Project Summary:Exploring Bias in Model

Introduction:

This project aims to evaluate bias in deep learning models, specifically focusing on image classification using the CelebA dataset. The task involves classifying whether individuals in images are wearing eyeglasses, while also analyzing model performance across demographic subgroups such as gender and skin tone. The objective is to identify and quantify potential biases in the model's predictions to ensure fair and equitable performance for all demographic groups.

Model Architecture:

The model used in this project is a **Convolutional Neural Network (CNN)**, which is effective for image classification tasks. The architecture includes:

- **Input Layer:** Accepts RGB images resized to 128x128 pixels.
- **Convolutional Layers:** Three convolutional layers with increasing filter sizes (32, 64, 128), each followed by a max-pooling layer to reduce spatial dimensions and extract features.
- **Flatten Layer:** Converts the 3D feature maps to 1D feature vectors.
- **Dense Layers:**
    - One dense (fully connected) layer with 64 units and ReLU activation.
    - A final dense layer with a sigmoid activation function for binary classification (eyeglasses: yes/no).

Traning Procedure:

1.**Data Preparation:**
The CelebA dataset was filtered to include 1000 random samples, balanced across demographic groups. Image pixel values were normalized, and labels were prepared for binary classification.

2.**Data Splitting:**
The dataset was divided into **80% training** and **20% testing**.

3.**Model Training:**
The model was compiled using the **Adam optimizer** and **binary cross-entropy loss function**, suitable for binary classification. Training was conducted over **5 epochs** using mini-batches of size 32.

4.**Data Loading:**
TensorFlow's data pipeline was used with efficient prefetching, shuffling, and parallel processing to handle large image datasets smoothly.

Inference System:

The trained model was used to predict whether new images contained individuals wearing eyeglasses.
The following steps were involved in the inference system:

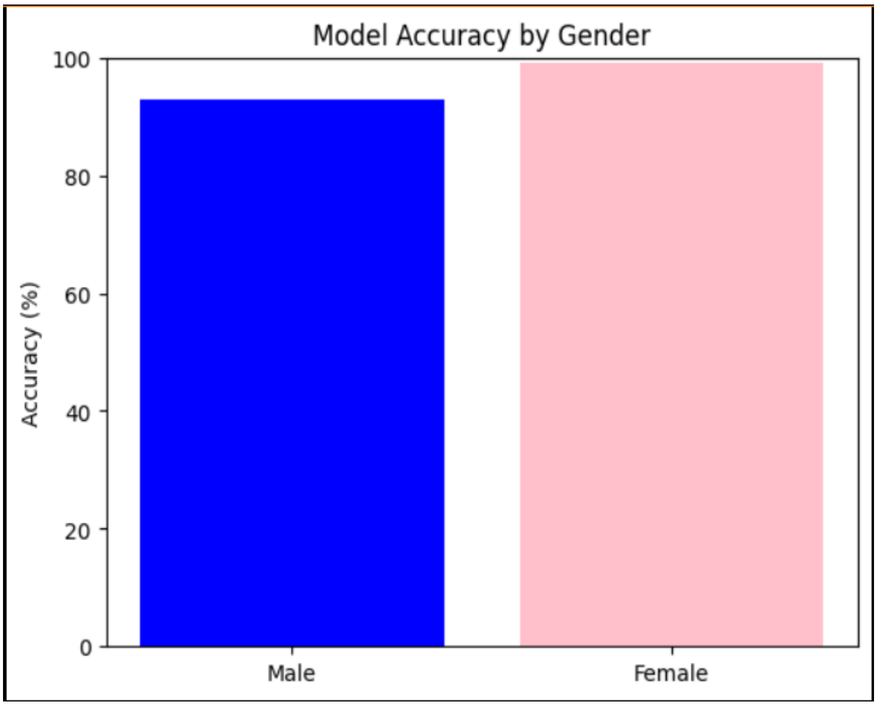
- **Image Preprocessing:** Input images were resized to 128x128 pixels and normalized.
- **Batch Prediction:** The model produced probabilities which were thresholded at 0.5 to classify images as either 'with eyeglasses' or 'without eyeglasses.'
- **Bias Analysis:** Model accuracy was separately calculated for different demographic groups (gender and skin tone) to assess fairness. Disparities in accuracies revealed potential biases,

particularly higher accuracy for female images compared to male images.

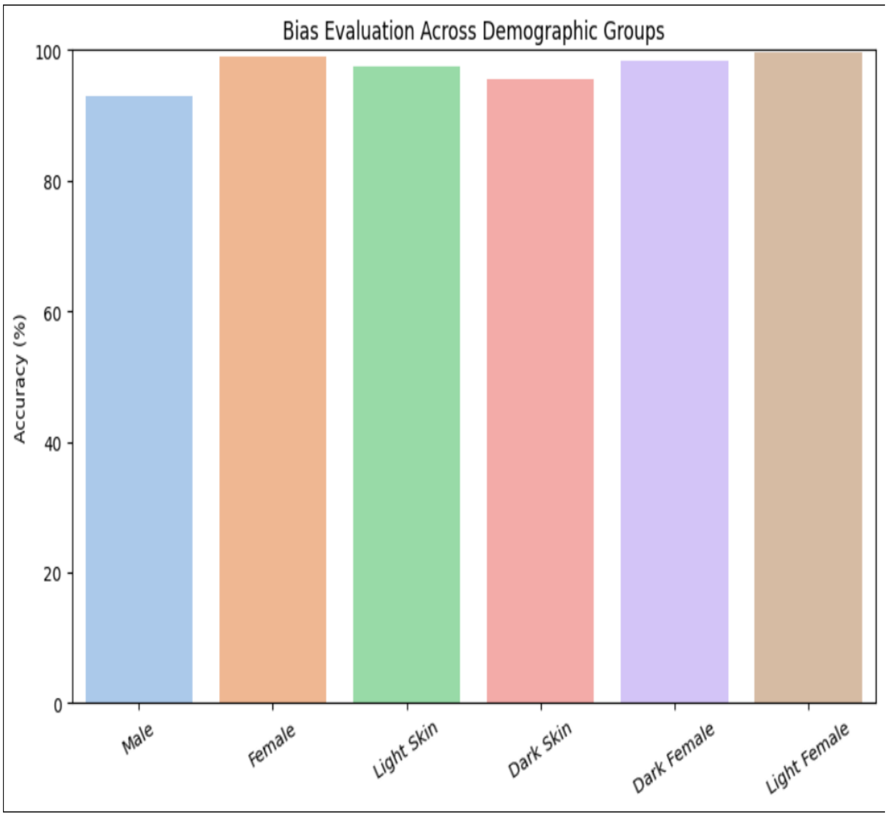
Conclusion:

The bias evaluation of our CNN model revealed performance differences across demographic groups. The model showed slightly lower accuracy for males compared to females, indicating a potential gender bias. Additionally, minor variations were observed across different skin tones, but these differences were not highly significant. Overall, the model performs well but demonstrates measurable demographic disparities that highlight the importance of fairness-aware model development and evaluation in deep learning systems.

Output:



We can conclude that the model performs better on female images than male images. There is a noticeable accuracy gap, with females having higher accuracy. This suggests the presence of gender bias in the model, where it is more effective in predicting eyeglasses for females than males. Addressing this bias is important to ensure fair and balanced model performance across genders.



The model shows higher accuracy for females compared to males, indicating a possible gender bias. Accuracy across light and dark skin tones is fairly balanced, suggesting minimal skin tone bias. The best performance is seen in female groups, but the consistent gap between genders highlights the need to improve fairness.