CS 674 Final Project

Rebecca Jones, Ethan Walker

April 21, 2021

1 Overview

Facial recognition algorithms are well known to be biased against minority groups[6][7]. Not only are minority groups misclassified more often, they are also shown to be recognized in databases in which they are not present more often than majorities. We build on previous work that has been done on investigating bias, [9][8] and look at how imbalanced datasets affect the accuracy and misclassification of different demographics.

2 Methods

2.1 Architecture

The different implementations of facial recognition algorithms were primarily focused on the use of different loss functions. So as a base we used a 50 ResNet equipped with the following different loss functions. We implemented each of these loss functions as described in the associated papers.

2.2 Ring Loss

Ring Loss is a feature normalization method[1]. We used Ring Loss in conjunction with Cross Entropy Loss in training, and found rather good results with this method. The main intuition of this method is that it encourages the model to place the final encoded features on a ring of a learned radius. The paper argues that this reduces test error as it discourages low norm features, and it encourages a better angular differentiation between the features.

The computation of the Ring Loss is extremely simple and is as follows

$$L_R = \frac{\lambda}{2m} \sum_{i=1}^{m} (\|\mathcal{F}(x_i)\|_2 - R)^2,$$

where $\mathcal{F}$ is the neural network, R is the learned radius parameter, and λ is a hyperparameter giving weight to the Ring Loss.

2.3 Angular Softmax

Angular Softmax has a similar intuition in that it seeks to create representations that are very different with respect to some angular measure. Specifically, A-Softmax can be interpreted as requiring the model to use the angular distance metric on a hypersphere to learn[2]. This is done in the following way

$$L_{ang} = \frac{1}{N} \sum_i -\log \left(\frac{\exp(\|x_i\| \cos(m\theta_{y_i,i})}{\exp(\|x_i\| \cos(m\theta_{y_i,i}) + \sum_{j \neq i} \exp(\|x_i\| \cos(\theta_{j,i})} \right),$$

where m is a hyperparameter. The θ values are computed by an inner product of x with a learned weight matrix W, which is why they are specified with two indices.

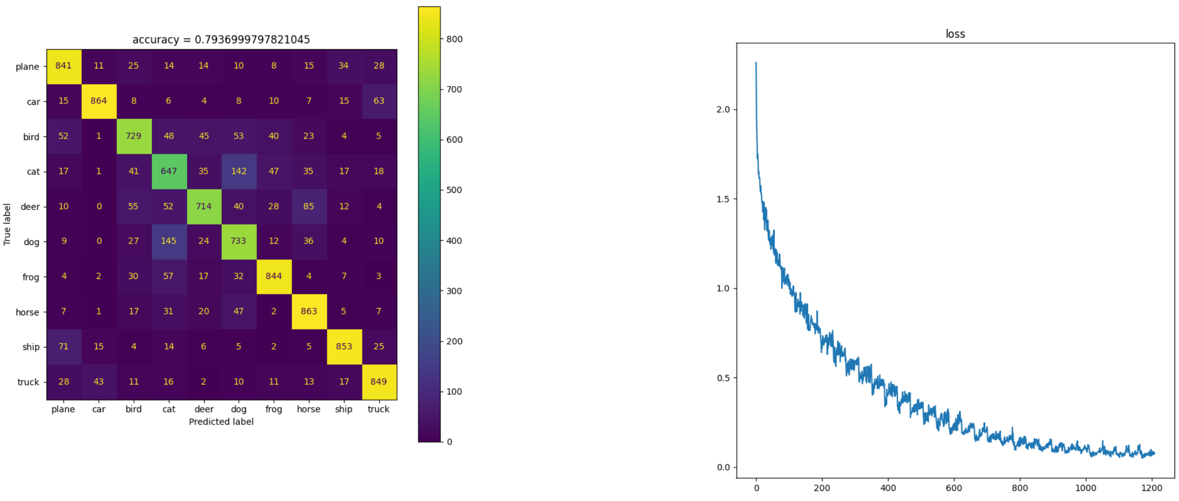
This loss function was difficult to work with, and required a lot of trial and error to understand and implement correctly.

2.4 Model Testing

To get a baseline understanding of how these different models would perform, as well to test them to ensure they work correctly, we trained each of these models on CIFAR10 and got the following results.

2.4.1 Ring Loss

In order to get a baseline understanding of how well Ring Loss can perform, we tested initially with CIFAR10. As you can see in Figure 2, we were able to get good performance with only a 50-layer ResNet and 50 epochs of training.



	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	841	11	25	14	14	10	8	15	34	28
car	15	864	8	6	4	8	10	7	15	63
bird	52	1	729	48	45	53	40	23	4	5
cat	17	1	41	647	35	142	47	35	17	18
deer	10	0	55	52	714	40	28	85	12	4
dog	9	0	27	145	24	733	12	36	4	10
frog	4	2	30	57	17	32	844	4	7	3
horse	7	1	17	31	20	47	2	863	5	7
ship	71	15	4	14	6	5	2	5	853	25
truck	28	43	11	16	2	10	11	13	17	849

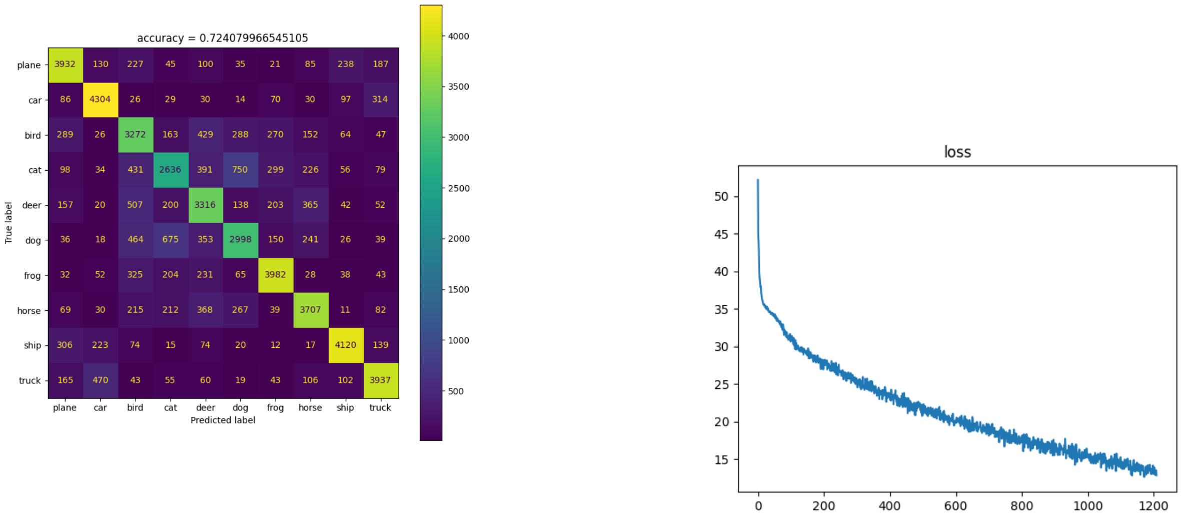
(a) Confusion matrix shows a 79.4% accuracy on CIFAR10

(b) Loss curve over 50 epochs

Figure 1: Ring Loss results on CIFAR10

2.4.2 Angular Softmax

We were able to get similar, though slightly worse results with the A-Softmax model. Because this model was more computationally complex, and therefore considerably slower to train, we could not train for the same number of epochs.



(a) Confusion matrix shows a 72.4% accuracy on CIFAR10

(b) Loss curve over 30 epochs

Figure 2: Ring Loss results on CIFAR10

3 Data

Testing the facial recognition model while also being able to observe the influence demographic attributes requires careful organization and representation of the data. We observe that while the datasets are unbalanced with respect to the attributes, it is balanced by class (people). To our knowledge, this type of unbalanced attributes has not been investigated in this way.

3.1 Dataset Creation

To create the datasets, we used the UTKFace dataset[4]. It consists of 23708 different images containing a single face. Importantly, each image is labeled by age, race, and gender. To create each dataset, we randomly selected images from each attribute to get the desired distributions. For gender, the datasets were split 50/50, 40/60, 30/70, 20/80 and 10/90 between the genders for a total of 10 datasets. Similar splits were done for the race.

Once we had the images, we augmented the data using albumentations[5] and a random assortment of resizing and cropping the images to 100x100 pixels, horizontal flipping, blurring, downscaling, coarse dropout, color adjustments, and noise. Each image was augmented 16 times to create the full dataset. The original resized image and another augmented copy were used in the test set.

Additionally, we randomly took 25 images that did not appear in the dataset to see how the model classified images not in the model.

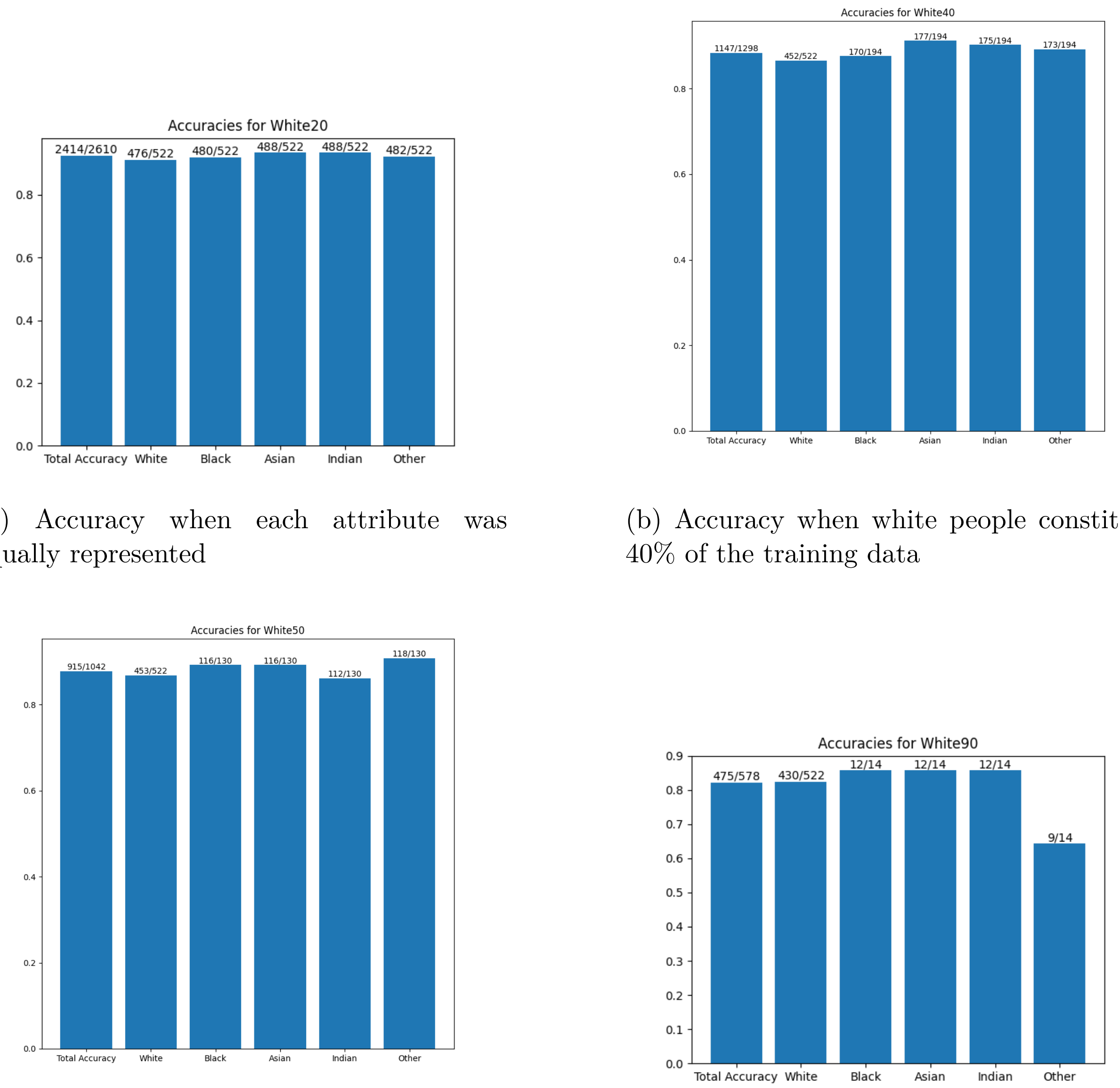
4 Results

Here we will examine each model's performance with respect to the different demographic attributes we considered. While many of the charts in these sections will be labeled with demographic attributes, the model was not classifying according to that attribute. Instead, these charts represent the model's performance on individuals among those demographic groups.

4.1 Race

4.1.1 Ring Loss

First we will examine the accuracy of the model. In order to examine the resiliency of Ring Loss to imbalanced data, we trained it with different data sets where one demographic becomes increasingly over represented. In the trials recorded in Figure 3, we see the accuracy results when the data set was 20%, 40%, 50%, and 90% white individuals. Ring Loss is highly resilient to data imbalance, allowing it to maintain high accuracy, even among demographic groups that deeply under represented in its training data.



(a) Accuracy when each attribute was equally represented

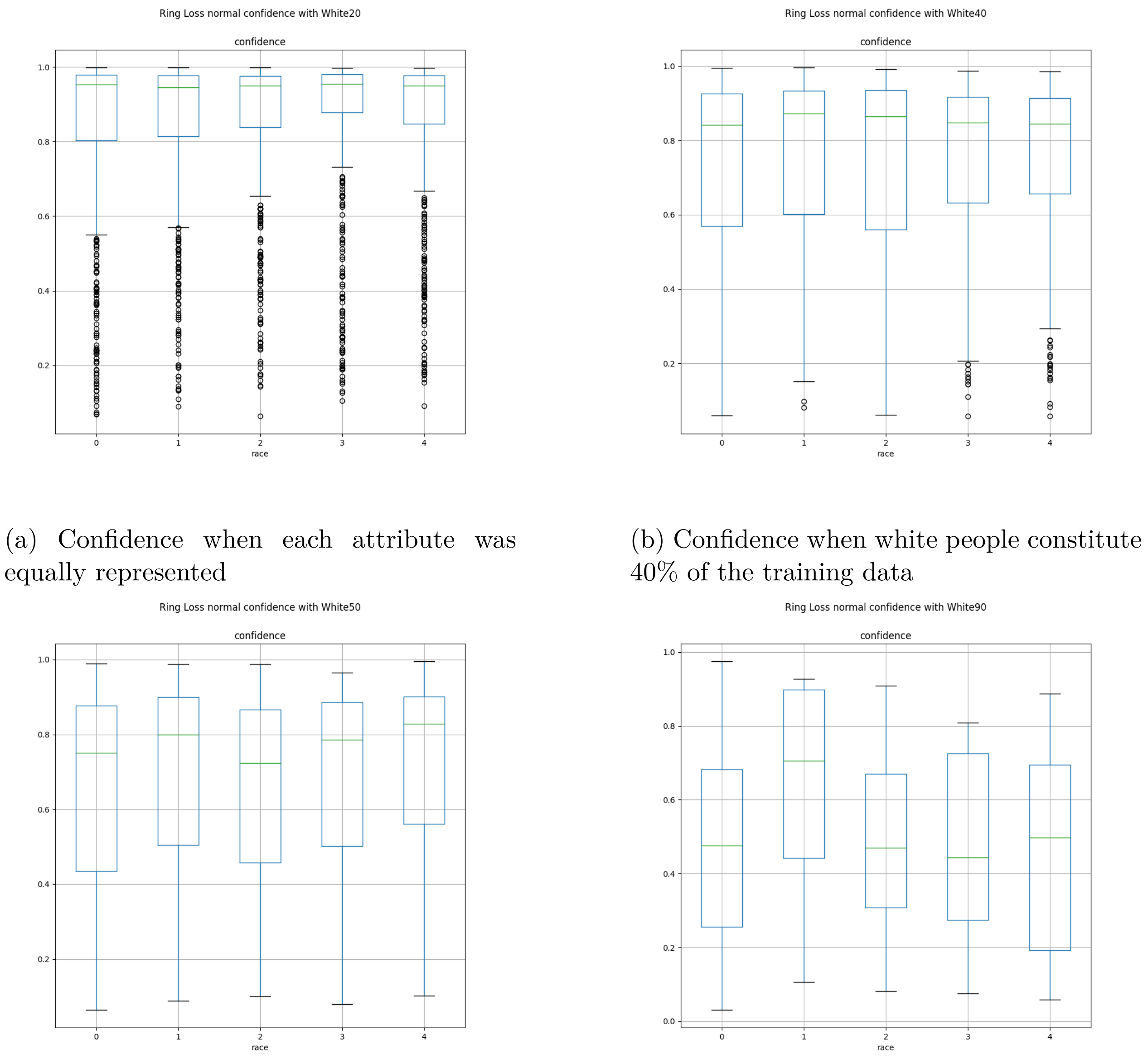
(b) Accuracy when white people constitute 40% of the training data

(c) Accuracy when white people constitute 50% of the training data

(d) Accuracy when white people constitute 90% of the training data

Figure 3: Ring Loss results on Face Data

Next we examined how the model treated individuals it had not seen before. In this case, we showed the model images of individuals that it had not yet seen and examined the confidence the model had in its labelling of those images. But before we examine Figure 5 we need to see the confidence of the model with faces that it has already seen. We can find this in Figure 4 the model is generally very confident of faces that it has seen before, however when the training data is unbalanced the overall confidence of the model decreases.



(a) Confidence when each attribute was equally represented

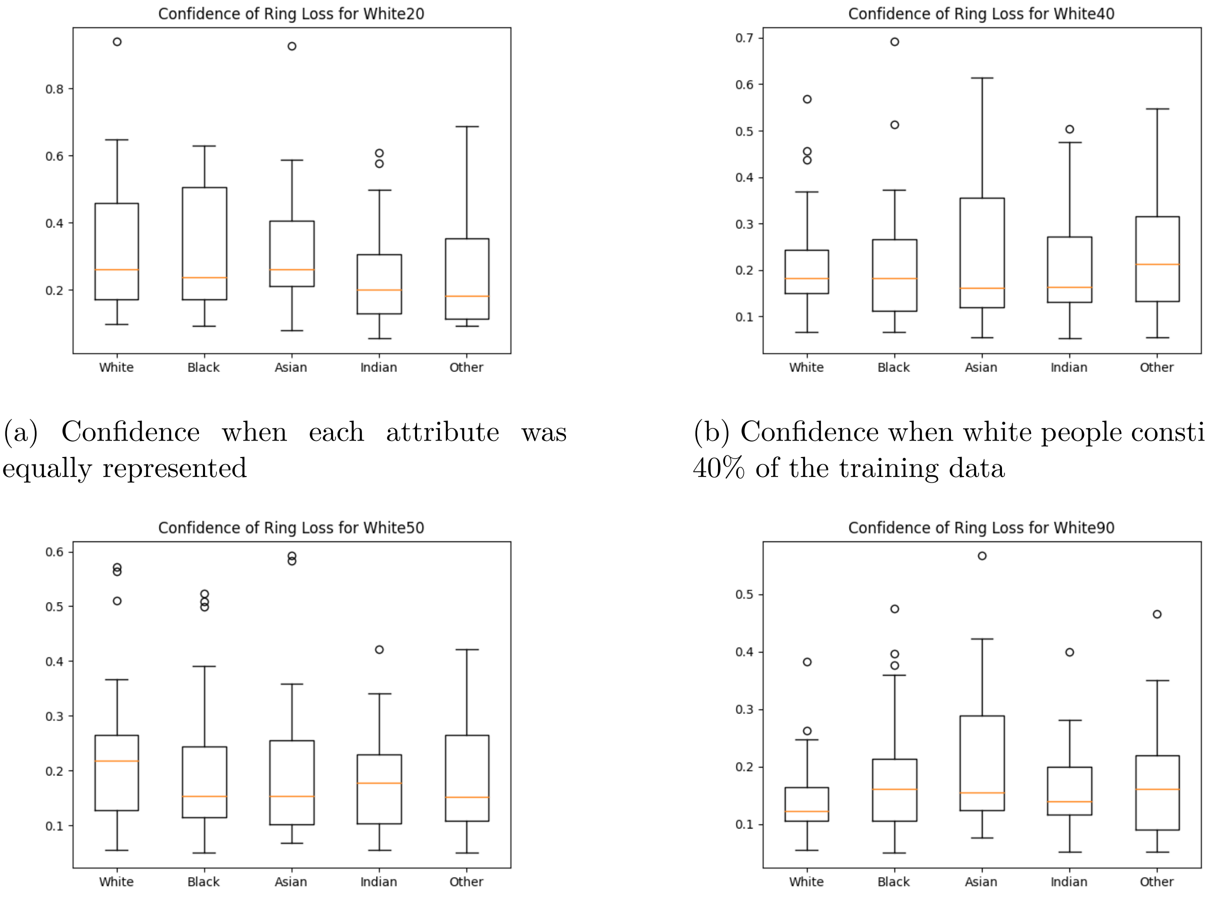
(b) Confidence when white people constitute 40% of the training data

(c) Confidence when white people constitute 50% of the training data

(d) Confidence when white people constitute 90% of the training data

Figure 4: Ring Loss confidence on seen face data

Comparing these two figures we can see that the model is significantly less confident when it has not already seen the faces it is asked to classify, but it does seem like there could be a slight over confidence when it comes to demographics that are not in the majority of its training data. This overconfidence is most visible in Figure 5d, and it can be seem in Figure 5b as well. Despite this indication, it is not clear that the trend actually exists.



(a) Confidence when each attribute was equally represented

(b) Confidence when white people constitute 40% of the training data

(c) Confidence when white people constitute 50% of the training data

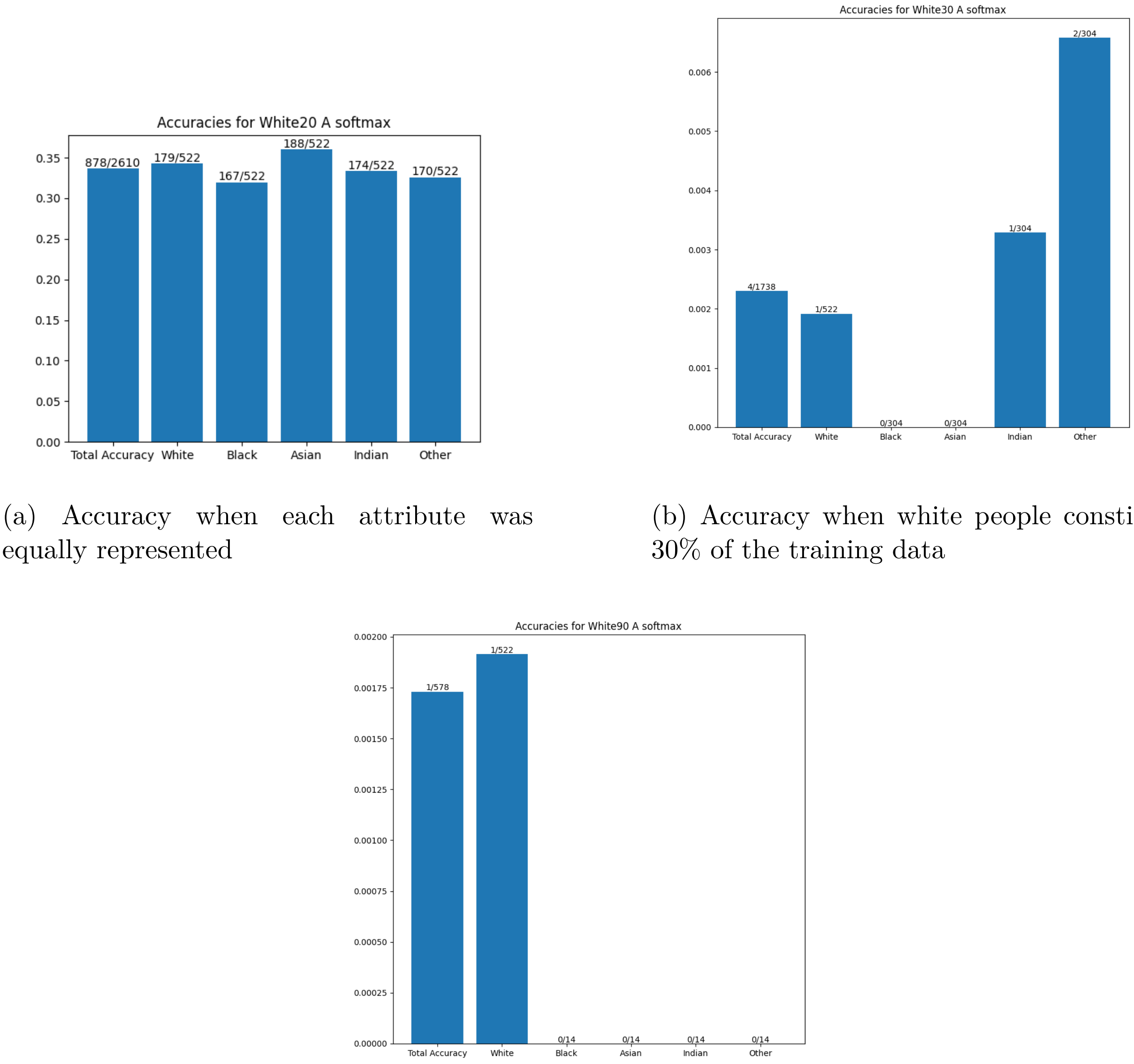
(d) Confidence when white people constitute 90% of the training data

Figure 5: Ring Loss confidence on unseen face data

4.1.2 A-Softmax

Angular Softmax performed very poorly in general. In Figure 6 we see that even in the balanced model the accuracy was never better than 35% overall. However we can see that on some occasions, when a group is over represented, the model does especially poorly on the under represented group, however that result is not consistent enough to make any conclusions.

These poor results are likely caused by a lack of hyper-parameter tuning. A-Softmax was slow to train and because of that, we did not spend as much time tuning the hyper-parameters. Because of these poor results we did not do a confidence analysis on the A-Softmax model.



(a) Accuracy when each attribute was equally represented

(b) Accuracy when white people constitute 30% of the training data

(c) Accuracy when white people constitute 90% of the training data

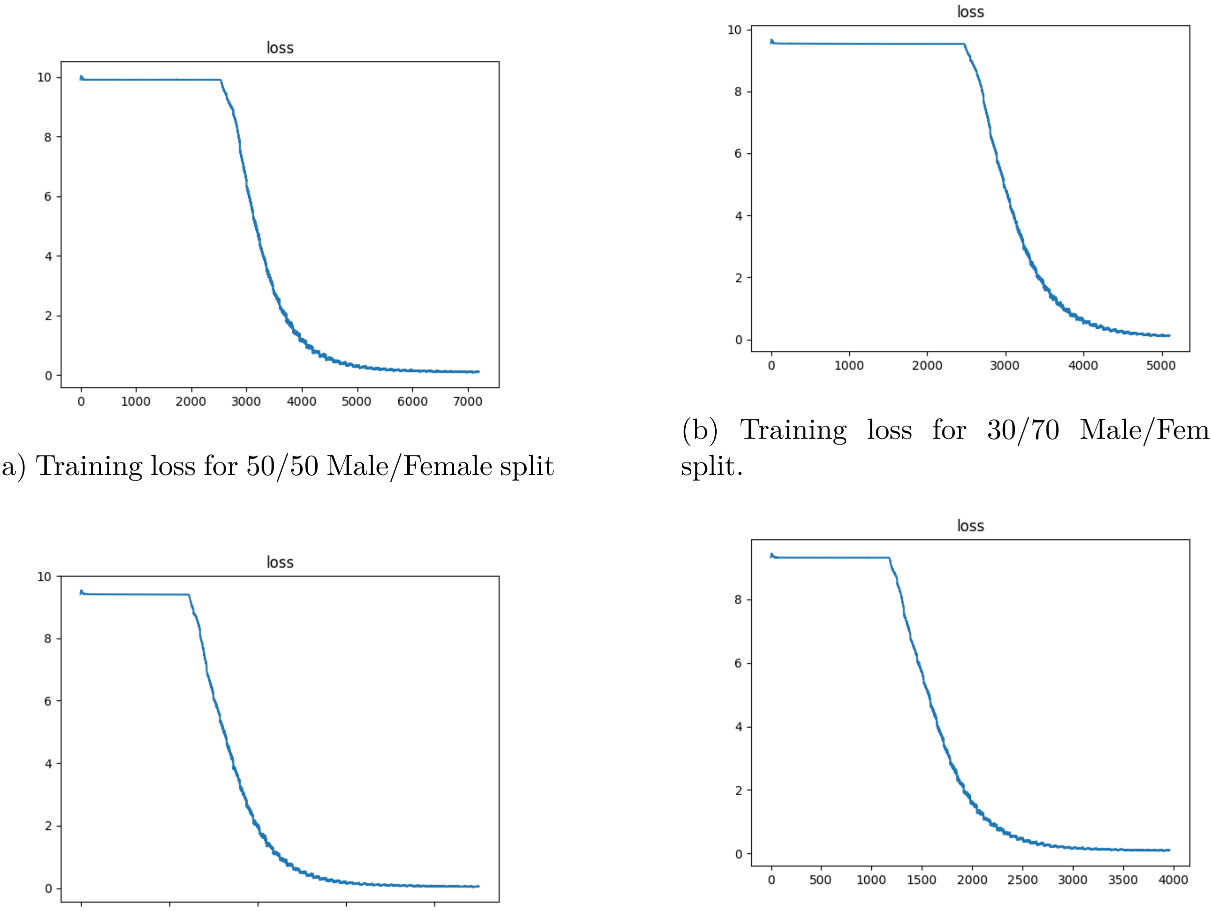
Figure 6: A-Softmax results on Face Data

4.2 Gender

Similar to the Race analysis, the Gender analysis looks at the accuracy of the model compared to the accuracy of classifying demographics. Overall, each model attained about a 90% accuracy using the Ring Loss. See Figure 8. The accuracies for the demographics were within 1% of the total accuracy, and unexpectedly, the demographic that was least represented performed better overall than the demographic that was respresented more.

A few examples of the training loss are shown in Figure 7. One interesting observation is that the training loss converged sooner as the demographics became more unbalanced.

To simulate identifying people in a database, we calculated the softmax of the model outputs on the datasets of people who weren't in the training or test sets. Any person who had a maximum softmax of greater than 90% was considered identified. Results can be seen in Figure 9. The results were not conclusive. We couldn't find out how this part of facial recognition worked or what a good threshold for identification was, so that may have played a part.



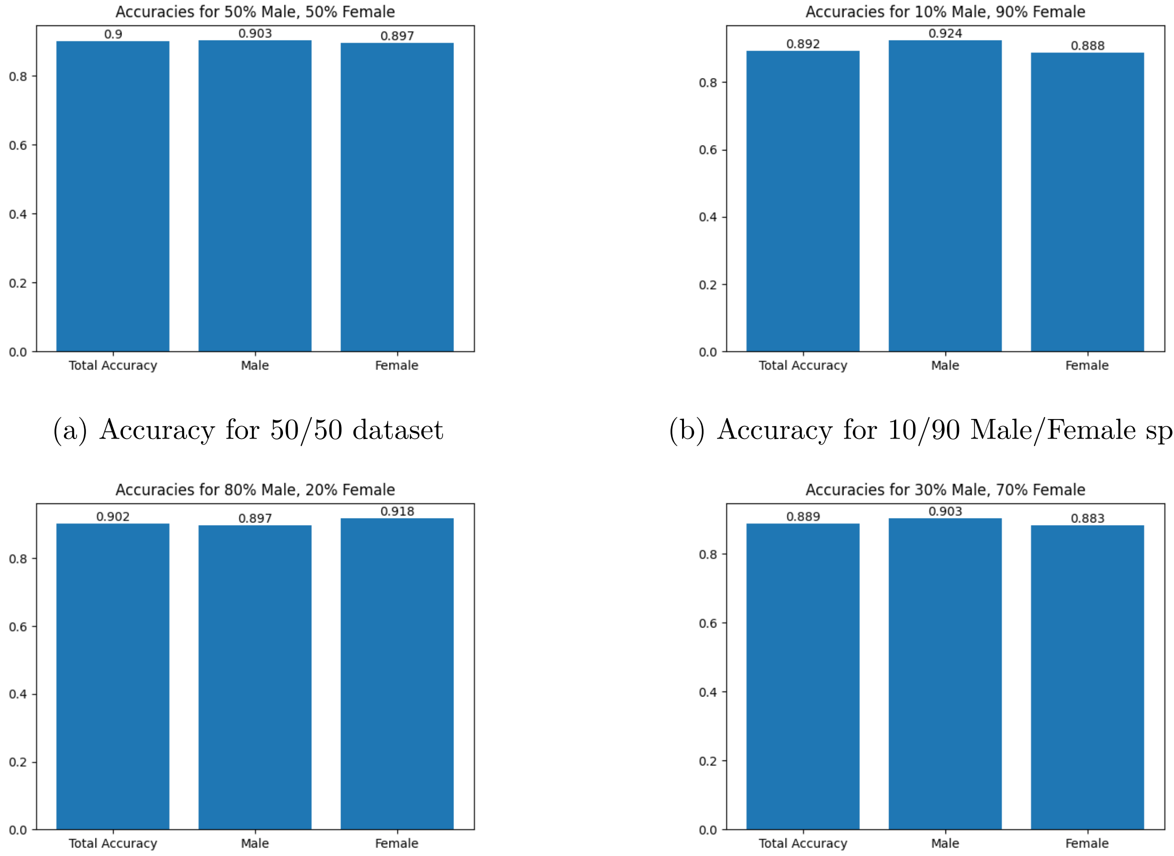
(a) Training loss for 50/50 Male/Female split

(b) Training loss for 30/70 Male/Female split.

(c) Training loss for 80/20 Male/Female split

(d) Training loss for 10/90 Male/Female split.

Figure 7: Figure showing Ring Loss with different proportions of male and female class



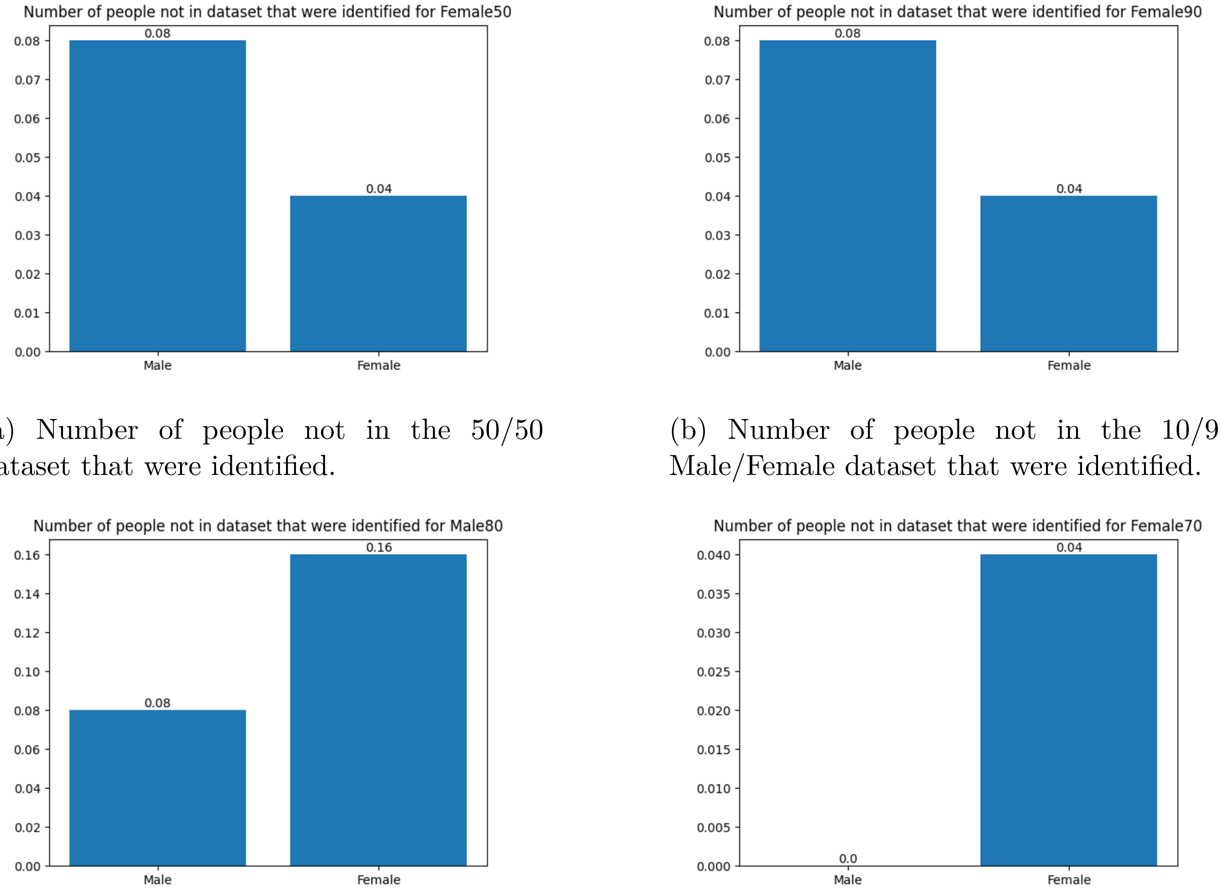
(a) Accuracy for 50/50 dataset

(b) Accuracy for 10/90 Male/Female split.

(c) Accuracy for 80/20 Male/Female split

(d) Accuracy for 30/70 Male/Female split.

Figure 8: Accuracies for models trained on imblanaced datasets



(a) Number of people not in the 50/50 dataset that were identified.

(b) Number of people not in the 10/90 Male/Female dataset that were identified.

(c) Number of people not in the 80/20 Male/Female dataset that were identified.

(d) Number of people not in the 30/70 Male/Female dataset that were identified.

Figure 9: Percent of people not in datasets who were positively identified

References

[1] Youtong Zheng, Dipan K. Pal, Marios Savvides. Ring loss: Convex Feature Normalization for Face Recognition. February 28, 2018. https://arxiv.org/pdf/1803.00130.pdf

[2] Weiyang Liu, et. al. SphereFace: Deep Hypersphere Embedding for Face Recognition. January 29, 2018. https://arxiv.org/pdf/1704.08063.pdf

[3] Chen, Kornblit, Swersky, Norouzi, Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. October 2020. https://arxiv.org/abs/2006.10029

[4] Zhang, Zhifei, Song, Yang. https://susanqq.github.io/UTKFace/

[5] Buslaev, Iglovikov, Khvedchenya,Parinov, Druzhinin, Kalinin. Albumentations: Fast and Flexible Image Augmentations https://albumentations.ai/

[6] Patrick Grother, et. al. Face Recognition Vendor Test Part 3: Demographic Effects. National Institute of Standards and Technology. December 2019. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

[7] https://www.csis.org/blogs/technology-policy-blog/problem-bias-facial-recognition

[8] Khan and Fu. One Label, One Billion Faces: Usage and Consistency of RacialCategories in Computer Vision https://arxiv.org/pdf/2102.02320.pdf

[9] Steed and Caliskan. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. https://arxiv.org/pdf/2010.15052.pdf