Transfer Learning Versus Regular Learning for Race Predictions on Medical Images

James Wen, QJ Yap
Harvard T.H. Chan School of Public Health
677 Huntington Ave, Boston, MA 02115
{james_wen, qijingyap}@hsph.harvard.edu

Abstract

As the growth of neural network applications increases, so does the need to assess the ethics of such methods. AI ethics is a burgeoning field especially with regards to medical applications where the need for fair algorithms is paramount for clinical use. Recent research has found that neural networks are able to predict the race of patients from chest x-rays despite no apparent racial biomarkers. This discovery is alarming as such results could indicate that neural network models can produce racially biased predictions. The purpose of this paper is to explore whether transfer learning for an unrelated task can predict race as accurately as a model explicitly trained to predict race. We took downsampled chest X-Ray (CXR) images and applied sets of data augmentations before running them through transfer learning models and explicitly trained models. This analysis attempted to predict Asian, Black/African-American, Native American, and White racial groups. We find that both transfer learning models and explicitly trained models were able to predict race based on chest X-Ray images. Specifically our models were quite accurate at predicting race for White patients. From an ethics standpoint, a neural network model that can extract patient race as a feature is highly concerning. Should a model be able to identify race, it may later bias results (potentially life threatening diagnoses) accordingly which may further historical inequities in medical research and clinical practices.

1. Introduction/Idea Overview

With the state of deep learning research in the field of computer vision progressing by the day, issues of algorithmic bias have come into greater prominence. In particular, AI models in the field of medical imaging often display racial disparity in the field of medical imaging, despite the absence of race biomarkers in these images. In 2021, a team of researchers found that even when not explicitly trained to do so, several popular deep learning model architectures were still able to distinguish for the new target variable of self-reported race based on looking at images of CXRs with extremely high performance. [4] In addition, the authors were unable to distinguish why this was the case, since human experts were not able to determine racial identity from the same images.

In the previous study, the models used to predict race used pre-trained weights with ImageNet. For our project, we are interested in comparing the effectiveness of a model created from scratch with the explicit goal of predicting race to another model created through transfer learning - incorporating pre-trained weights before further finetuning. We hope that this process might uncover some light on the effectiveness of transfer learning for an unrelated task as well as how potential racial biases can form.

2. Literature Review

In the paper by Banerjee et al., the same models were run again after attempting to account for different possible confounding factors, such as stratifying for age/sex, bone/tissue density, presence of disease, BMI, as well as trying to decrease image resolution and even removing the portions of images identified by Grad-CAM as regions of interest. Despite all this, performance across models remained relatively strong. [4] Surprisingly, the phenomenon of models trained on a different task still performing well is not uncommon. While it is not completely clear exactly why transfer learning appears to work so well, prior research has shown that using ImageNet to pre-train deep learning models perform well even with less labeled ImageNet training data, lesser images per class and reducing fine-grained recognition did not cause a large drop-off model performance, pointing to the astounding versatility of transfer learning methods. [9, 10]

In a more extreme case, another study found that adapting frozen pretrained transformers (FPT) initially built for Natural Language Processing problems were able to per-

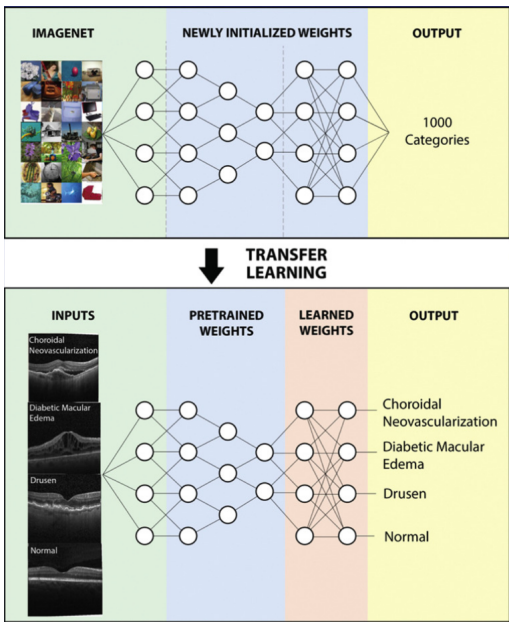


Figure 1. Taken from Kermany et. al. 2008 [3]

form tasks on different modalities, such as classifying the MNIST digit dataset up to test accuracy scores of 98.0%. [7] The same paper suggested that "FPT models underfit the data, which lends them to further improvements by increasing model capacity", which gives us inspiration to experiment further with adding learned weights to pre-trained weights. This is a concept that has been tested in the field of healthcare specifically, as illustrated in Figure 1 from D. Kermany et. al. (2018). [3]

3. Data Sources

Our project is limited by the accessibility of CXR data freely available online. Due to this limitation, our main dataset was obtained from CheXpert, a large dataset containing 224,316 chest radiographs of 65,240 patients. [5] This data was collected from studies at Stanford Hospital between October 2002 and July 2017 and its machine-learned label method was validated against board-certified radiologists to ensure accuracy in the labels. Due to the nature of our target task of race, potential labeling issues was not a factor in our analysis.

4. Method

4.1. Model Architecture

One goal of the project was to partially replicate the method used by Banerjee et al. in using pre-trained weights on ImageNet. [1] Since Banerjee et. al. selected Resnet34 as the model architecture for external validation due to the similar excellent model performance on all architectures,

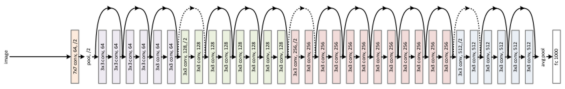


Figure 2. ResNet34 Architecture, [8]

we intended to do the same, with the additional complexity of fine-tuning the fully connected layer as seen in Figure 2. [8]

In addition to the above, we also initialized and tuned another Resnet34 model from scratch, but this time built specifically with the purpose of training specifically for the task of race. By looking at the Average Precision (AP) scores (area under curve for Precision-Recall Curve) as well as the test classification accuracy between the two models, we will hopefully be able to obtain a better understanding of how a pre-trained model performs compared to a model from scratch, even in the situation of a target variable difficult to perceive with the human eye.

4.2. Transfer Learning Models

To enhance our ability to compare models, our team decided to apply the same model structure and training hyperparameters to all of our transfer learning models. As previously noted, we froze the ResNet34 weights (trained on Imagenet) and added a dense layer of 4 nodes (one for each racial group). We then applied a softmax activation function to arrive at the final prediction probabilities. Our loss function to train the dense layer was categorical cross entropy loss and our optimizer was Adam with a learning rate of 0.00001. All models utilized 100 epochs.

4.3. From Scratch Models

Due to the computational complexity of training ResNet34 models entirely from scratch, some modifications to the prior process had to be made. The models were initialized with random weights across all parameters unfrozen. Just like the transfer learning models, we also initialized the scratch models with the parameter include_top = False, then added the same dense layer of 4 nodes with softmax activation function. While our loss function to train the dense layer was categorical cross entropy loss and our optimizer was Adam once again, due to the sheer number of parameters that needed to be trained, we utilized a learning rate of 0.001, a magnitude of 100 times higher. Finally, the computational demands led to a much higher training time for the scratch models compared to the transfer learning models. Thus, the scratch models only went through 30 epochs of training for the purpose of this project.

4.4. Data Preparation and Pre-Processing

For the sake of reproducibility, we ensured that we standardized three different seeds: a Tensorflow seed of 261,

random seed of 1126 and NumPy seed of 1030. The race labels from CheXpert were read in from a separate file, and initially had multiple categories for similar race designations (eg. two different labels indicating Black or African American and Black). We thus compressed all the different labels into 4 main categories: Black/African American, White, Asian, and Native American. After removing missing values from the dataset, we were left with 125491 White samples, 23372 Asian samples, 11961 Black/African American samples and 3346 Native American samples. It should be noted that the dataset provides a Hispanic designation as an additional label to race. For example, someone could be classified as White-Hispanic, Asian-Hispanic, Black-Hispanic etc. For the sake of simplicity, we ignored this feature of the data. The dataset also did not provide labels for those who identify as Latino as other datasets do.

Due to the imbalanced nature of our data classes, it is possible that the process of randomly splitting our dataset into train, validation and test sets could result in our models training on a dataset with a severe under representation of one of the race labels. We used a stratified train_test_split done on the race categories in order to avoid this problem. There were a total of 103364, 44299 and 16407 train, validation and test samples respectively after this process.

Finally, the code for augmenting data, training and benchmarking the models was written in the Keras framework with TensorFlow backend. The models were trained with the computing resources provided from the Google Colab Pro Plus GPUs, and were benchmarked by looking at overall Accuracy, F1 scores in predicting each individual race label, as well as AP.

4.5. Data Augmentation

As with many computer vision neural network models, data augmentation was a key consideration for our project. Given that CXRs are not necessarily standardized and image quality issues can occur, we decided to run our models utilizing different sets of data augmentations. Further, all images were rescaled to 150 by 150 to 1) see if the model can accurately predict race despite a lower resolution and 2) to scope the project given computational constraints. For our model built with pre-trained weights from ImageNet and further fine-tuning, we applied the following four sets of transformations:

Version 1:

- rescale = 1./255, rotation_range = 90, width_shift_range = 0.2, horizontal_flip = True, vertical_flip = True, shear_range = 0.3, zoom_range = 0.1

Version 2:

- rescale = 1./255, fill_mode = 'constant',rotation_range = 10,horizontal_flip = True, zoom_range = 0.05

Version 3:

- rescale = 1./255 (no further data augmentation besides re-scaling)

Version 4:

- rescale = 1./255, horizontal_flip = True

After training the four different pre-trained models, as a result of much higher computational cost, we decided to use only two different models from scratch. We would compare the model trained with the data augmentation technique that yielded the highest performance on our chosen metrics of test accuracy and PRAUC, as well as Version 3 (only rescaling). This eventually turned out to be Version 2 with parameters as listed above. This process allowed us to save on resources both in hardware as well as training time, while still giving us an opportunity to compare pre-trained and scratch models under similar conditions.

5. Results

5.1. Transfer Learning Models

Table 1. Accuracy

Version	Accuracy	Asian	Afr/Am	Native American	White
1	0.397	0.16	0.14	0.00	0.57
2	0.559	0.30	0.18	0.02	0.72
3	0.692	0.36	0.19	0.00	0.77
4	0.346	0.25	0.16	0.01	0.49

This table provides overall accuracy of the model and the F1 scores by race for each version of our transfer learning models.

Table 2. Average Precision

Version	Overall	Asian	Afr/Am	Native American	White
1	0.38	0.18	0.09	0.02	0.80
2	0.57	0.25	0.13	0.03	0.83
3	0.66	0.29	0.14	0.03	0.86
4	0.34	0.24	0.12	0.03	0.84

This table provides the Average Precision (AP) by race for each version of our transfer learning models.

For our transfer learning models, Version 3 (without any data augmentation) performed the best with an accuracy of 0.629 and AP of 0.66. Overall, all of the transfer learning models performed poorly. Further, there was no parity in F1 scores between the racial groups. The models seemed to be able to accurately classify White patients but suffered with

minority racial groups. This could be due to the class imbalance in the dataset though it should be noted that Banerjee et al. were able to achieve parity in accuracy metrics despite the same class imbalances.

5.2. Models from Scratch

Table 3. Accuracy

Version	Accuracy	Asian	Afr/Am	Native American	White
2	0.822	0.60	0.48	0.12	0.90
3	0.769	0.42	0.32	0.06	0.87

This table provides overall accuracy of the model and the F1 scores by race for each version of our models built from scratch.

Table 4. Average Precision

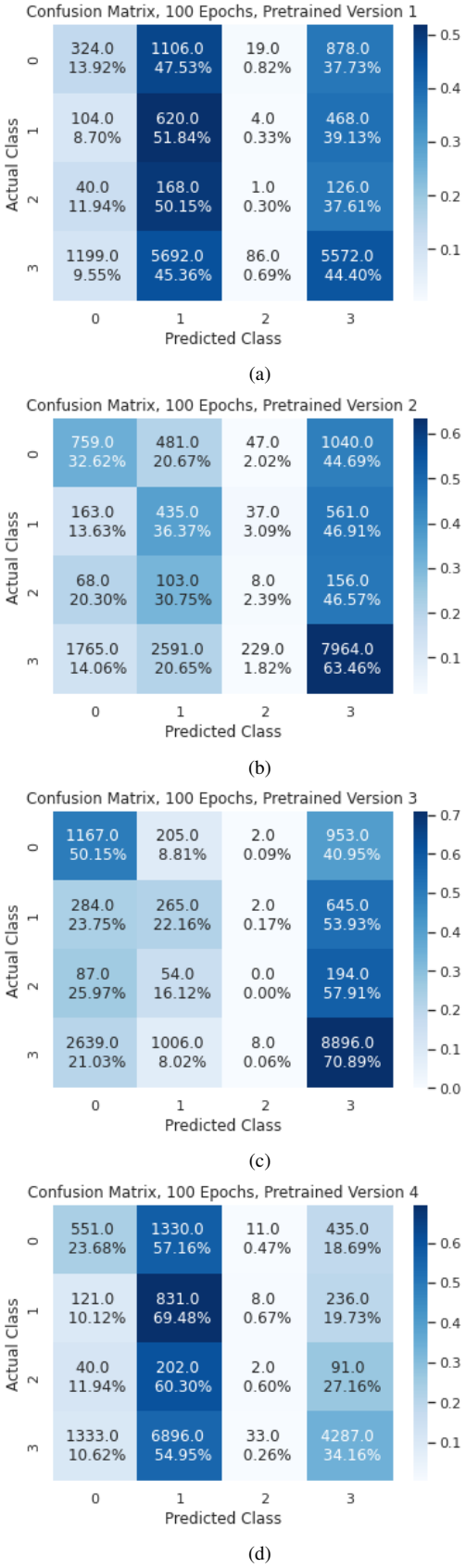
Version	Overall	Asian	Afr/Am	Native American	White
2	0.89	0.65	0.50	0.12	0.95
3	0.82	0.45	0.30	0.06	0.90

This table provides the Average Precision (AP) by race for each version of our models built from scratch.

Our models built from scratch performed much better than their counterparts in the transfer learning section. Interestingly, unlike the transfer learning models, it was the version with data augmentation (Version 2) that performed the best with an accuracy of 0.822 and AP of 0.89. Further, the average precision and F1 scores for individual racial groups improved compared to the transfer learning models but still suffered from lack of parity.

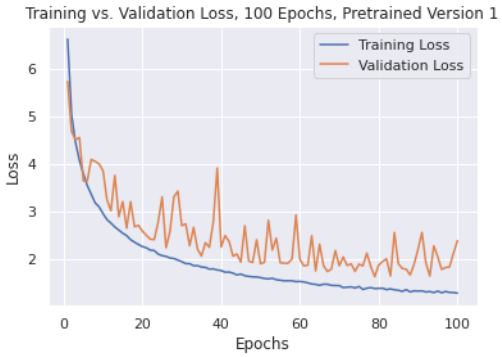
6. Discussion

Overall, we found that the models built from scratch performed far better than the transfer learning models in all of the measured metrics (overall accuracy, F1 score by race and AP). While the exact mechanism that could explain this disparity remains largely unknown, we can offer up some possible hypotheses. Firstly, due to training time constraints, we were unable to tune some of our hyperparameters such as learning rate (0.00001 and 0.001 for the transfer and scratch models respectively). The choice of an extremely low learning rate for the transfer learning models was a decision informed by the need to only fine-tune the final dense layer, but a marginally higher learning rate could have decreased the bias of the models. Looking at the plot in Figure 4c comparing training and validation loss for the best performing transfer learning model with no augmentation, it looks like the validation loss was only slightly higher

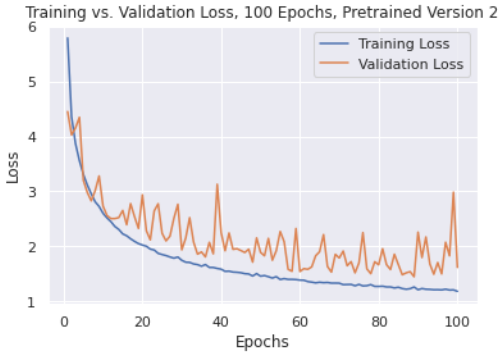


(a) (b) (c) (d)

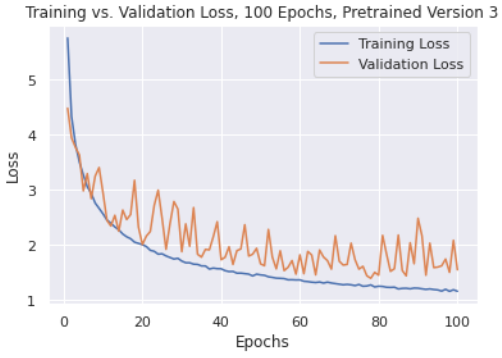
Figure 3. 0 = Asian, 1 = Black/African American, 2 = Native American, 3 = White



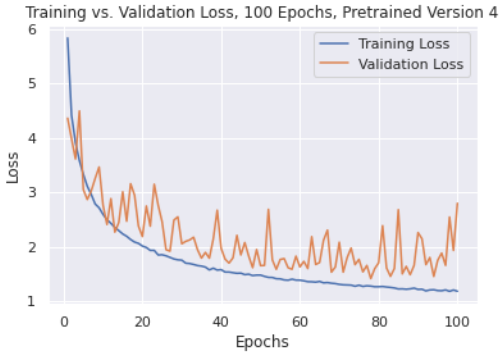
(a)



(b)

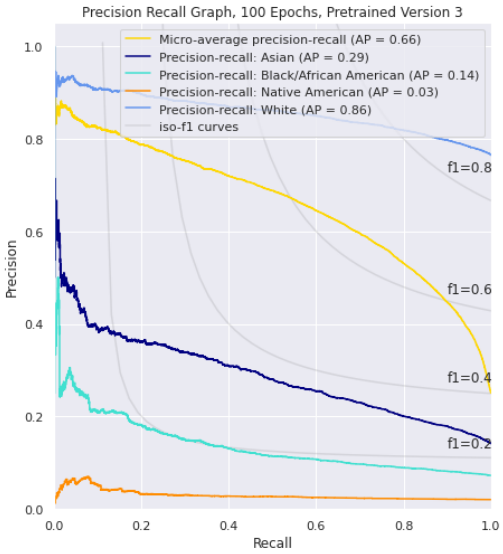


(c)



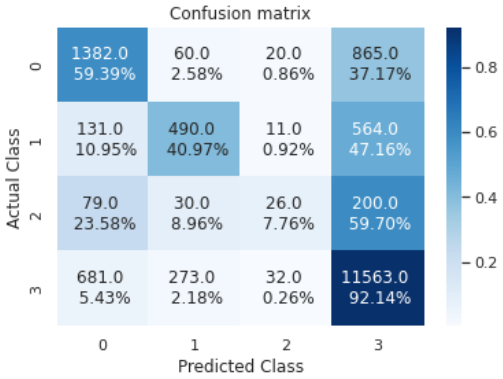
(d)

Figure 4

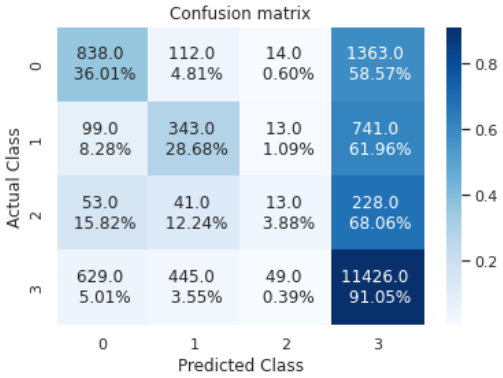


(a)

Figure 5. Version 3: Pre-trained Model, Precision Recall Curve

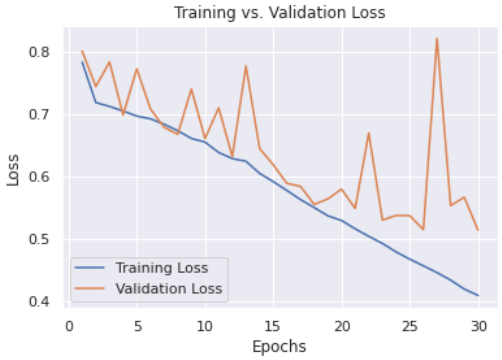


(a) Version 2: Model from Scratch

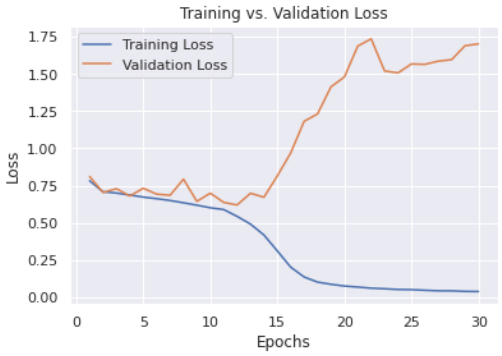


(b) Version 3: Model from Scratch, No Data Augmentation

Figure 6. 0 = Asian, 1 = Black/African American, 2 = Native American, 3 = White

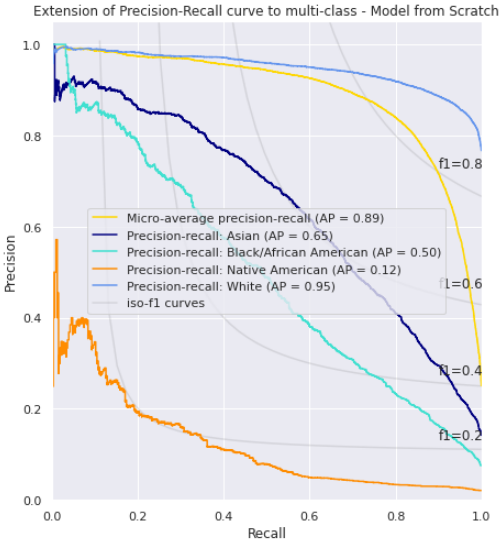


(a) Version 2: Model from Scratch



(b) Version 3: Model from Scratch, No Data Augmentation

Figure 7



(a)

Figure 8. Version 2: Model from Scratch, Precision Recall Curve

than training loss. While this does show that the model does a good job at not overfitting the training data, it is possible that there might have been more variation in the data that the model was not able to capture.

Another reason why this discrepancy exists could be that downsampling the data to a lower resolution hurt the performance of the transfer learning models far more than that of the scratch models. This is a theory that makes sense, since the further away a new task is from the original task the model was trained from, the further the performance of transfer learning would be expected to drop as well. With less information provided in a given image, there might be a greater need to train more model parameters in the full ResNet34 architecture in order to preserve that information. Since we only modified the final dense layer of our transfer learning models, it is likely that the ImageNet trained weights performed poorer than expected at a task even further from what was planned originally.

Interestingly, the best transfer learning model was the one that did not use data augmentation, but the best model built from scratch was the one that used data augmentation, an unexpected result. We attribute this to the fact that the models built from scratch had a much poorer baseline to begin with before training, as well as the lower number of epochs in training the models. This would have necessitated the need for data augmentation to help improve training results. On the other hand, the ResNet34 pre-trained weights were obtained by training the network on ImageNet data with relatively minimal augmentation, with the authors of in the original paper using scale augmentation, horizontal flip for some images and a standard color augmentation. [8] On the other hand, Version 2 of our data augmentation methods which ultimately proved to be the best transfer learning model also incorporated a constant fill around the image, 10 degrees of random rotation, and a zoom range of five percent. These additional parameters might have thus hindered the transfer learning models when compared to using no data augmentation since the method of image augmentation used was dissimilar to the image augmentation that resulted in the ImageNet pre-trained weights.

Comparing our results to that of Reading Race, it is clear that there was a significant dropoff in performance compared to the findings of that paper when it came to predicting race. For example, the paper achieved an ROC-AUC score of 0.98 for black patients using the CheXpert dataset for internal validation, indicating extremely high performance results. On the other hand, our best performing transfer model had suboptimal results on our performance metrics, achieving a paltry F1 score of 0.19 for the Black/African American race label. A possible explanation for this could be the downsampling of image data that was needed due to the lack of computational power available. However, this explanation does not fully explain the discrepancy. Banerjee et al. also experimented with downsampling images, working with resolutions ranging from 4x4 to 512x512. They found that while performance did degrade as resolution decreased, the model performance remained

strong even at resolutions slightly under 100x100, unlike our transfer learning results. [4]

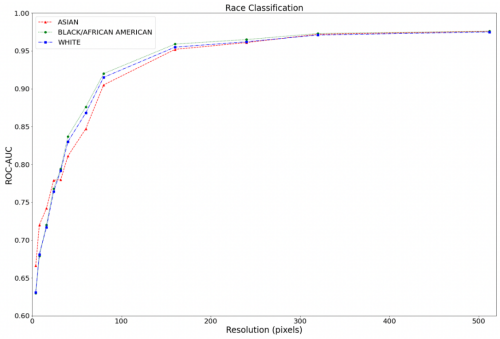


Figure 9. Increase in ROC-AUC has resolution increases. Pic taken from [4]

In addition to the previously mentioned differences, all of our models, regardless of whether built from scratch or through transfer learning, were not able to achieve parity between the different racial groups based on our metrics. This was a result that Banerjee et al. were able to achieve, as can be observed from Figure 9. This may be explained in part since we decided to include indigenous groups as a fourth race label in an effort to recognize an historically significant demographic group in the United States, while Banerjee et al. only assessed Asian, Black/African-American, and White patients. Given that the Native American labels only amounted to approximately 2.04 percent of the CheXpert dataset, in hindsight excluding these samples might have been a better decision. However, this does not completely explain how the original authors were able to achieve their exceedingly high level racial parity, and it would be illuminating if we had access to their original code to understand their methodology in greater detail.

Despite the shortcomings covered above, the overall results of both the transfer learning and scratch models were still promising. In the case of the best transfer learning model, an accuracy of 0.629 could be explained by the overwhelming proportions of White race labels, with all other labels not able to achieve a precision, recall or F1 score above 0.50. However, the best model built from scratch had an F1 score of 0.90 and AP of 0.95 for White race labels, and an F1 score of 0.60 and AP of 0.65 for Asian race labels, which are encouraging results that improve on a simple naive majority class predictor. These metrics in absolute may not seem promising by themselves, but in context, our project revolves around predicting race labels using only CXRs, which seemingly have no racial biomarkers and are indistinguishable from each other in terms of race classification by human experts. Considering the inputs to our models, the fact that the models were able to partially distinguish racial labels in the first place is already a surprising result.

7. Limitations

Our project faced unique challenges that required us to simplify and scope down our original proposal. The first issue we encountered was data access. Medical data is not as publicly available compared with datasets in other fields. Many datasets are proprietary to hospitals or require certifications for access. Thus, our project had to rely on CheXpert, one of the few readily available CXR datasets with race labels. While CheXpert is a thorough dataset, we would have liked to build models incorporating other data sources for potential image diversity and quality.

The disparities between racial groups likely originate from the highly imbalanced nature of CheXpert. We would be curious to see if such imbalance remained if the classes were more even. A potential idea we had to solve this issue given more computation power was to run a third and fourth iteration of all of the models by 1) over-sampling images belonging to minority groups and 2) under-sampling images belonging to White patients.

In terms of modeling, we did encounter issues related to computational power due to the lack of access to cloud computing servers. The CheXpert dataset comes in two versions: 1) the original dataset containing images of about 500x500 and 2) the downsized version with images of about 350x350. Knowing we would only have access to a GPU via our Google Colab accounts, we opted to work with the downsized dataset. Our original plan was to produce models and to compare accuracy scores on sets of images with different resolutions. However, even after upgrading to Pro Plus, we did not have enough power to run any models with a resolution higher than 150x150. (150x150 was originally our lowest resolution threshold planned for analysis) Thus, we decided to pivot and focus on data augmentation instead.

In addition to being able to experiment with different image resolutions, our original plan also included running experiments on different models besides ResNet34. It would be interesting to assess if predicting race is unique to ResNet34 or if other models trained on ImageNet such as AlexNet, VGG, or DenseNet can achieve similar results. Base model experimentation could further help clinicians and AI researchers root out what is allowing for racial predictions to happen. Should other models not be able to detect race, this result would lead researchers to look more into the features ResNet34 extracts and the structure of that model. Alternatively, if multiple pre-trained models are able to capture race, that could potentially push researchers to look more at data collection methods such as the X-Ray machines themselves or potential issues with data labeling. For example, one could hypothesize that X-Ray machine calibration itself could be racially biased which if true would likely be reflected in results from any pre-trained

model.

Finally, while we were attempting to partially replicate the methods used in Reading Race, we did not use the same metrics of ROC-AUC that Banerjee et al. relied on, instead using a combination of accuracy, F1 score and AP. We had chosen the latter three measures because we wanted to understand our results from different perspectives, but in hindsight we should have also taken ROC-AUC scores into account when conducting our analysis.

8. Conclusion

Our findings indicate that while transfer learning is powerful, explicitly training a model to predict race will still produce better results. The transfer learning models performed poorly on minority racial groups but did well on predicting White patients. Models explicitly trained on race had better accuracy predicting White and Asian patients but struggled with other racial groups. Again, such discrepancies in race were not observed in Banerjee et al.. We attribute these differences to a dataset imbalance and limited computational capabilities which forced us to downsample an already downsampled dataset and cut down on training time. Such issues present their own questions related to AI ethics.

The disparity in results between the different race categories has some implications in the sphere of ethical AI. In this study, we were using the race labels as the primary target of interest here as opposed to predicting other outcomes that may have significant medical implications, such as various diseases. Thus, the issue of racial bias might not necessarily be a factor directly here. However, our findings do give us an indirect window of insight into why racial bias might be prevalent in much of medical research. One of our biggest limiting factors in model performance was the presence of imbalanced race labels in CheXpert. Chakraborty et al. proposed a modification of the popular oversampling algorithm Synthetic Minority Over Sampling Technique (SMOTE), Fair-SMOTE, which attempts to solve data imbalance by taking into account protected attributes like race, gender, and age when oversampling a class of the target variable. [6] In the current age of rapid advancement in deep learning research, developing methods such as Fair-SMOTE is crucial when it comes to mitigating bias in AI.

Further, retraining large models for explicit tasks presents environmental issues related to computational power. Strubell et al. found that training an NLP neural network with tuning and experimentation produced over two times the carbon dioxide that the average American emits over one year. [2] Given contemporary discussions on climate change, AI researchers not only need to assess the direct practical consequences of their model (in this case, how patients are treated) but also environmental externalities brought on by training large neural networks. In the case of this project, transfer learning with frozen weights would cut down on computational power and thus reduce environmental impact. However, transfer learning did not prove to be as powerful as explicitly training all of the weights. Thus, more robust future research on this topic may need to carefully weigh the environmental impacts of a similar project versus unearthing new information as to how a neural network can predict race. Arguably, as this is a topic related to clinical patient outcomes, such environmental consequences are more justifiable than in other more traditional areas of AI research.

Overall, this project elevated many issues related to ethics both in an AI setting and in a clinical setting. Further research is needed to more specifically identify what is allowing neural networks to pick up race on data that no clinician would deem as reliable for such classification problems. Should the root cause of the problem remain elusive, the next steps are to pivot to harm control measures to prevent future neural network models from biasing clinical diagnoses on race. While finding the root cause of the issue is of course ideal, given the uncertainty surrounding this topic, harm control methods are inevitably needed in the short term. Such harm control measures could be closely monitoring diagnoses recall scores for different racial groups to ensure equity in prediction power. Again, recall requires special attention due to the extreme consequences a false negative prediction can have in a medical setting. Overall, we hope that future experimentation in this area is conducted and believe that this question, while remaining open, presents a unique opportunity for AI and medical researchers to come together and produce more insights into how computation can further improve patient outcomes.

9. Personal Contribution

- what i did number 1
- what i did number 2

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009. 2

[2] Ananya Ganesh Emma Strubell and Andrew McCallum. Energy and policy considerations for deep learning in nlp., 2019. 8

[3] D. Kermany et. al. Identifying medical diagnoses and treatable diseases by image-based deep learning., 2018. 2

[4] Imon Banerjee et. al. Reading race: Ai recognises patient's racial identity in medical images., 2021. 1, 7

[5] Michael Ko Yifan Yu Silviana Ciurea-Ilcus Chris Chute Henrik Marklund et al. Jeremy Irvin, Pranav Rajpurkar. Chex-

pert: A large chest radiograph dataset with uncertainty labels and expert comparison., 2019. 2

[6] Suvodeep Majumder Joymallya Chakraborty and Tim Menzies. Bias in machine learning software: why? how? what to do?., 2021. 8

[7] P. Abbeel I. Mordatch K. Lu, A. Grover. "pretrained transformers as universal computation engines., 2021. 2

[8] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2016. 2, 6

[9] A. Efros M. Huh, P. Agrawal. What makes imagenet good for transfer learning?, 2016. 1

[10] Quoc V. Le S. Kornblith, J. Shlens. Do better imagenet models transfer better?, 2019. 1