DETECTING RACIAL BIAS IN FACIAL RECOGNITION SOFTWARE

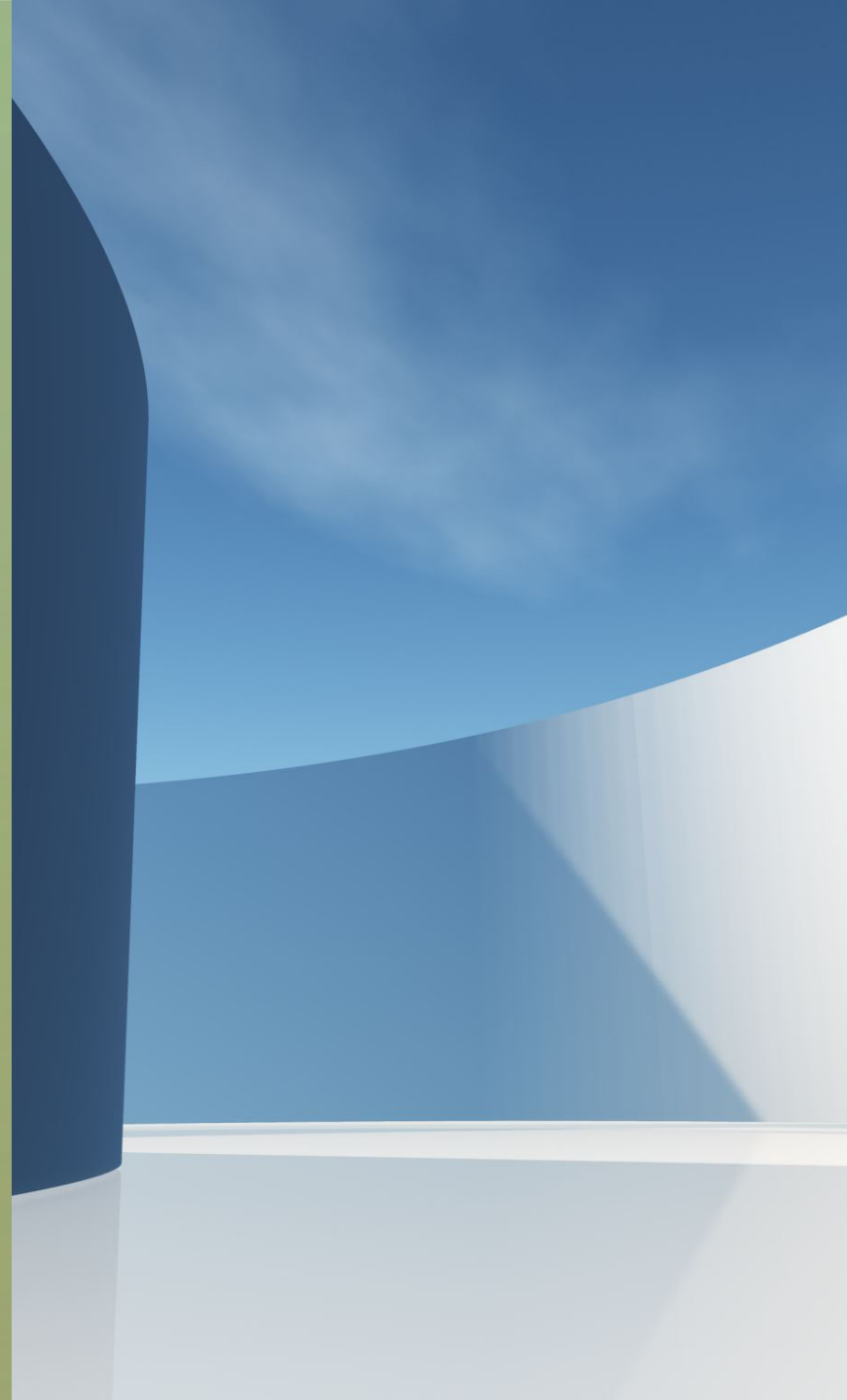
MACHINE LEARNING FINAL PROJECT

SPRING 2024

BIANCA GUNAWAN

ELENA GUALDA

MAYA MIRELES RIOS



RACIAL BIAS IN FACIAL ANALYTIC SYSTEMS

- Facial recognition algorithms have higher false positive rates for Asian and African American faces compared to Caucasian faces (The National Institute of Standards and Technology, 2019)

- Facial recognition systems from major tech companies exhibited higher error rates for darker-skinned women (up to 34.7%) compared to lighter-skinned men (0.8%) (Buolamwini and Gebru, 2018)

- Existing public face image datasets predominantly feature Caucasian faces, leading to underrepresentation of other races
- Efforts to diversify train data have been made to help mitigate bias

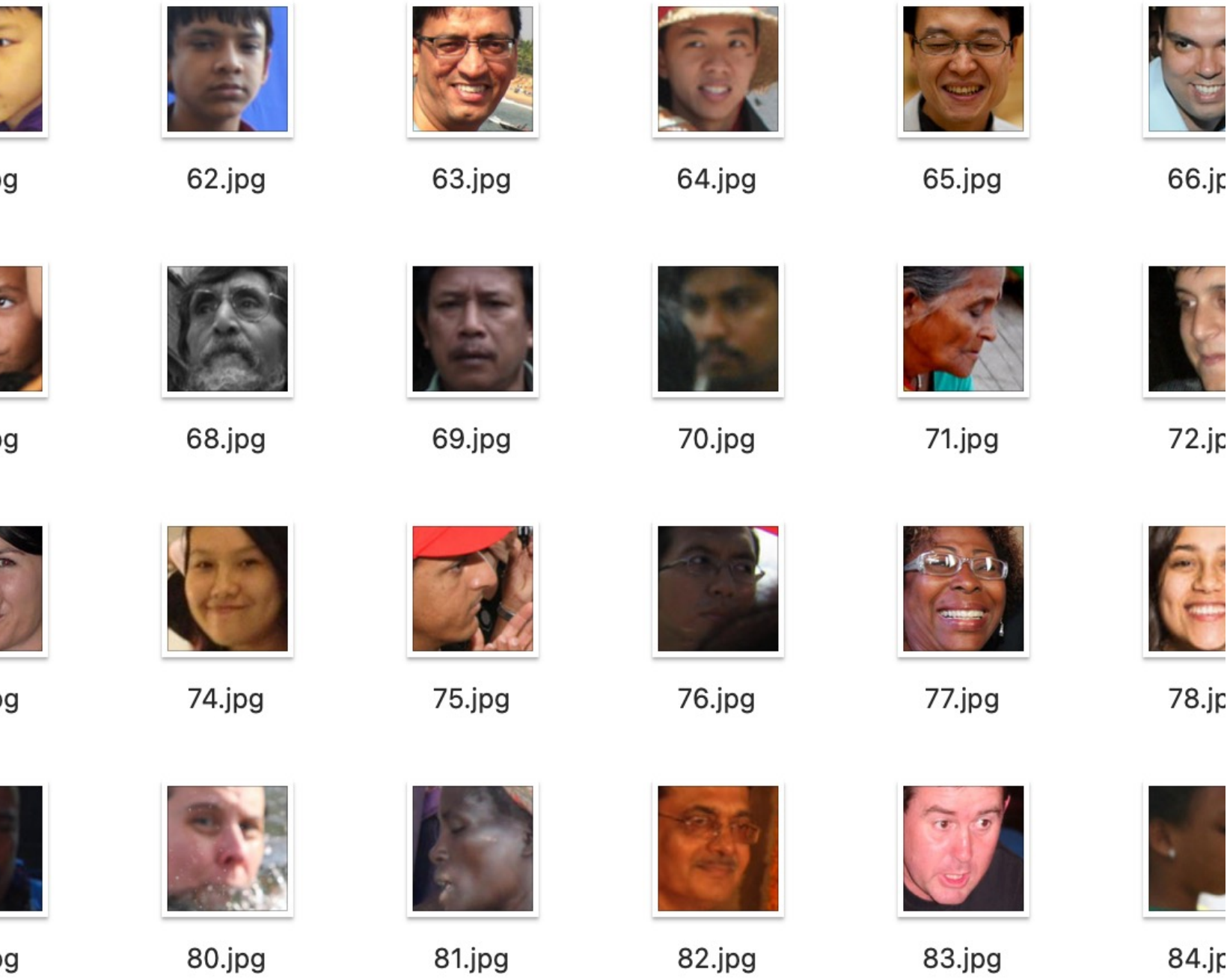
The Problem

Where does racial bias exist in deep learning software?

What role does the dataset play in the performance of these models?

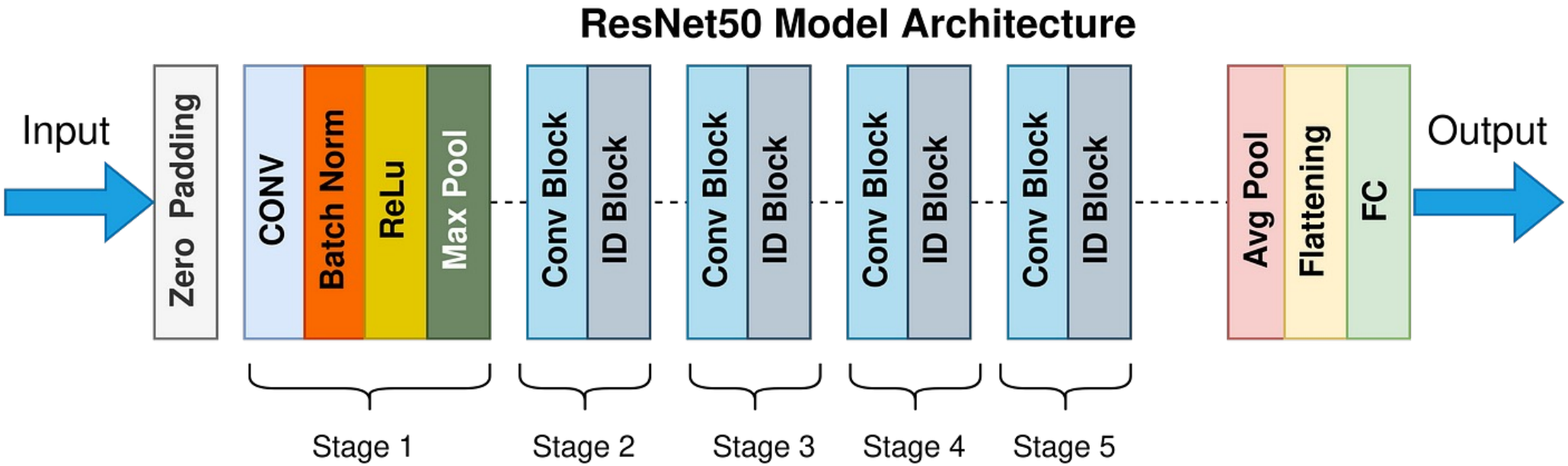
Goals of Analysis

- Explore the impact of more diverse data sets on model performance
- Compare the performance of different models given same data



DATA SET & DESCRIPTION

- Data set was sourced from GitHub.
- Racially diverse data set called FairFace
- 86,744 Facial JPEG Images labeled
  - 7 races
  - Age range 0-70+
  - Gender: Female & Male
- Due to computational power and efficiency, we took 3000 facial images for our Deep Learning analysis
- Resized Images to 224 x 224

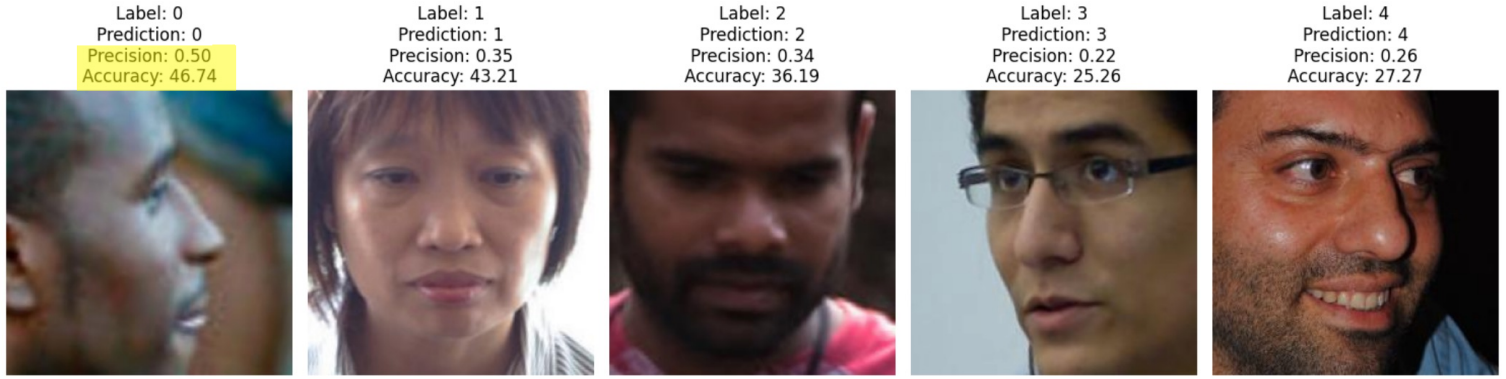


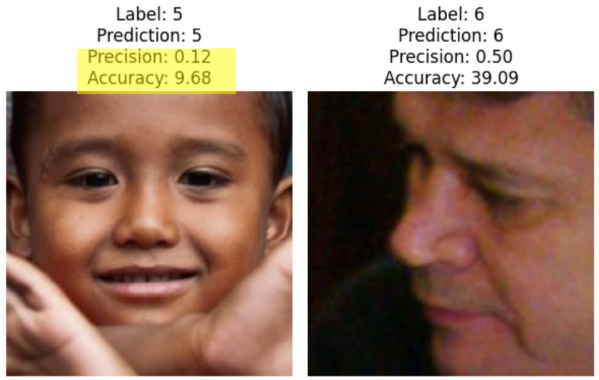
- 50-layer convolutional neural network
- Forms networks by stacking residual blocks, allow for the direct flow of information through the skip connections, mitigating the vanishing gradient problem
- Fewer filters and is less complex than a VGGNet

RESNET-50 SUMMARY

DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION

RESNET-50 RESULTS





- 3,000 images into 80/20 Train-Test Split
- Overall Test Accuracy: 34.0%
- Significantly low accuracy and precision for Southeast Asian, but quite balanced for other races.

INCEPTION RESNETV1– PRETRAINED WITH VGG MODEL - EXPLAINED

Inception: Multiple convolutional filters to process the image which captures different features for image.

- Each **filter** is a **different size** (e.g., 1x1 for fine details, 3x3 for medium patterns & max pooling, 5x5 for large patterns)
- All the outputs from **filters are concatenated** which combines features to a **single tensor**(multi-dim array).

ResNetV1: Creates shortcuts so model can jump layers to gather info from earlier layers and pass it on.

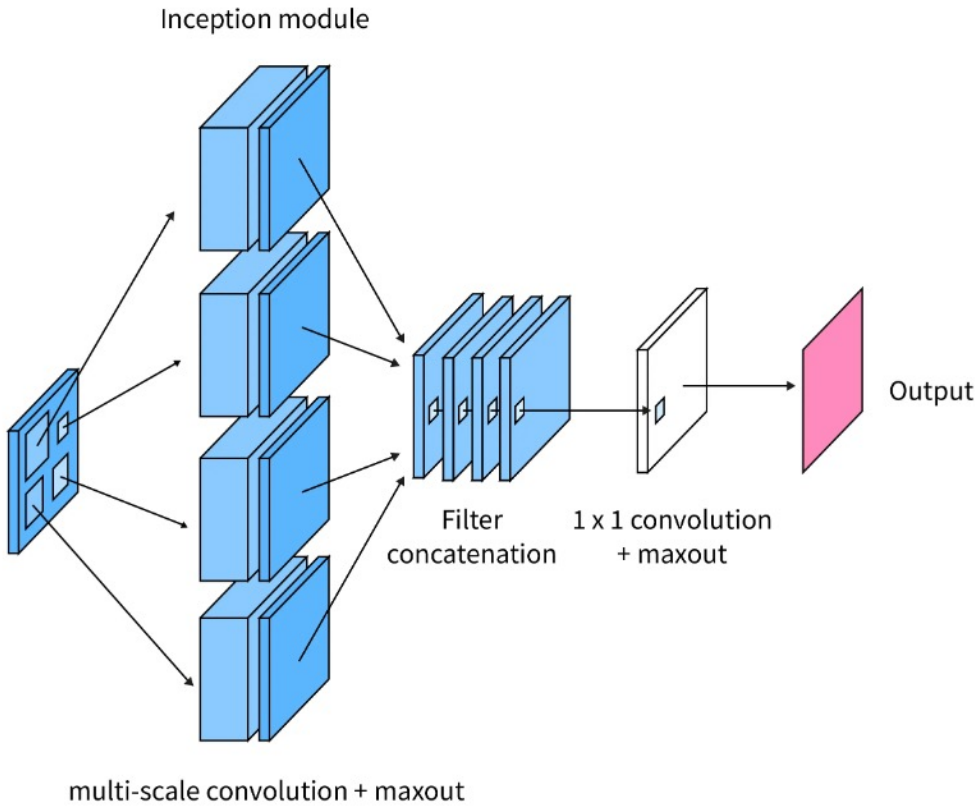
- Gradients flow backward easily during backpropagation which allows the **neural network to be much deeper.**

Pretrained – VGGFace2: Dataset from Oxford with 3.3 million images. Deep CNN Layers are used in VGG model architecture.

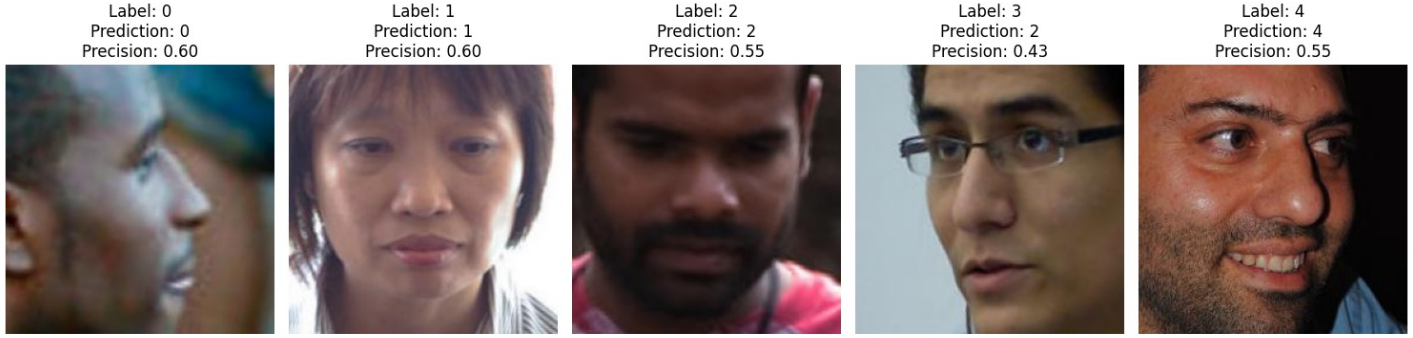
- InceptionResNetV1 is **trained** on the VGGFace2 dataset and is a prime example of **transfer learning.**

INCEPTION RESNETV1 – PRETRAINED WITH VGG MODEL - SUMMARY

- **Summary**: Each model feature has its own contribution to the neural network architecture for understanding images:
  - **Inception**: Feature Extraction
  - **ResNetV1** (Residual): Training of Deep Networks
  - **Pretrained** VGG: Weighs the model in transfer learning



INCEPTION RESNETV1– PRETRAINED WITH VGG MODEL – METHODOLOGY & RESULTS



- 3,000 images into 80/20 Train-Test Split
- Overall Test Accuracy: 52.50%
- Accuracy for:

  Label 0 (Black): 73.91%

  Label 1 (East Asian): 62.96%

  Label 2 (Indian): 45.71%

  Label 3 (Hispanic): 31.58%

  Label 4 (Middle Eastern): 25.45%

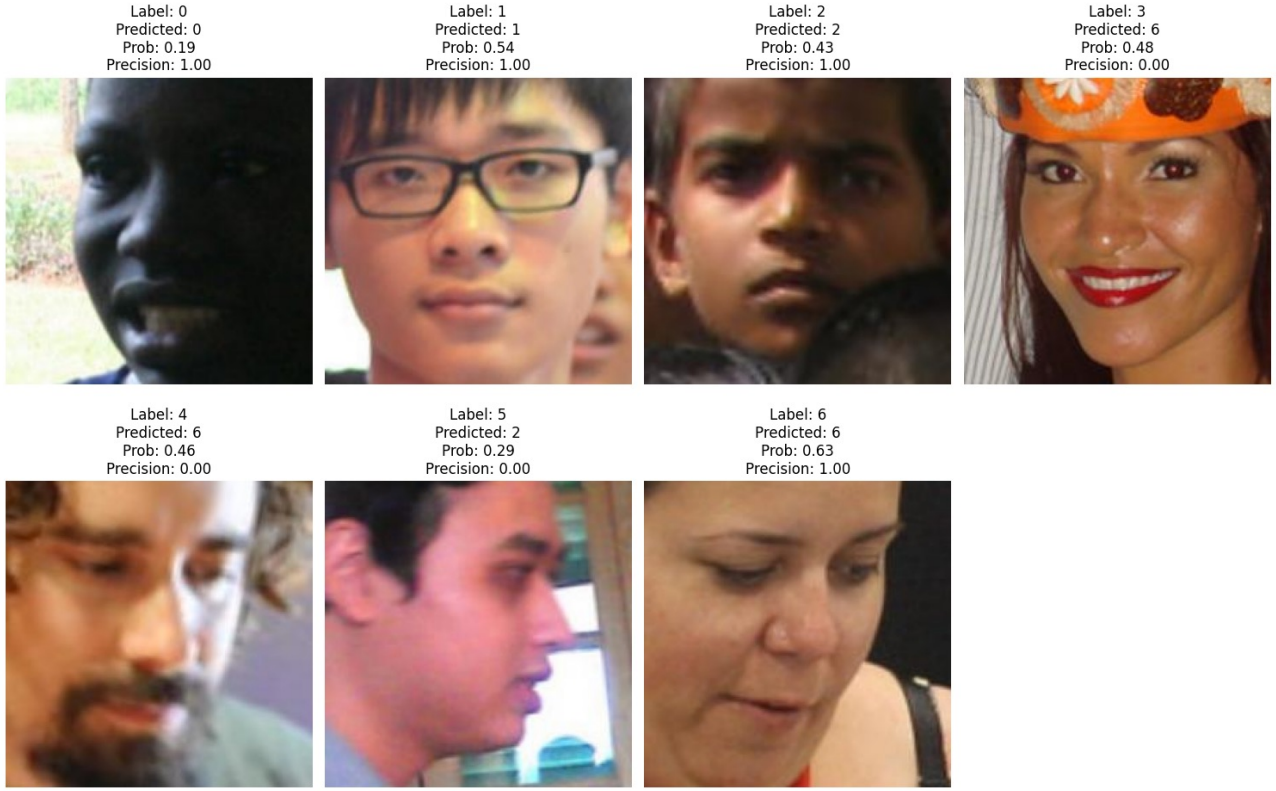
  Label 5 (SE Asian): 50.00%

  Label 6 (White): 66.36%

INCEPTION RESNETV1– PRETRAINED ON CASIA-WEBFACE

- Same **Hybrid Architecture:** Combines Inception modules and Residual Connections to leverage both depth and width
- Casia-Webface: large-scale face dataset that contains over 10,000 subjects and 494,414 images, making it one of the most popular datasets for training face recognition systems

INCEPTION RESNETV1(PRETRAINED WITH CASIA-WEBFACE) RESULTS



- 3,000 images into 80/20 Train-Test Split
- Overall Test Accuracy: 40.50%

Accuracy for:

Label 0 (Black): 64.13%

Label 1 (East Asian): 48.15%

Label 2 (Indian): 29.52%

Label 3 (Hispanic): 29.47%

Label 4 (Middle Eastern): 0%

Label 5 (SE Asian): 27.42%

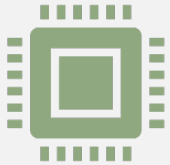
Label 6 (White): 62.73%

CONCLUSION

- RESNET50 is trained on object images, resulting low and varying facial recognition accuracy for different race class. This also suggests presence of racial bias
- INCEPTION RESNETV1(PRETRAINED WITH VGG2) is very robust and continually improves its accuracy the higher sample size its trained with. It demonstrates varying levels of accuracy based on race, suggesting that racial bias exists.
- INCEPTION RESNETV1(PRETRAINED WITH CASIA-WEBFACE) demonstrates varying levels of accuracy based on race, suggesting that racial bias exists

FUTURE DIRECTIONS

Accuracy improves the more data its fed. We could not increase train size due to RAM & GPU as of now



Fine-tune the pre-trained models with manually added layers

Explore other models that are more specific to facial recognition, such as FaceNet and DeepID

THANK YOU!

RESOURCES

Data

https://github.com/joojs/fairface

References

https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

https://www.scaler.com/topics/inception-network/