Skeletal Hands Group: Trevor Brokowski, Weiqian Lin, Nikhil Patel, Anish Patel, Shiqin Tan

Professor Wollman, TA Perrie, TA Hemminger

C&S Biology 185

10 March 2022

Final Report

# **ABSTRACT**

Inspired by Chen et al. (2020)’s incorporation of gender information in bone age assessment (BAA), we searched for other possible factors that influence the outcome of BAA and found ethnicity to be key information in the literature. We examined whether the addition of an ethnicity feature could improve the original model's performance at bone age prediction. We preprocessed the data from Digital Hand Atlas (DHA) to adjust to the original model, and one-hot encoded the ethnicity feature to concatenate it alongside the gender and the image features in the age regression network. Using the hyperparameters that yield the optimal performance in the paper, we trained both the DHA dataset and a subsample of the RSNA dataset with the same size and compared the resulting test mean absolute errors (MAE). We found that there was no significant improvement with the inclusion of ethnicity in the model, but further research involving hyperparameter tuning is needed to confirm.

# **INTRODUCTION**

Bone age assessment (BAA) is often performed to determine subjects' bone age, which indicates their biological and structural maturity better than the chronological age. Traditionally conducted by radiologists, The process of BAA includes referencing and labeling the X-ray images by expert radiologists in both traditional manual methods and recent deep learning-based approaches. To solve this problem, Chen et al. (2020) proposed an attention-guided approach to automatically localize the regions of interests (RoIs) and output age distribution instead of single age labels. The team’s model achieved performance comparable to state-of-the-art methods.

One of the measures taken to improve the model is the incorporation of gender information in the age regression network. Because of the physiological differences between men and women that affect bone development, the inclusion reduced the error in predicted age by approximately a year. Inspired by such an approach, we searched for potential influencers of the BAA outcome from literature. Based on recent research of possible differences in bone density (Zengin et al.,2015) and rate of bone maturation among different ethnicities, we attempted to concatenate ethnicity features into the model of Chen et al. in the same way as gender information and evaluated the efficacy of the resulting model by mean absolute errors (MAE) of the test set.

Aside from the RSNA BAA Challenge dataset with 12,611 labeled images provided in the study by Chen et al., we obtained the dataset of this research from the Digital Hand Atlas of University of Southern California (DHA) dataset, which hosts 1390 hand radiographs with gender and ethnicity information.

# **METHODS**

Our initial attempts at reimplementing the original model on the entire RSNA dataset were obstructed by the limited RAM provided by Google Colab, which cannot load all of the data. The subsequent change to progressive loading with the Image Data Generator resolves the issue, but the total training time would still be too long with the GPU resources we had. We then decided to randomly select 1390 images from the RSNA dataset, a size equivalent to that of DHA, so we could evaluate the incorporation of ethnicity information when the dataset size stays the same.   
 To preprocess the DHA data, we iterate through each folder of the DHA dataset and extract the ethnicity and gender information from the filename. Using a custom python dictionary, the genders were binarized: 1 for male and -1 for female. The 4 ethnicities: Asian, African American, Caucasian, and Hispanic, were one-hot encoded because the nature of the ethnicity feature is categorical with no ordinal relationship. The age of the patient was computed by finding the difference between the date of the study and the patient’s date of birth listed in the DICOM file. Although Pydicom provides the functionality needed to load and save an image as an array, we were unable to utilize this as Pydicom loads in a one-channel image wherein a 3 channel image is needed to work with the base models. Thus, OpenCV was used to load and save the JPEG images as 3-channel image arrays to the correct dataset.

Ideally after loading the DHA data, we should pass it through the entire modified localization and regression networks. However, with the limited computing resources, we directly train the original images, instead of three cropped images as original paper, through our modified model with both gender and ethnicity concatenated to the image feature before the FC layers in the age regression network with Xception backbone.

# **RESULTS**

Bone Age Regression result:

|  | DHA  Original image | RSNA (size = 12,611)  H+R1+E | RSNA (size = 1309)  H+R1+E |
| --- | --- | --- | --- |
| Test MAE | 9.4 | 4.7 | 8.3 |

H: localized hand region, R1: region of interest I, E: original image with R1 masked out, MAE: mean age error

Training algorithm to determine BAA accuracy exceeded 6 hours of runtime before experiencing runtime error. Each epoch took approximately 300 seconds with Google Cloud Virtual Machine instance. Due to CPU and GPU limitations of the Google Cloud free trial, runtime errors were expected.

# **DISCUSSION**

Algorithmic bias and the inclusion of racial/ethnic data in models involving artificial intelligence is of high concern, garnering much attention in the field of computational research due to the potentially dangerous implications it can have in a healthcare setting. AI has shown the capacity to recognize a patient's racial identity only from objective image features in radiological imaging (Banerjee et al., 2020). In many instances, uses of deep learning in the medical setting have also proven to reproduce health care disparities along racial and ethnic lines, especially in radiology (Sorin, 2021). This is because an algorithm’s performance is dependent on the training data that may be imperfect and reflect biases and inequities in society at large. Even though our training data is roughly evenly distributed between the four ethnicities, we must be aware that the algorithm may perform worse on ethnicities that constitute a minority of data available, such as people from Pacific islands, and amplify existing biases.

This is of even larger concern when known biological differences exist along ethnic lines, which AI models can fail to account for. Recent studies highlight that there are significant differences in bone density between different ethnicities (Zengin et al.,2015)and that the rate of bone maturation in children differs across ethnic lines (Ontell, 1996). Thus, if a model is used to assess bone age by objective image features like bone density, there could be large ethnic discrepancies if this information is not included within the model. There have been a number of proposed methods for addressing the aforementioned issues: including entirely separate models for different ethnicities, including racial/ethnic information in various ways as inputs or outputs to models, or manually correcting images for known differences in objective image features (like HU adjustment for bone density differences). To this point, there is no consensus in the literature regarding a most optimal solution.

The present study tackles this issue by concatenating the ethnicity feature in the final regression network, just before the softmax function. Our preliminary results do not demonstrate that there was a significant improvement in the age regression accuracy with the inclusion of ethnicity. However, we did not have the capacity to fine-tune the model hyperparameters and determine the most optimal ones due to the limited time constraints for conducting this research project and the lengthy running time inherent in training multistage image processing models. Hyperparameter optimization may have resulted in greater accuracy with the ethnicity features, representing a starting point for future directions with regards to this research.

Other limitations included processing power issues with google colab slowing data loading speed and capping available resource utilization after a certain amount of time. This could be improved by using Kaggle or other computing resource platforms in the future. Other improvements that could be employed with additional time include generating a stat CSV for the DHA dataset for ease of analysis as it did not have one, and training the RNSA model again with masked images to get region 2.

In summary, this method of ethnicity inclusion represents a possible solution to improve bone age assessment by deep learning. However, additional research involving hyperparameter tuning with greater computational resources is required to assess whether this represents a true improvement compared to the baseline model. Confirmation of our preliminary hypothesis would be a significant contribution to the field of computational radiology, showing an optimistic path forward to handling algorithmic bias and accounting for ethnic differences to ensure the most equitable treatment of patients that rely on radiological tools such as bone age assessment in the future.

# 

# REFERENCES

1. Banerjee, Imon, et al. "Reading Race: AI Recognises Patient's Racial Identity In Medical Images." arXiv preprint arXiv:2107.10356 (2021).
2. Sorin, Vera, and Eyal Klang. "Artificial Intelligence and Health Care Disparities in Radiology." *Radiology* 301.3 (2021): E443-E443.
3. Zengin, Ayse et al. “Ethnic differences in bone health.” *Frontiers in endocrinology* vol. 6 24. 17 Mar. 2015, doi:10.3389/fendo.2015.00024
4. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol. 1996;167:1395–1398.

# 

# APPENDIX

Datasets Used:

1. The RSNA dataset:

[kaggle.com/kmader/rsna-bone-age](http://kaggle.com/kmader/rsna-bone-age)

1. The Digital Hand Atlas of University of Southern California: [ipilab.usc.edu/research/baaweb/](https://ipilab.usc.edu/research/baaweb/)

Our commented code, presentation, and scripts can be found in:

<https://drive.google.com/drive/folders/1clBYgN39kyPXsoWOzqe4nxfJC6Gkd5Le?usp=sharing>

Implementation details:

Implemented on Google Colab Pro with about 24GB RAM, 2x vCPUs, K80/T4/P100 GPU (no guaranteed resource). The localization network uses InceptionV3 as the baseline model, is initialized with weights pre-trained on ImageNet, batch size of 32, and trained using the Adam optimizer with a batch size of 32 and the learning rate set to 0.0003 for the first 60 epochs, then 0.0001 for the next 30 epochs. The localized hand region (threshold = 50), RoI 1 (threshold = 20), and original image with RoI 1 masked-out, are fed into the regression network for the RSNA training. The regression network uses Xception as the baseline model and is trained using the Adam optimizer with a batch size of 16, and learning rate = 0.0003 for the first 40 epochs, then 0.0001 for the next 20 epochs. The dataset is split into test/valid/train based on the original paper’s proportion (55/55/1280)