, Cooper DG. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *Applied Linguistics*, 5(4):377–390, 12 2014.