Helen Le

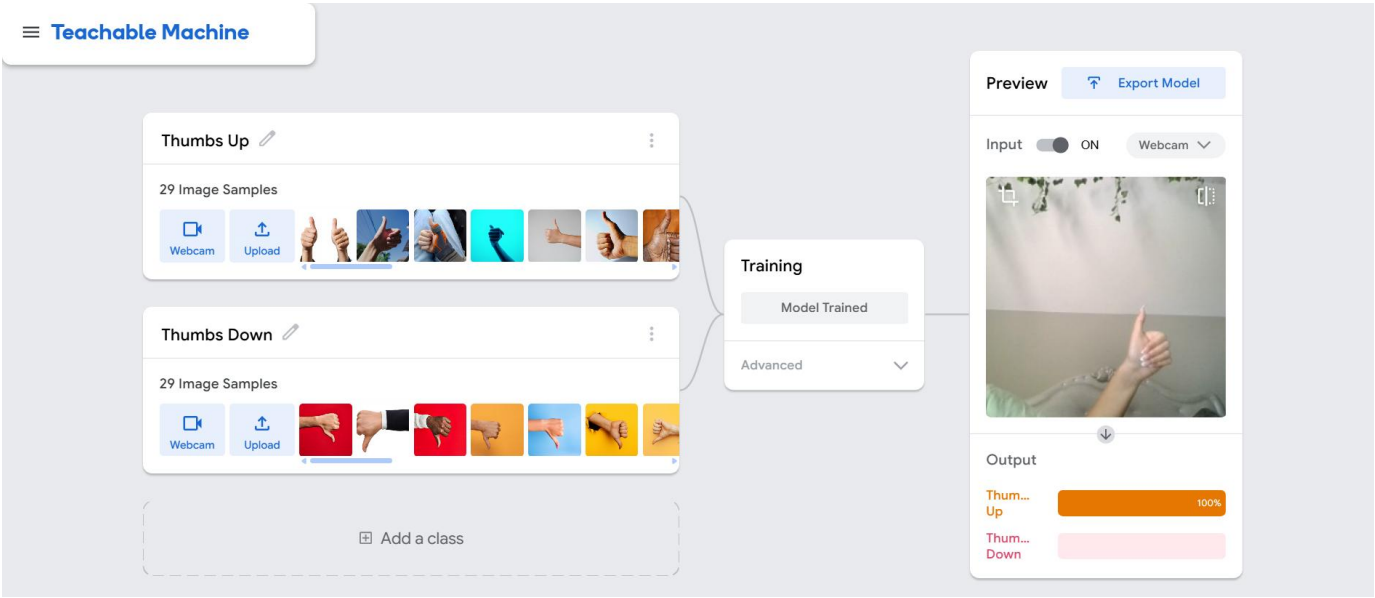
Professor Davidson

CIS 290-51 Emerging Topics in CIS: AI Ethics

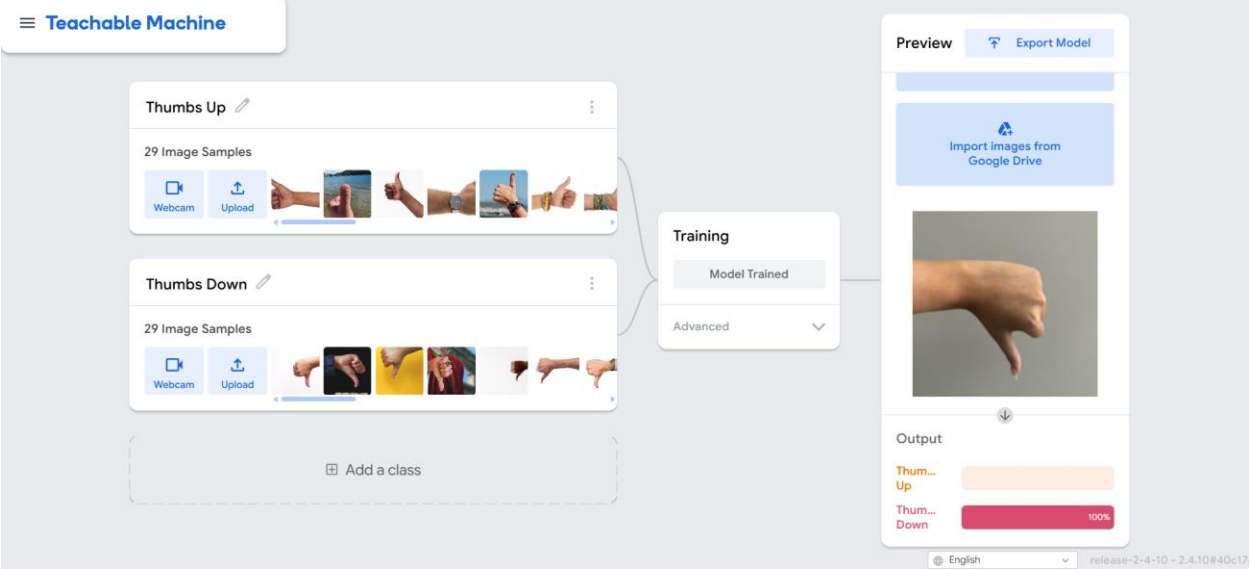
5 October 2025

Midterm

Balanced Dataset

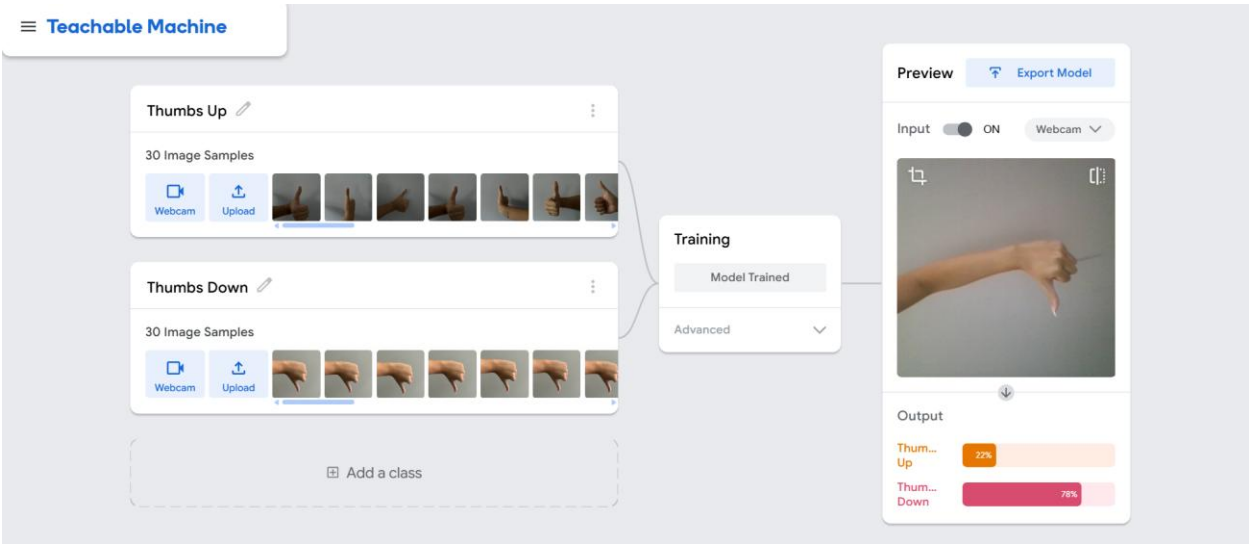


Correctly predicting a Thumbs Up outside of given data



Correctly predicting a Thumbs Down outside of given data

Biased Dataset



Correctly predicting thumbs down from same model and angle the machine was trained on

Balanced Dataset: images gathered from the web using different angles, lighting, accessories, genders, skin colors

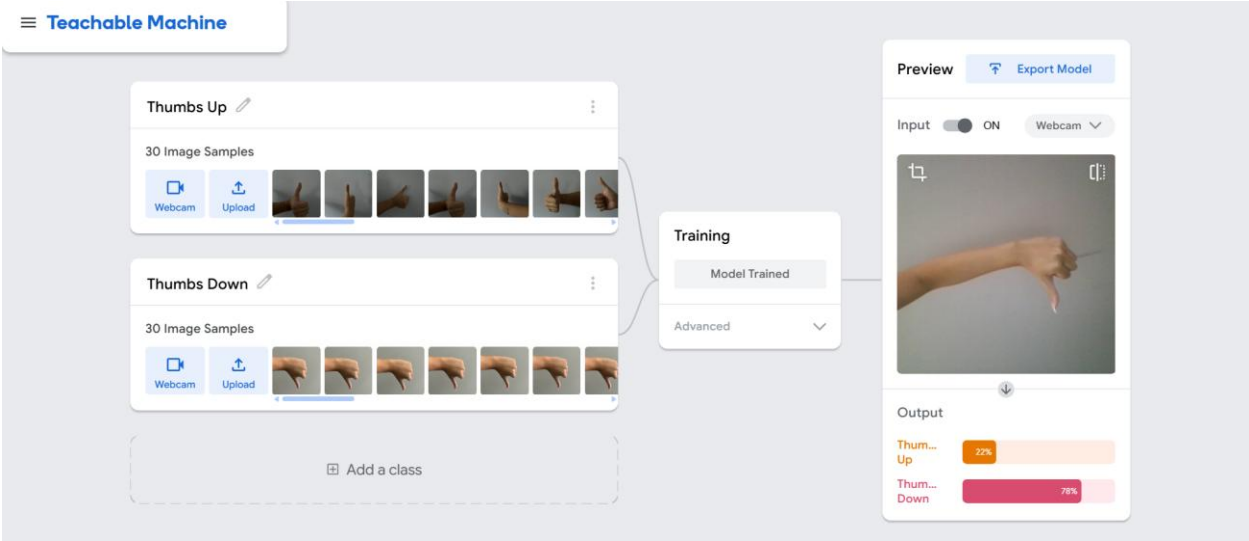
Biased Dataset: thumbs up was my friend's hand at different angles with different accessories and thumbs down was my hand at only one angle and with no accessories

Reflection

The balanced model had been able to correctly predict both thumbs up and thumbs down for images with different angles, models, lighting, and accessories. When I was rotating my thumb in the input, I did find that angles not represented in the dataset had proved to be less accurate, emphasizing the need for truly representative data that portrays the entire picture and not just part.

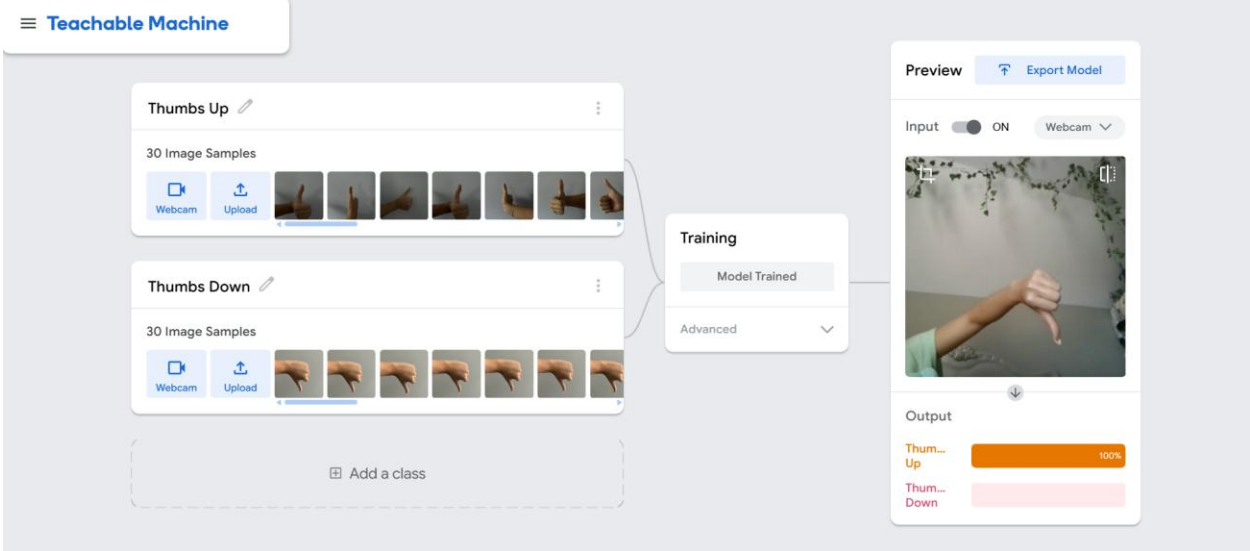
On the other hand, the biased model consistently made inaccurate conclusions.

EXAMPLES TESTING THUMBS DOWN



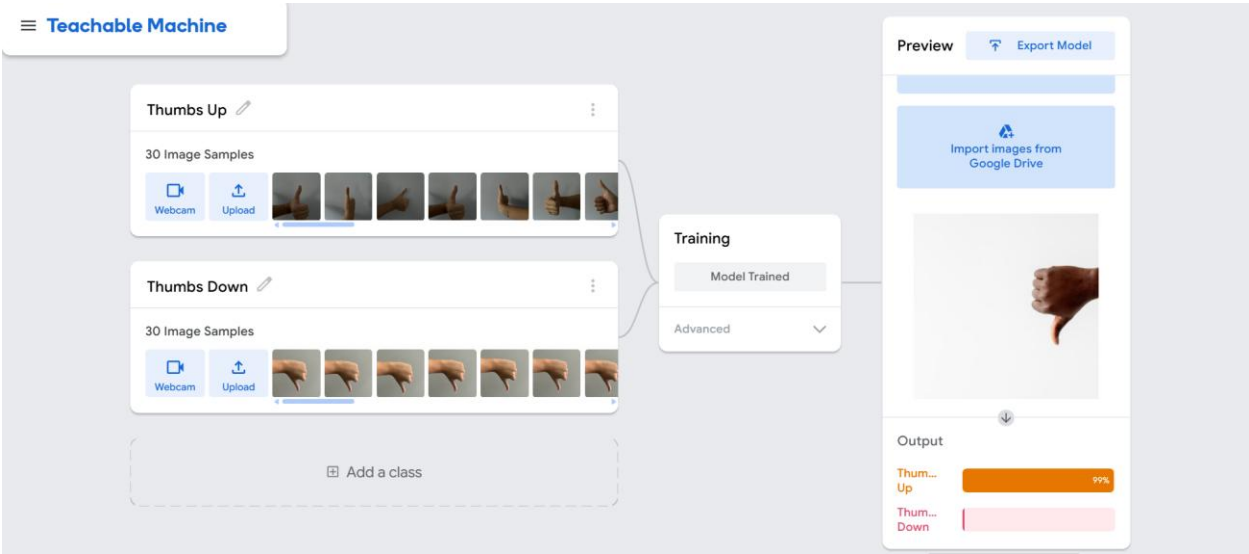
Correctly predicting thumbs down from same model and angle the machine was trained on

As shown in the image above, it had been mostly correct about my own thumb being a thumbs down when I used the same background at a similar thumbs down position.



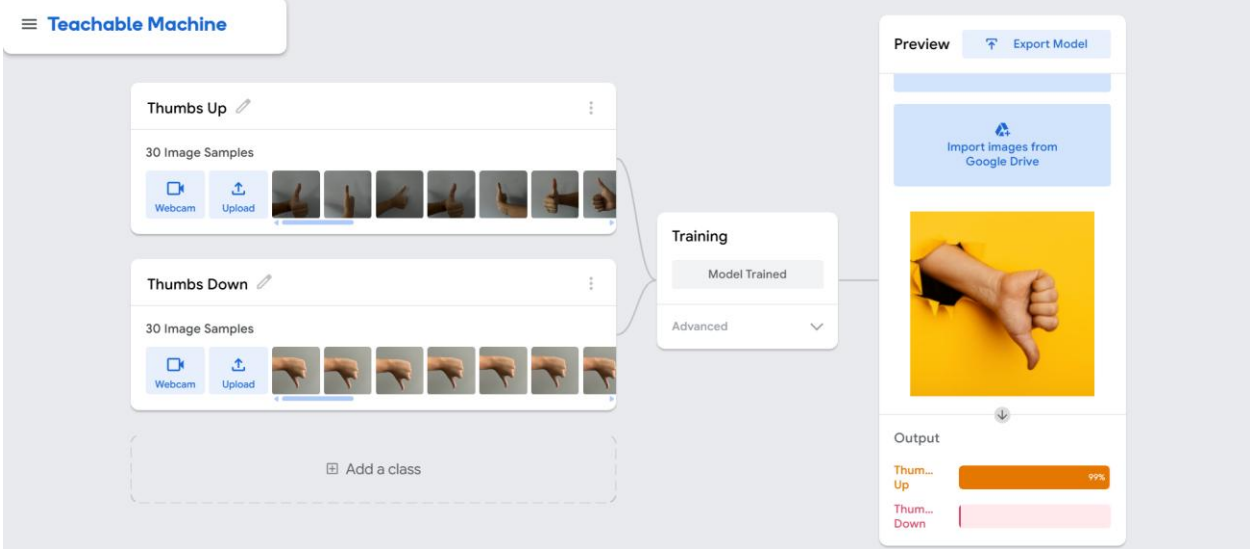
Incorrectly predicting thumbs up for my thumbs down

However, when I used my thumb again but at a different angle with a different background, it had been 100% confident that it was a thumbs up.



Incorrectly predicting thumbs down using different skin color

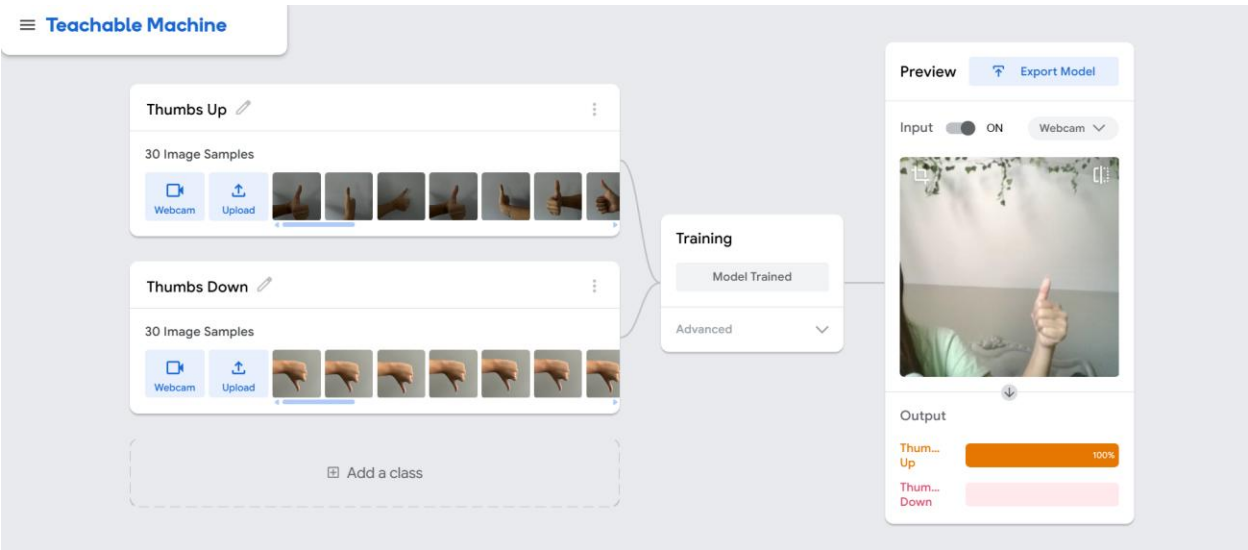
When using unseen models, it had incorrectly predicted a thumbs down from a person with a different skin color to be a thumbs up.



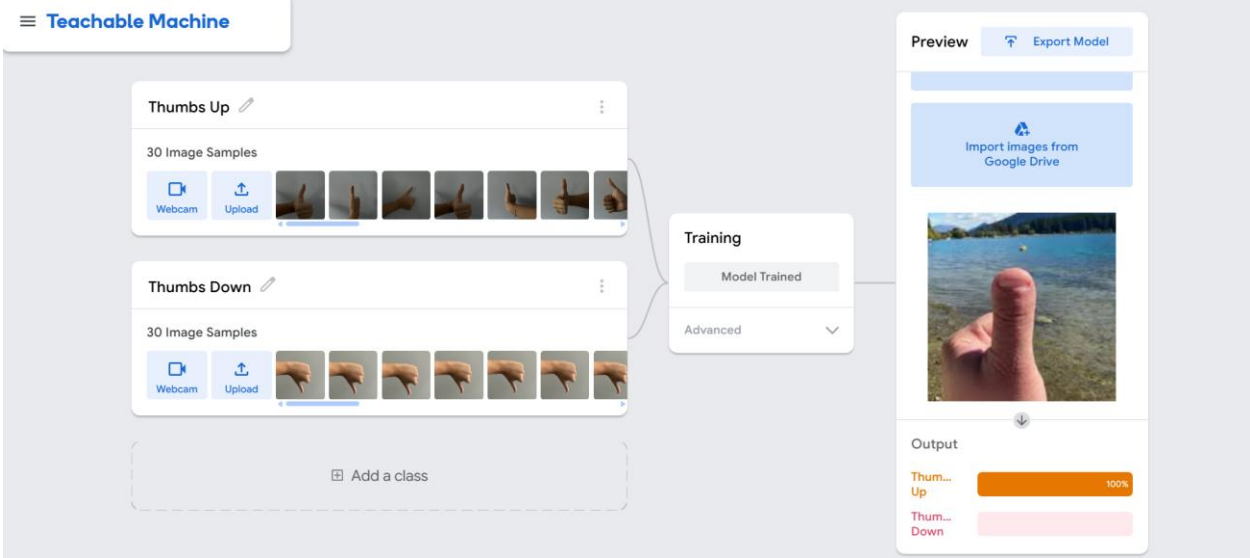
Incorrectly predicting thumbs down using similar skin color but different model from original dataset

When using an unseen model with a similar skin color, it had still incorrectly predicted a thumbs down to be a thumbs up. It is clear the biased model consistently misclassified thumbs downs as a thumbs up. It had only been remotely correct when I mimicked the same angle as the training data used.

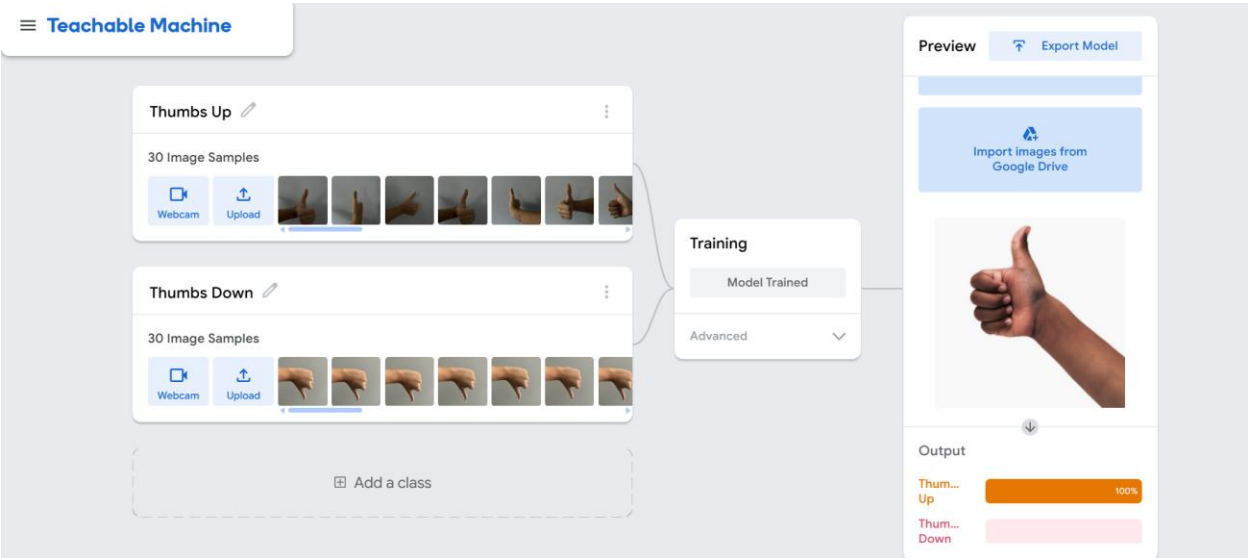
EXAMPLES TESTING THUMBS UP



Correctly predicting new model (my own thumb) to be a thumbs up



Correctly predicting new model with similar skin tone to be a thumbs up



Correctly predicting model with a darker skin color to be a thumbs up

When testing the accuracy of thumbs up with different models, it had been extremely accurate. This can be explained by the diverse dataset used to train the Thumbs Up class. As I used different angles with different accessories, the model had been forced to not just learn the conditions but learn the thumbs up gesture entirely to be able to make accurate predictions.

The bias in the model comes from the way training data was collected. For the thumbs down class, I had only used images of **my hand** at **one specific angle**, while for the thumbs up class, I used pictures of my **friend's hand** at **many different angles** and **wearing varying accessories**.

This imbalance caused the model to simply learn and memorize what thumbs down looked like in that one specific situation, rather than truly understanding thumbs down from all angles, skin colors, and other conditions. In short, instead of learning the gesture, it learned the conditions. The biased model had been overfit to one specific version of a thumbs down. Compared to the balanced model, it had a worse performance.

To make AI models less biased, always ensure that more diverse training data is used, meaning include different people, angles, locations, lighting, and more. I also found it key to note that when trying to represent different races, most images out there were of Caucasian descent, leaving other races to be unrepresented. This is where data augmentation or data preprocessing would come into place. In doing this, also ensure this diversity goes for **all classes used** and not just one, like this example had done. Since only one was diverse, the other one had simply learned my thumb at that one location and one angle, rather than understand that it is a thumbs down. Once you believe your dataset is fully representative of all people, locations, and conditions, **test before deploying.** Doing test examples as done here prevents a system like this from going out in the real world and causing harm. If the system clearly makes repetitive mistakes such as these, it should not be implemented until the issue is resolved.

If these systems were to be used in the real world, it would create false expectations in how inaccurate it is. For instance, if analyzing customer emotions concerning a product, the machine would misread emotions from one group as it is outside of what it is trained from. The MIT article linked had explored how darker-skinned women had significantly higher error rates than light-skinned men. From these results, it was clear that the database the model was trained on consisted primarily of light-skinned men. In fact, the database was “more than 77 percent male and more than 83 percent white.” This caused it to be more accurate for that group and less accurate for those outside of it. In terms of security access, a user that *should* be authorized to a system might not be properly recognized from their hand gesture because the dataset was not fully representative of all races, lightings, locations, and angles.

In conclusion, this experiment emphasizes the importance of obtaining a well-rounded dataset and demonstrates the impact on the AI model. While my balanced model was able to predict a variety of new models, my biased model had failed to predict new models and was extremely confident, despite being inaccurate. The bias had come from the Thumbs Down dataset only

being one model at one specific angle and essentially replicated 30 times, while the Thumbs Up dataset was more diverse, consisting of varying angles and accessories. If bias were ignored for real world use, it would replicate the biases learned and cause extreme harm to misrepresented races.

Sources

https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Balanced: https://teachablemachine.withgoogle.com/models/PE_blkwky/

Biased: https://teachablemachine.withgoogle.com/models/qx68BVw0F/