MIE429
Machine Intelligence Capstone Design:
Team 11

Client:
AI Development and Evaluation Lab at Trillium Health Partners

Students:
Benjamin Cheng (1004838045)
Katarina Chiam (1004908996)
Anton Liu (1005102407)
Haoran Jayce Wang (1004927163)
Richard Yang (1004908870)

Introduction

Background

Pediatric bone age estimation helps evaluate the maturation of skeletal structures within youth under 18 in order to aid in the diagnosis of conditions impacting physical growth and development [1][2]. This process is commonly done on the hands (Figure 1) and measured in months or years. Radiologists compare a given hand X-ray to a standard atlas, a set of X-rays with labeled bone ages, and identify the closest matching images in the atlas to the patient's X-ray, thereby determining its bone age [2]. Notably, bone development differs based on the individual's sex, which is reflected in the atlas [3]. This comparison process is done by radiologists, and its manual nature makes it time-intensive. As such, a strong motivation exists to automate this process [4]. Specifically, recent studies have shown Deep Learning (DL) models can be used to generate reliable bone age predictions [5].

Figure 1: Example of an X-ray hand scan [5]

Problem

Our client, Trillium Health Partners (THP), has requested a machine learning (ML) solution to improve the efficiency of radiologists while maintaining accuracy in bone age estimation. The solution will utilize an input hand X-ray and may leverage the patient's sex as well. Additionally, this model must be deployed within an automated pipeline. The pipeline will fetch patient data in the form of DICOM, the standard data format within medical imaging [6], from an image server and return its predictions along with various visual tools to further help the radiologist.

Deliverables

Requirements

- Should train and deploy a DL model to predict the bone age of hand X-ray images
- Should generate visuals to inform the radiologist of the prediction:
  - Visual means of explaining the model prediction
  - The closest matching X-rays from the bone age atlas
  - Plot of model prediction over a standard bone age growth chart
- Should integrate the DL model and visual outputs into a pipeline as a DICOM node

Constraints

- Must train a model with Mean Absolute Error (MAE) within 5-12 months compared to the radiologist's estimate in the dataset used
  - The human-labeled dataset has around 5 months of error between sources [5]. Producing a model with a lower MAE may suggest the model learning potential errors within the dataset.
  - The client has requested the MAE to be under 12 months

Data

The RSNA Dataset

In order to train our model, we required a dataset with examples of X-ray images with their associated bone age prediction from radiologists (ground truth), as well as the patients' sex. To

that end, our client suggested that we utilize the dataset provided by Radiological Society of North America (RSNA) in their 2017 Pediatric Bone Age Challenge [5].

Since there is no way to identify bone age objectively, multiple radiologists may predict different bone ages for the same patient. To produce the singular prediction in the dataset, RSNA has computed the ground truth values as a weighted consensus from 6 sources [5]. As a reference, the Mean Absolute Error (MAE) for these radiologists from the resulting ground truth was in the range of 5 to 7 months. Because of this deviation in human accuracy and lack of alternative data sources, we consider 5 months a lower bound in MAE for our model.

The data from RSNA was generally of good quality and was usable as-is. As the dataset is designed for machine learning tasks, it is distributed in a training and validation split. The images in the dataset are grayscale in PNG format, with the bone age prediction and patient sex provided in a CSV file. Other important characteristics are summarized in Table 1 with a more detailed histogram illustrating bone age distributions in Appendix D.

	Training	**Validation**
Number of Examples	12611	1425
Patient Demographics (% Female)	46%	46%

Table 1: Summary of Dataset Characteristics

A test set of data acquired by THP will be used by their clinicians to evaluate the model's performance. Data provided by THP will be in DICOM format, from which the X-ray image and sex will be extracted. This will be further elaborated in **Integrated Application**.

Despite the good quality of the RSNA data, some X-rays are very low in contrast between the background and hand. Consultation with THP revealed that these circumstances will not occur with their equipment, and need not be considered. Accordingly, we narrow the validation set to better represent the test performance using Contrast-Based Filtering which will be elaborated on in **Methods**.

Data Manipulation

Data Preprocessing

Despite good quality data from RSNA, we performed a few preprocessing steps to help improve the performance of our model.

Signature Removal

Within the RNSA dataset, images contained a "signature" that was burned into the X-ray images (Figure 2). The intention of these is to label the images for the radiologists and the images we expect THP to send to our model will contain this signature as well. These signatures however do not provide any meaning when it comes to detecting bone age and may distract the model or accidentally include the model's output.

Figure 2: Examples of common signatures [5]

After a handful of alternatives were explored, feature matching yielded the most reliable results. This approach requires two inputs: a query image and a search image. With the goal of matching the features of the query image onto a subset of the features in the search image, thereby locating the query image within the search image. Since all signatures appear similar, we use the same query image across the entire RSNA dataset (seen in Figure 3). The results of the feature matching can be seen in Figure 4. More details on the implementation of this code can be seen in Appendix G.

Figure 3. Query image used for feature matching [5]

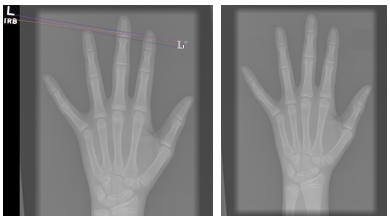


Figure 4: Top keypoint matches (left) and image after signature has been masked out (right) from *RSNA* [5]

Data Transformations

In training, validation, and the final implementation of our model, we utilized three transformations. Resizing and normalization of input images were essential to standardizing the numerical distribution of input values for any computer vision pipeline involving Convolutional Neural Networks (CNNs) [7]. Contrast enhancement was selected as it was clear through visual assessment that many images had very low contrast, making it difficult for even the human eye to distinguish bone features from the background. Increasing the contrast would enhance the disparity between the features and background of the X-ray and provide a better pixel distribution for the identification and learning of features (Figure 5).

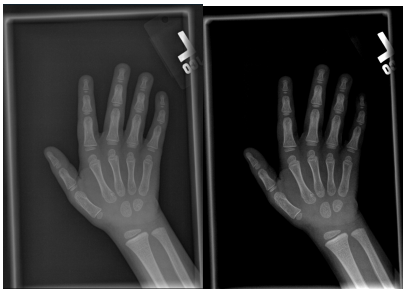


Figure 5: original X-ray (left), contrast enhanced X-ray (right) [5]

Data Augmentation

From the RSNA Pediatric Bone Age Challenge literature [5] and empirical baseline results, data augmentation proved to be a vital step in the training pipeline to promote model generalizability and reduce overfitting.

To do so, we decided to include Gaussian noise and random affine transformations as part of the data augmentation stack during training (Figure 6). The Gaussian noise augmentation adds each pixel with a Gaussian distributed random value to combat overfitting and increase the chance that models learn high-level features from pixel distribution. The random affine transformation further promotes better generalizability as it randomly rotates and translates the image to ensure the model learns features that are position invariant within the image frame. During the evaluation phase, we ensured to only use resizing, normalization, and contrast enhancement to retain the same input pixel distribution during training without adding any transformation to obscure features.

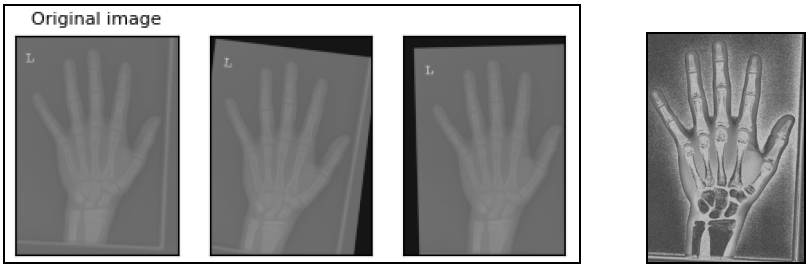


Figure 6: Original with 2 random affine transformations from the original (left). Contrast enhancement and Gaussian noise augmentations (right) [5]

Methods

Models

The main task involves image processing, which suggests the use of neural networks. The client has also asked us to implement deep learning methods due to their proven effectiveness in the RSNA challenge [5]. CNNs are preferred over simple Multi-Layer Perceptrons for their speed by reducing the number of trainable parameters.

To explore the different models, we first set a baseline with ResNet34, and then implemented two other models based on the winner of the official RSNA challenge, and the winner of the Kaggle RSNA challenge. To speed up experiments, we have used pre-trained weights wherever possible, and also set up a training pipeline outlined in Appendix B.

Baseline (ResNet34)

Our baseline model is ResNet34, which is one of the most popular and powerful models in computer vision and therefore serves well as a baseline feasibility evaluation of using neural networks for this task [8][9]. However, the original ResNet model is a classifier model with 1000 neurons in the final dense layer. To convert this into a regression model for our task, we appended another fully connected layer with a single neuron output. (See Appendix B for image)

Kaggle Winner (VGG-16 w/ Attention)

Another model that we ran experiments with is the top model on the Kaggle RSNA challenge by K Scott Mader (seen in Figure 7) [10]. In his implementation, it has achieved 14 months MAE, which is close to our client's requirements. It is based on the VGG-16 model [11], however, the three dense layers at the end of the model are removed. Additional layers are appended to form an attention mechanism, which gives relevant regions on the image to determine bone age. The attention mechanism contains more convolutional and pooling layers and ends with two dense layers before the output. The model is originally written in TensorFlow, which we have converted to PyTorch to train it in our pipeline. There were some difficulties, as some layer types such as LocallyConnected2D, were not readily available in PyTorch.

RSNA Winner (Bilbily)

We have replicated the model used by the RSNA challenge winner, Bilbily et al [5]. It is built around an InceptionV3 model, with a binary encoding of patient sex and then incorporated via a 32-neuron dense layer [12]. The sex embeddings were concatenated with the inception output, which is finally passed through two more dense layers with 1000 neurons, generating the final output. In literature, this architecture was able to achieve an MAE of 4.2 months, which is better

than the client's expectations [5]. However, the model does not have components to explain how it made its decisions. An overview of the model architecture can be seen in Figure 7.

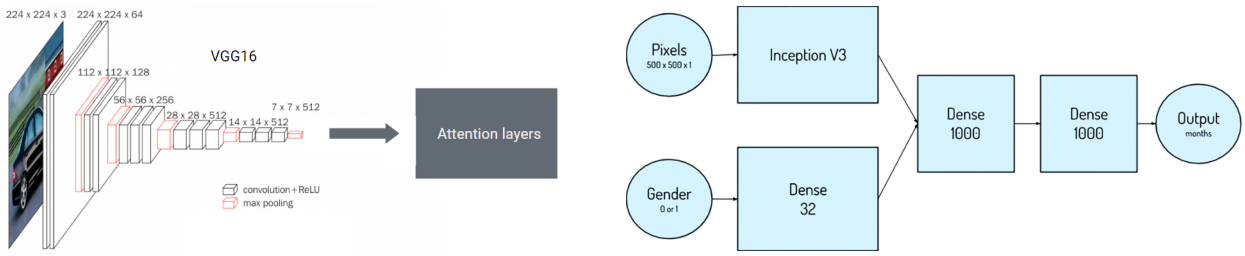


Figure 7: Mader's model based on VGG16 (left) [13], Bilbily's model based on InceptionV3 (right) [5]

Error Metrics

Two primary metrics are used to evaluate the success of our models. The first is Mean Squared Error (MSE), evaluated in each epoch during both the training and validation phases. In general, MSE loss is a straightforward metric for regression tasks to monitor the learning and optimization behavior of our models over time [7]. Additionally, this is used as the loss function for training. The second metric is Mean Absolute Error (MAE), which measures the average absolute difference between the bone age in months between the prediction and the ground truth label. This metric provides a direct means to assess our progress with respect to the ideal error margin proposed by our stakeholders.

Results

Experimental Summary

Since our baseline attempt at this regression task, we have iteratively made changes throughout the training pipeline and model architecture to enhance our prediction performance. The training plots of our best model can be seen in Figure 8. The major iterations of our results are presented below in Figure 9.

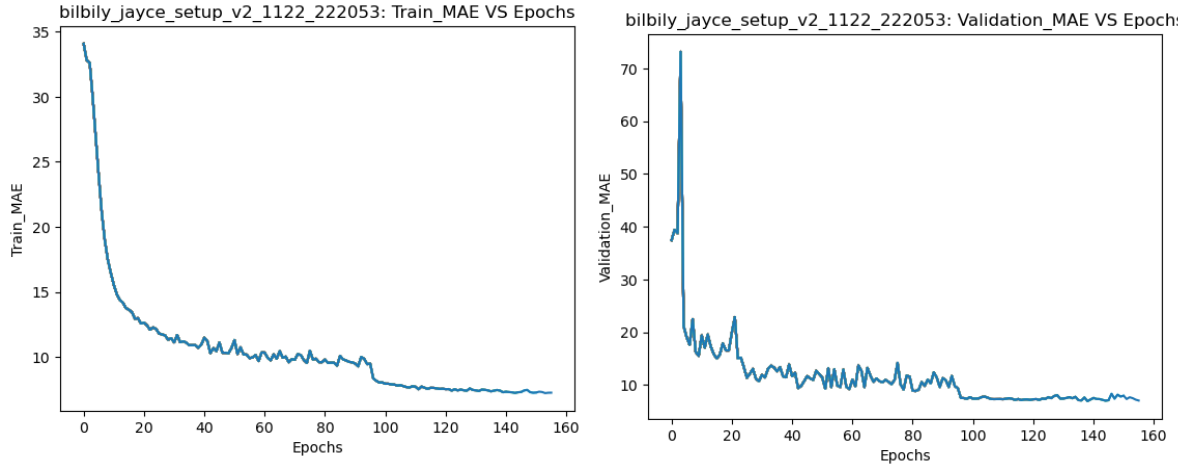


Figure 8: Progression of training and validation MAE of our best model throughout training

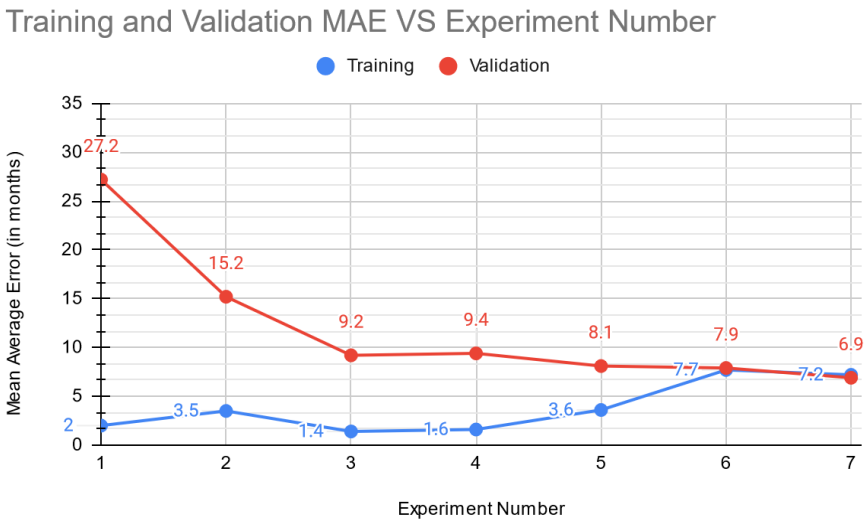


Figure 9: Progression of training results in terms of training and validation MAE across the entire timeline of all experiments conducted for this project. Data and additional training details in Appendix F.

Our baseline experiments (experiments 1, 2) used ResNet34 with weights pre-trained for ImageNet classification. We explored both pre-trained (experiment 3) and orthogonal weight initializations [14](experiments 4-7) for the InceptionV3 architecture that resides within the Bilbily model. Additionally, we explored two encoding types for sex input including one hot encoding (experiment 3,4) and a binary encoding (experiments 5-7). Moreover, training setup and hyperparameter changes were tuned for each new iteration of experiments. Most notably, the Adam [15] optimizer was used for experiments 1-5 due to its immense popularity and proven capabilities. With more research however, we discovered that the newer AdamW [16] variant claimed to offer substantially better generalization performance [16], and thus was used for our latest experiment models 6 and 7. To help optimize convergence rate, we used the ReduceLROnPlateau learning rate scheduler, which reduces the learning rate by a set factor after loss does not decrease for some specified amount of epochs [17].

From Figure 9 we can see direct improvements from our changes in model architecture and data augmentation not only in our validation MAE but also in reducing overfitting. Compared to our first iteration of experiments, our current best results reduced our validation MAE by ~75% (27.2 to 6.9 months) and reduced the gap between best training and validation MAEs since the first experiment by ~99% (25.2 to 0.3 months). This showed clear improvement in our model performance as a result of our efforts and also achieved the drastic reduction in overfitting that we had originally aimed to minimize. Our successful efforts of preventing overfitting and enhancing generalizability is further evident through the observed increase in training MAE throughout experiments 1 to 7 but juxtaposed by the consistent decline in validation MAE. A validation MAE of 6.9 months by our best model falls well below the stakeholder minimum expectation of 12 months set at the beginning of the project. All of our training plots can be observed in Appendix E.

Contrast-Based Filtering

When the final deliverable is handed off to THP, their production data will differ in quality from the RSNA dataset. Namely, the contrast of THP's images will be higher thus providing a clearer picture of the hand. To study the expected behavior on these anticipated production images, we closely analyzed the validation dataset images and related model behavior. Immediately we could see in Figure 10 that certain images in the validation dataset were inferior in quality and contrast and features are hard to see even by the human eye.

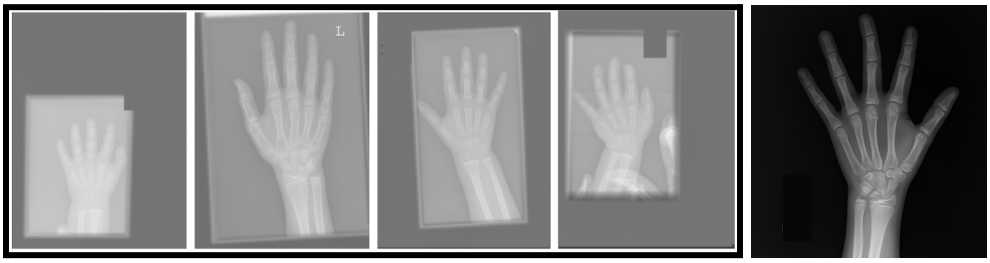


Figure 10: Four examples of low-contrast images (left) and high-quality images (right). Images are from the validation dataset [5]

As a result, we visualized histograms of our error ranges on different images in the validation set starting from the original validation set to validation sets with contrast filtering applied. Contrast filtering drops a certain percentage of the lowest contrast images during evaluation to study how much the poor validation data affected our validation MAE overall.

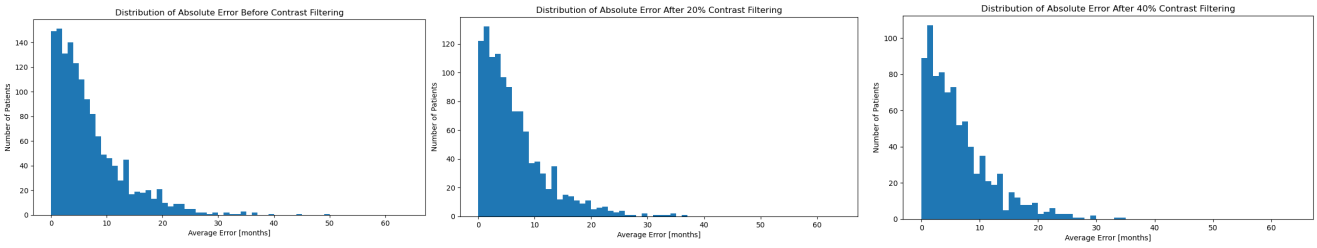


Figure 11: Histogram of validation prediction error on the original validation set (left), 20% (middle) and 40% (right) contrast filtered validation set[1].

From Figure 11, we can observe an exponential dropoff in data samples where the prediction exceeds 10 months of error in all three histograms. As a result, it is evident that we can be confident our best model will perform with less than 10 months of MAE in the majority of cases. Performing evaluation with low contrast data removed shows a slight reduction in the amount of outlier data points that result in a prediction of 30+ months although some still remain even after 40% of the lowest contrast images are removed. This could indicate further areas of improvement that could be addressed to lower overall validation MAE. Ultimately, evaluating using 20% contrast filtering produced a validation MAE of 6.4 months while using 40% produced a diminishing return of validation MAE of 6.3 months, suggesting better image contrast will not produce meaningful model prediction improvements past MAE of ~6.4 months. Analyzing performance using contrast-based filtering provides a positive insight into our model's potential results on production images knowing they will be higher in quality and contrast.

[1]y-axis scale is adjusted to show distribution shape as the amount of data used for filtered set is reduced

Implementation

Required Outputs

Explainable AI

Given the client's requirement to explain the results generated from the model, we have generated an Explainable AI (XAI) heatmap based on the concept of GRADient-weighted Class Activation Maps (GRAD-CAM) to visually identify important regions in the image where the CNN is paying attention to the most [18] when estimating the bone age. In addition to THP recommending GRAD-CAM, heatmap visualizations are a popular tool among radiologists in bone age prediction interpretation [19]. Additionally, activation maps are commonly used in image-based DL models. An example of a heatmap we generated is found in Figure 12.

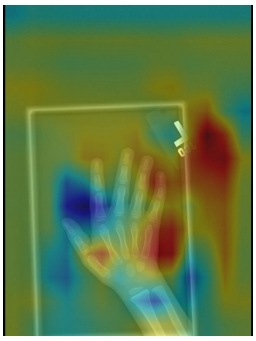


Figure 12: Heatmap generated by our XAI pipeline; underlying image taken from RSNA [5]

Growth Chart

We have additionally included a growth chart in the output, using the mean (green line) and 95th confidence intervals (blue lines) of bone age, taken from *Skeletal Development of the Hand and Wrist: A Radiographic Atlas and Digital Bone Age Companion* [21]. As seen in Figure 13, chronological age is plotted against bone age, and includes the model's prediction (red dot). Using the model's prediction, radiologists can use this graph to determine if a patient's predicted bone age is outside the range of healthy bone age for a patient's chronological age. Both the bone age and chronological age axes are in years, as typical growth charts for bone age have measurements in years.

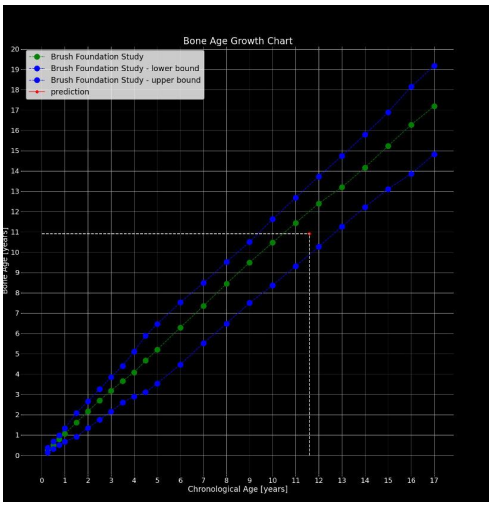


Figure 13: Bone age growth chart for a male patient including model prediction and values from *Skeletal Development of the Hand and Wrist* [21]

Atlas

As previously mentioned in the **Background**, radiologists use an atlas of hand X-rays to compare a patient's X-ray to estimate bone age. The atlas contains images corresponding to a broad range of bone ages, organized by sex. Our client has provided us with images from *Hand Bone Age: A Digital Atlas of Skeletal Maturity* [20]. The burned-in bone age signature has been modified to have a year, month format as requested by THP.

As requested by THP, we have rendered a compilation that includes the lower and higher bone age estimates from the atlas alongside the input image. We additionally burned the model's prediction into the input image. This will allow the radiologist to have a side-by-side comparison to validate the model's prediction with images in the atlas. An example can be seen in Figure 14.

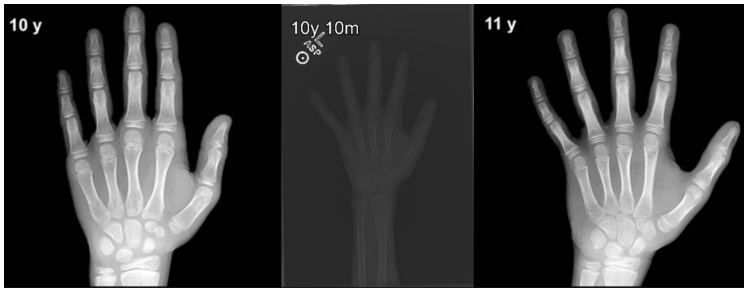


Figure 14: Sample output atlas compilation containing 2 closest atlas images from *Hand Bone Age: A Digital Atlas of Skeletal Maturity* [20] and patient image from *RSNA* [5] with model prediction

Integrated Application

In order to expose our model and associated outputs described previously in a useful manner to radiologists, THP has requested that we implement our solution as a DICOM node. We have created an open-source, fully-integrated application that receives inputs, performs inference with our model, generates results/outputs, and provides the outputs.

DICOM is a communications standard that has become ubiquitous in medical imaging. It defines an interface that allows for any DICOM-supporting image acquisition equipment (computed radiography in our case) to interact with DICOM-supporting computer systems [22].

Figure 15 depicts our application in the overall pipeline at THP. Our application acts as both a DICOM Service Class Provider (SCP) and a Service Class User (SCU), which means we both receive (as an SCP) and transmit (as an SCU) images. In the application, the following occurs:

1. Received data is sent to the predictor. Workers pick up tasks off the predictor work queue.[2]
2. Workers perform inference and generate the three required outputs. Outputs are sent to the predictor's completed queue.
3. Predictor removes items from the completed queue, and sends them via the SCU.

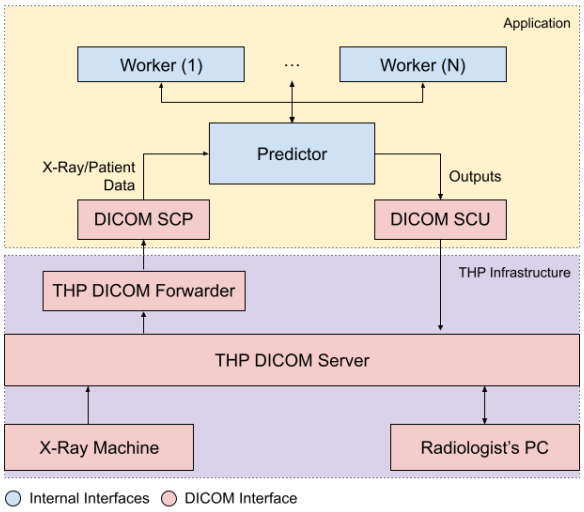


Figure 15: A Typical Deployment of the Application

Discussion

Bone age estimation is an important metric to determine healthy development in children, but is also a time-intensive process. In addition to the model, the visuals – XAI heatmap, atlas, and growth chart – we have developed are also useful, as they provide easy visualizations to give insight into the model's decision and compare our prediction with existing images and bone ages. Throughout the design, all data used was a publicly accessible anonymized dataset with an even split in patient sex [5]. However, radiologist labels are based on an atlas composed of only Caucasian children, meaning diversity in race is not accounted for [23][24]. While our model

[2] Multiple workers are used to take advantage of multi-threaded hardware

met the initial requirements set by THP and yielded reliable results, we have some limitations due to the nature of a regression model. We do not have confidence scores for the predicted age, nor a way to generate a ranked list of possible bone ages. This also made it difficult to implement a visual activation heatmap for XAI. GRAD-CAM is based on classification models, as it uses the gradients of target classes to produce heatmaps. However, our task is a regression task and thus we effectively have only one class. This may have played a large role in the uninterpretable result, as previously seen in Figure 12.

Currently in this field, *Physis* from 16Bit AI also uses a DL model and hand X-rays to automate bone age measuring [25]. They base their tool on the *Gruelich and Pyle* atlas, and their initial prototype won the 2017 RSNA challenge with an MAE of 4.3 months, but is a paid service. Similar to *Physis,* our project displays the closest matching atlas images and the growth chart to assess potential issues in the patient's bone development. However, our solution is an open-source alternative with similar performance levels and offers the same depth of information in its output. Additionally, our solution integrates into hospital infrastructure with DICOM, while *Physis* does not appear to have such integration. Furthermore, our solution does not require licensing as it is designed to be self-hosted, but *Physis* must be licensed by a health authority such as Health Canada [25] as it is a third-party provider handling health data. This means our solution is a free alternative that is deployable independent of geographic location.

Conclusions and Future Directions

The goal of our project was to produce a DL solution that would estimate pediatric bone age from a given hand X-ray and additionally includes explainable AI components and visualizations to aid radiologists in understanding and verifying model decisions. The model from Bilbily et al. was chosen out of 3 models based on its performance in estimating bone age, with a validation MAE of 6.3. Our end product integrates the data transformations, model, visual generation, and outputs into an application that can be deployed in any typical hospital system. Overall, our project met client expectations, with the exception of XAI.

In the future, improvements to the XAI component could be done by changing our model to a classification task by binning ages based on the bone ages in the atlas THP has provided us. Going forward, THP can re-train our model using THP patient data, as their X-rays would be of higher quality, and could also integrate more patient attributes that the student team was not privy to, such as other existing health conditions that may affect growth. For the student team, THP has suggested our work be submitted to a conference in radiology informatics.

Appendices

A. Attribution Tables

Project Contribution	
Anton	• Implemented and trained Kaggle VGG16 model ○ Explored other deep learning models • Improved on training framework for easy model switching and experimenting • Merged data manipulation methods into training pipeline
Benjamin	• Implemented RSNA dataset loaders • Implemented and trained the first version of the Bilbily model • Improved training framework with ability to vary model parameters in yaml • Designed and implemented DICOM pipeline • Assisted in debugging explainable AI code
Jayce	• Designed and implemented main training and evaluation framework code including main training loop, hyperparameter setup, saving of training checkpoints, training plots, model retrieval class, evaluation script • Implemented definable data augmentation function stack • Designed, explored, and tuned setup and training of result models for iterative improvement
Katarina	• Wrote initial code to convert DICOM to PNG and extract sex • Developed code to execute explainable AI ○ Researched explainable AI methods • Wrote code to generate atlas compilation • Wrote code to generate growth chart

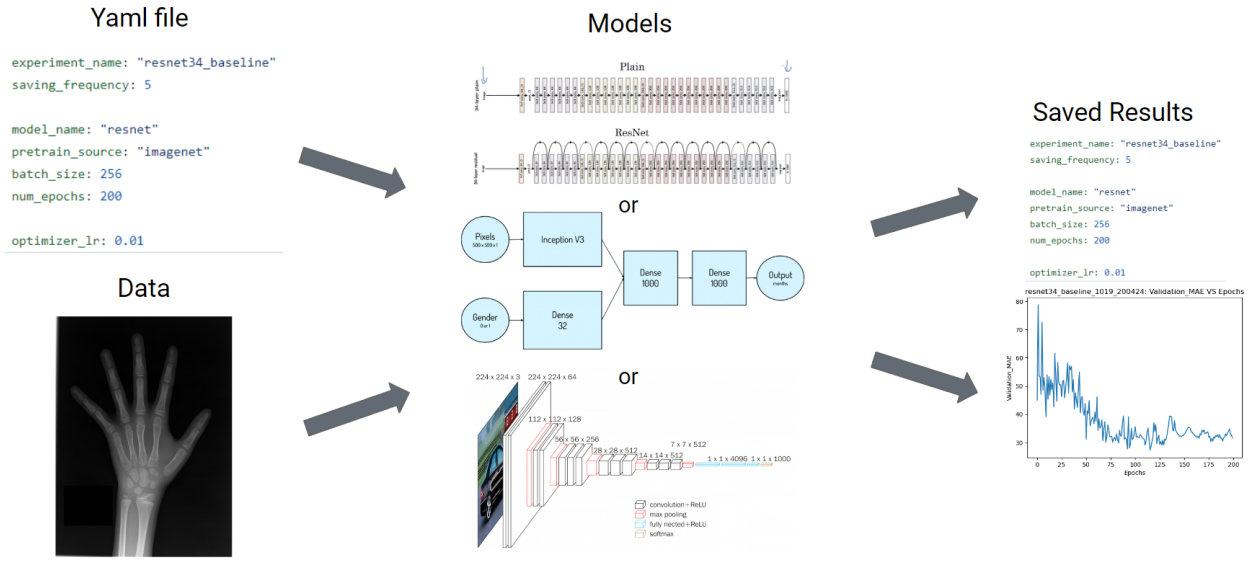
Richard	• Wrote prototype code to perform DICOM queries to the Orthanc Server
	  ○ Modified the Orthanc server configuration to allow DICOM requests
	• Designed signature removal tools and implemented it across the RSNA dataset
	  ○ Ensured code was executable as a script for simplicity
	• Worked on code to execute explainable AI

Report Contribution	
Anton	• Methods (Models) • Conclusions
Benjamin	• Data • Implementation (DICOM)
Jayce	• Data Augmentation/Transformation under Data Manipulation • Results

Katarina	• Introduction • Implementation (Deployment Pipeline) • Discussion • Conclusions and Future Directions
Richard	• Introduction • Data Preprocessing (Signature Removal) • Discussion

B. Training Pipeline & Model Diagram

We have designed a training pipeline that lets us experiment easily. This is done by using a yaml file that the code reads the hyperparameters from, which contains the model name, batch size, learning rate, etc. After each experiment, the results are saved along with a copy of the yaml file.



Figures taken from [5], [26], [27]

ResNet34 Model architecture

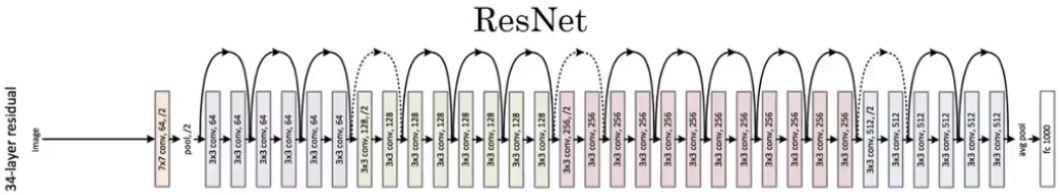
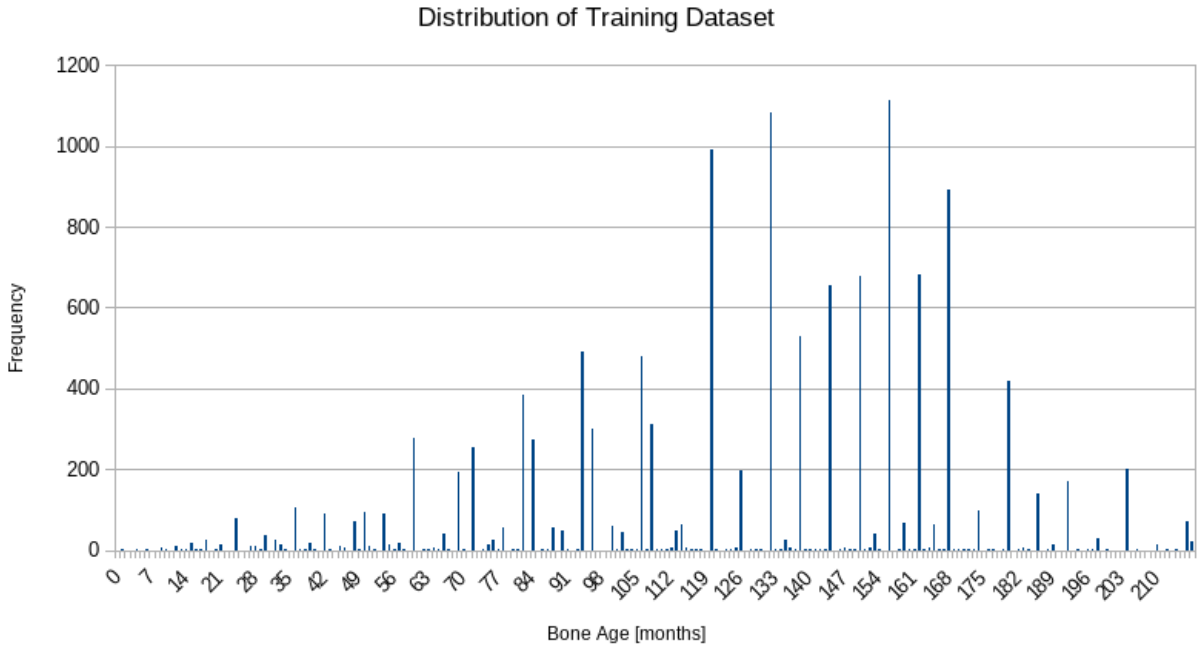


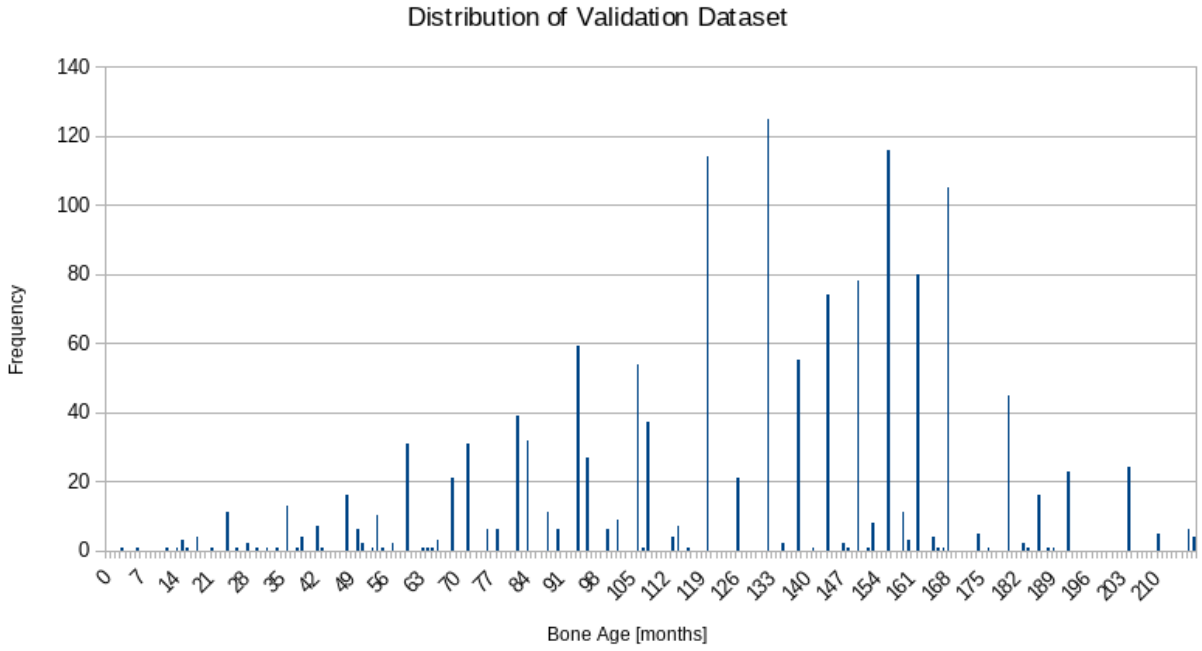
Figure taken from [26]

C. Source Code Repository

Github Repo: https://github.com/lolzballs/mie429

D. Dataset Distribution

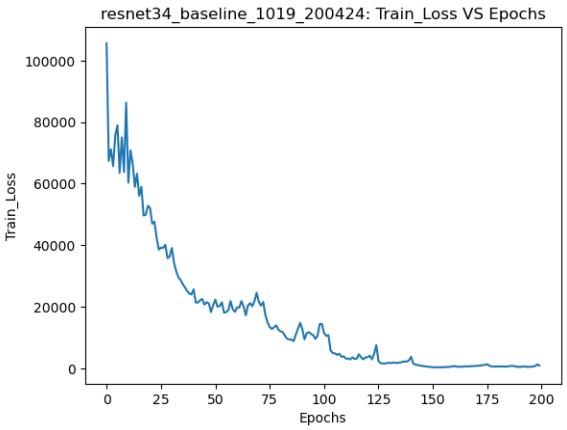


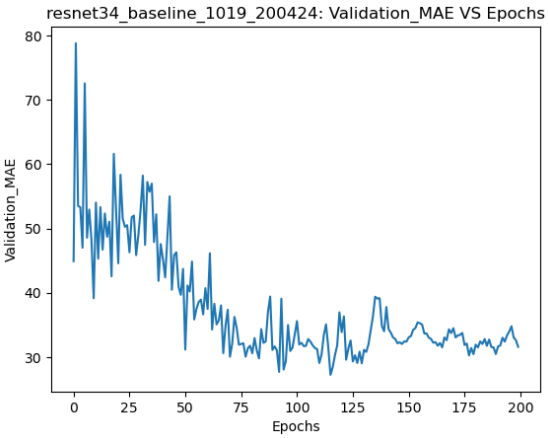


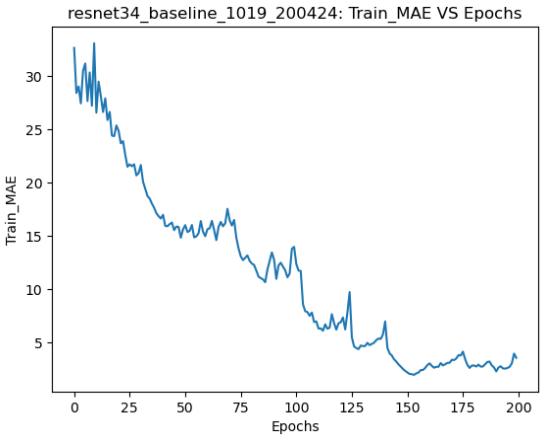
E. Model Training Plots

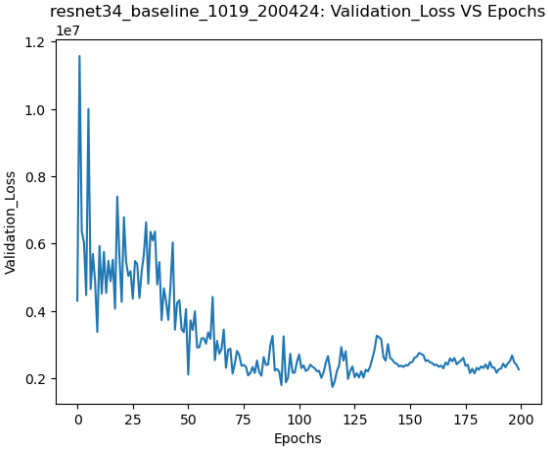
Below are all the plots produced during training tracking training MSE loss, validation MSE loss, training MAE, and validation MAE for each iteration of our models. Refer to Table X above for corresponding training details for each iteration number.

Iteration 1

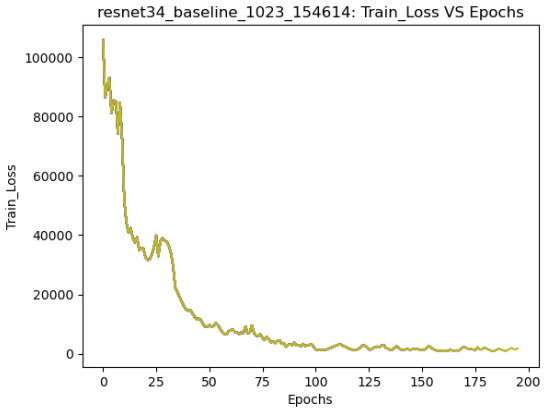


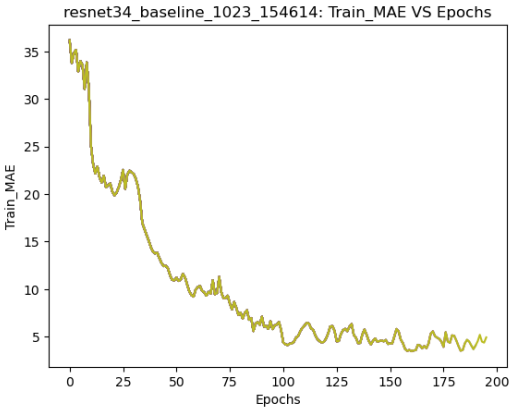


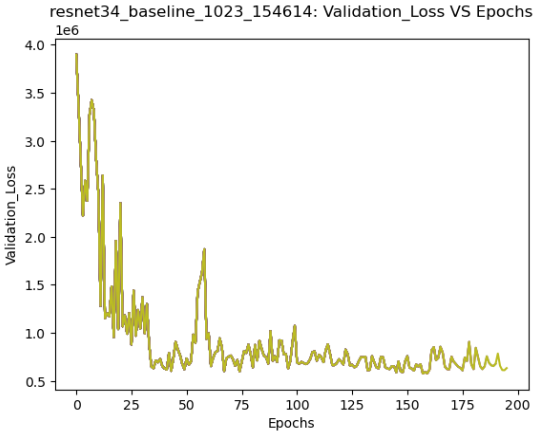


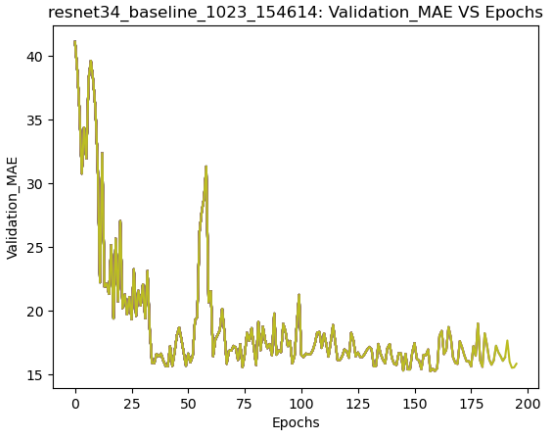


Iteration 2

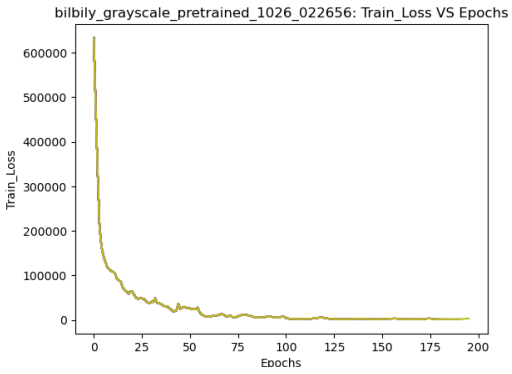


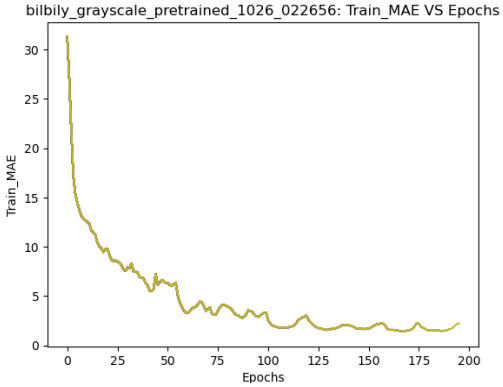


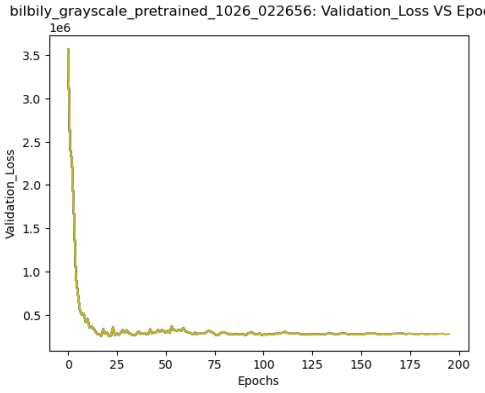

resnet34_baseline_1023_154614: Validation_Loss VS Epochs
1e6
Validation_Loss
Epochs

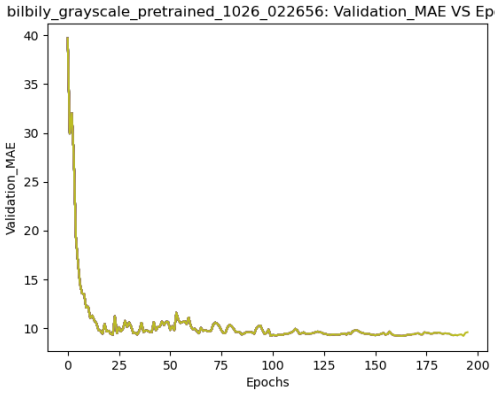

resnet34_baseline_1023_154614: Validation_MAE VS Epochs
Validation_MAE
Epochs

Iteration 3

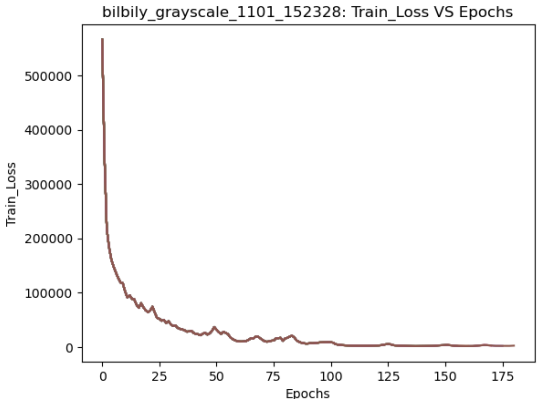

bilbily_grayscale_pretrained_1026_022656: Train_Loss VS Epochs
Train_Loss
Epochs

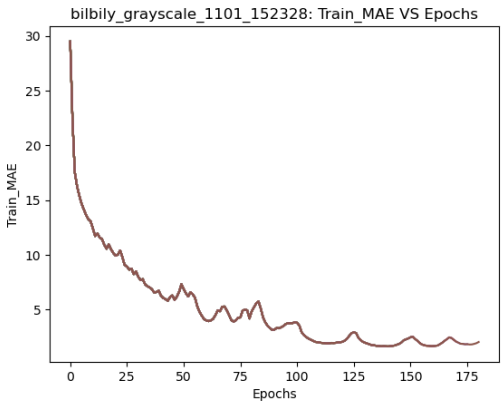

bilbily_grayscale_pretrained_1026_022656: Train_MAE VS Epochs
Train_MAE
Epochs

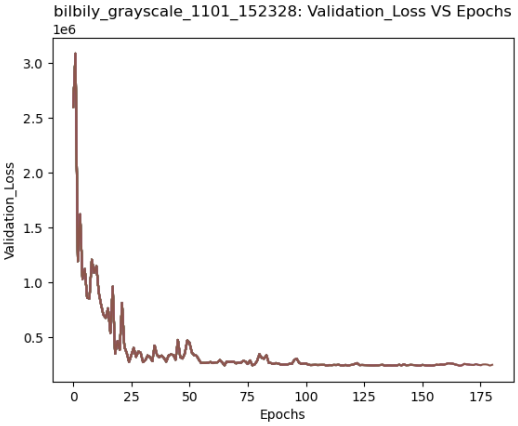

bilbily_grayscale_pretrained_1026_022656: Validation_Loss VS Epochs
1e6
Validation_Loss
Epochs

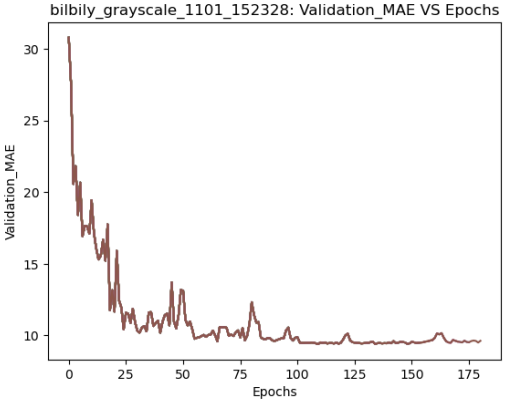

bilbily_grayscale_pretrained_1026_022656: Validation_MAE VS Epochs
Validation_MAE
Epochs

Iteration 4

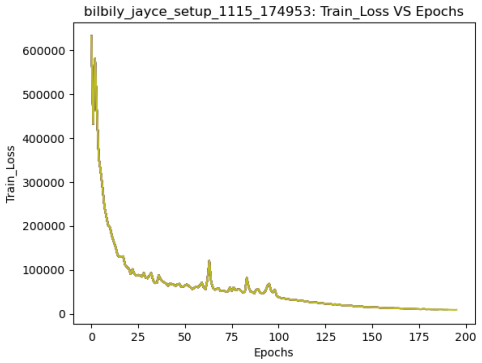


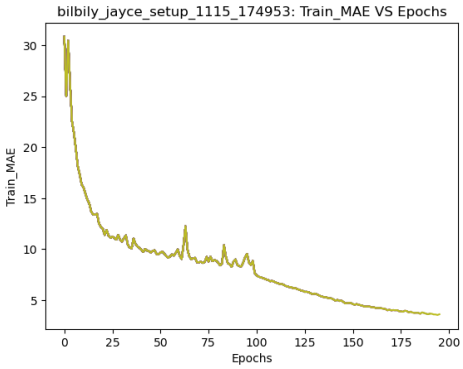


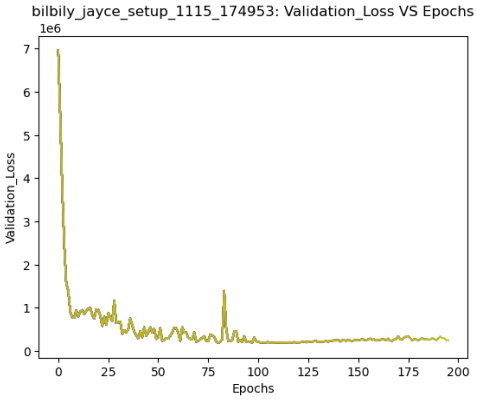


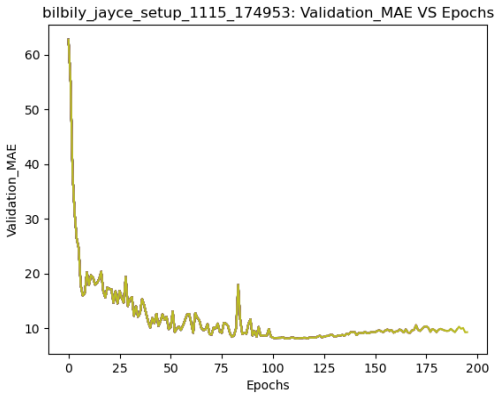


Iteration 5

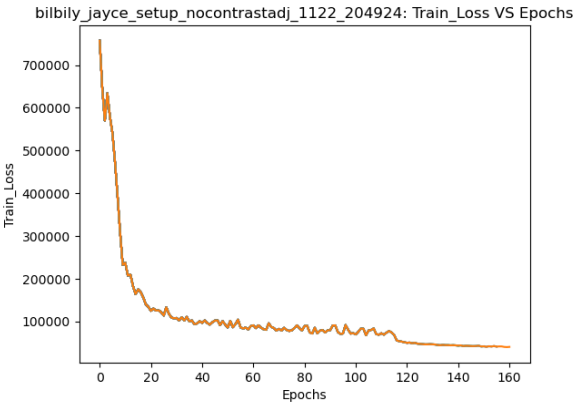


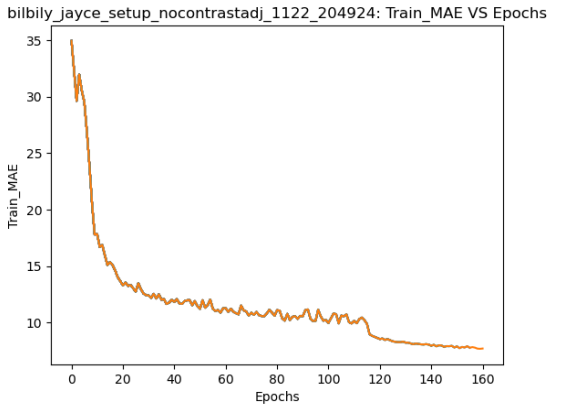


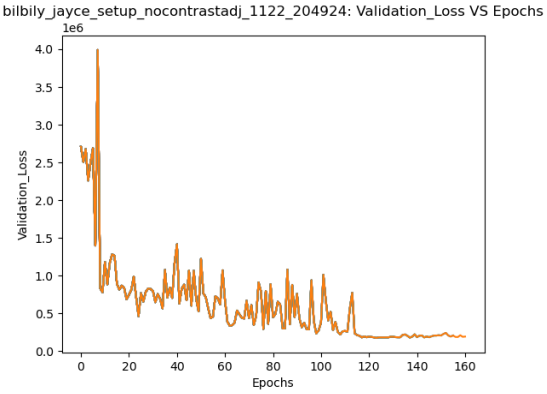


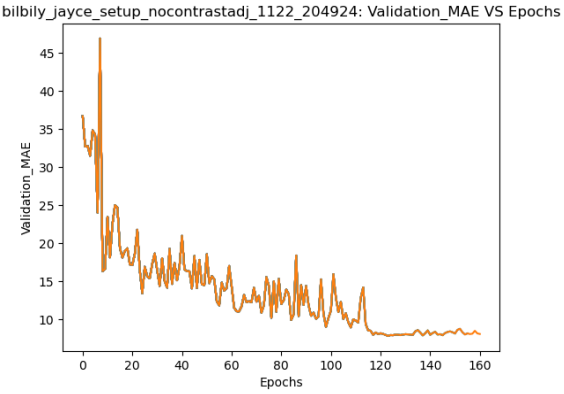


Iteration 6

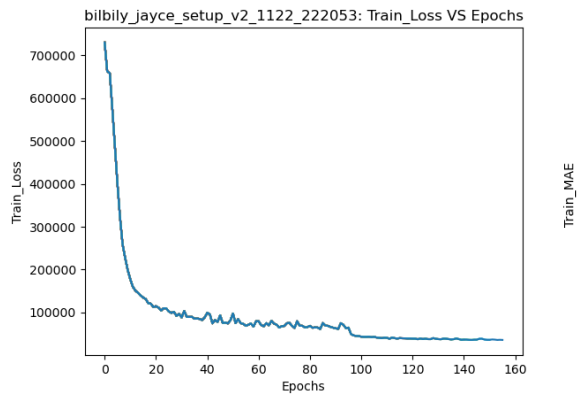


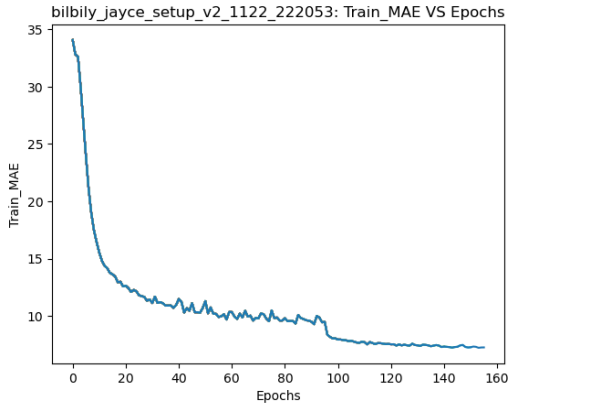


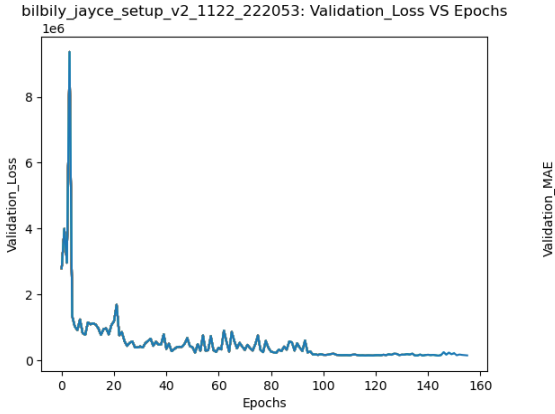


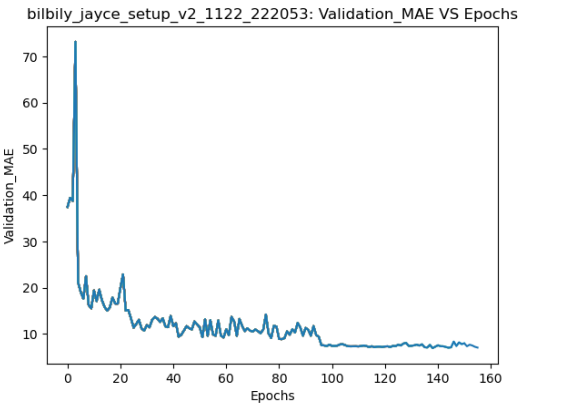


Iteration 7









F. Model Training Results Table

Iteration	Model Name	Data Manipulations Applied	Training MAE (months)	Validation MAE (months)
1	Resnet34 Pretrained Baseline	- Resized to 224x224x3	2.0	27.2
2	Resnet34 Pretrained Baseline	- Resized to 224x224x3 - Normalized	3.5	15.2
3	Bilbily Pretrained OneHot Encoding	- Resized to 512x512x1 - Normalized	1.4	9.2
4	Bilbily OneHot Encoding	- Resized to 512x512x1 - Normalized	1.6	9.4
5	Bilbily Binary Encoding	- Signature Removed - Resized to 512x512x1 - Contrast Enhanced - Normalized - Gaussian Noise	3.6	8.1
6	Bilbily Binary Encoding	- Signature Removed - Resized to 512x512x1 - Normalized - Gaussian Noise - Random Affine	7.7	7.9
7	Bilbily Binary Encoding	- Signature Removed - Resized to 512x512x1 - Contrast Enhanced - Normalized - Gaussian Noise - Random Affine	7.2	6.9

G. Feature Matching Implementation Details

Features, in the form of keypoints are identified in both the query and the search image. We used OpenCV's ORB (Oriented FAST and Rotated BRIEF) algorithm to generate these features. Features are generated on rescaled versions of the search and query image to introduce scale invariability when searching for our query image, helping to detect the query image even if the signature in the search image is larger or smaller in relation. Similar feature vectors within the two images are then matched together and used to locate the signature within the search image. The signature is then covered up by a bounding box rectangle and filled according to the search image's background color.

References

[1] "What is Bone Age?", *BoneXpert*. [Online]. Available: https://bonexpert.com/what-is-bone-age/. [Accessed: 24- Sep- 2022].

[2] L. Hirsch, "X-Ray Exam: Bone Age Study (for Parents) - Nemours KidsHealth", *KidsHealth*. [Online]. Available: https://kidshealth.org/en/parents/xray-bone-age.html#:~:text=The%20bone%20age%20study%20can,by%20pediatricians%20or%20pediatric%20endocrinologists. [Accessed: 24- Sep- 2022].

[3] S. Iuliano-Burns, J. Hopper, and E. Seeman, "The age of puberty determines sexual dimorphism in bone structure: A male/female co-twin control study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 5, pp. 1638–1643, 2009.

[4] R. R. van Rijn and H. H. Thodberg, "Bone age assessment: Automated techniques coming of age?," *Acta Radiologica*, vol. 54, no. 9, pp. 1024–1029, Nov. 2013.

[5] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, "The RSNA pediatric bone age machine learning challenge," Radiology, vol. 290, no. 2, pp. 498–503, 2019.

[6] W. D. Bidgood, S. C. Horii, F. W. Prior, and D. E. Van Syckle, "Understanding and using DICOM, the data interchange standard for Biomedical Imaging," Journal of the American Medical Informatics Association, vol. 4, no. 3, pp. 199–212, 1997.

[7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of Convolutional Neural Networks: Analysis, applications, and prospects," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999–7019, 2022.

[8] A. Joseph, "Unlocking resnets," Medium, 10-Nov-2021. [Online]. Available: https://medium.com/analytics-vidhya/opening-resnets-46bb28f43b25. [Accessed: 29-Oct-2022].

[9] "Papers with code - resnet explained," ResNet Explained | Papers With Code. [Online]. Available: https://paperswithcode.com/method/resnet?fbclid=IwAR17bSW_9yKxeb8MPZq_KvH-_Y5DDG-HoUnqeA8Zmy_B3A8G6rsj98GtyVY. [Accessed: 01-Dec-2022].

[10] K. S. Mader, "Attention on Pretrained-VGG16 for Bone Age," Kaggle, 11-Apr-2018. [Online]. Available: https://www.kaggle.com/code/kmader/attention-on-pretrained-vgg16-for-bone-age. [Accessed: 30-Nov-2022].

[11] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," arXiv, 10-Apr-2015. [Online]. Available: https://arxiv.org/pdf/1409.1556.pdf. [Accessed: 30-Nov-2022].

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, “Rethinking the Inception Architecture for Computer Vision,” arXiv, 11-Dec-2015. [Online]. Available: https://arxiv.org/pdf/1512.00567.pdf. [Accessed: 30-Nov-2022].

[13] K. Le, “An overview of VGG16 and NiN models,” Medium, 25-Mar-2021. [Online]. Available: https://medium.com/mlearning-ai/an-overview-of-vgg16-and-nin-models-96e4bf398484. [Accessed: 22-Nov-2022].

[14] W. Hu, L. Xiao, and J. Pennington, “Provable benefit of orthogonal initialization in Optimizing Deep Linear Networks,” arXiv.org, 16-Jan-2020. [Online]. Available: https://arxiv.org/abs/2001.05992. [Accessed: 01-Dec-2022].

[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv.org, 30-Jan-2017. [Online]. Available: https://arxiv.org/abs/1412.6980. [Accessed: 01-Dec-2022].

[16] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv.org, 04-Jan-2019. [Online]. Available: https://arxiv.org/abs/1711.05101. [Accessed: 01-Dec-2022].

[17] “Reducelronplateau,” ReduceLROnPlateau - PyTorch 1.13 documentation. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html. [Accessed: 01-Dec-2022].

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2019.

[19] F. C. Kitamura and I. Pan, “Artificial Intelligence class activation mapping of Bone age,” *Radiology*, vol. 303, no. 1, pp. 52–53, Apr. 2022.

[20] V. Gilsanz and O. Ratib, “Reference Images,” in *Hand bone age: A digital atlas of skeletal maturity*, Berlin: Springer, 2005, pp. 41–101.

[21] C. M. Gaskin, S. L. Kahn, J. C. Bertozzi, and P. M. Bunch, “Tables,” in *Skeletal development of the hand and wrist a radiographic atlas and Digital Bone Age companion*, New York, New York: Oxford University Press, 2011, pp. 8–9.

[22] W. D. Bidgood, S. C. Horii, F. W. Prior, and D. E. Van Syckle, “Understanding and using DICOM, the data interchange standard for Biomedical Imaging,” Journal of the American Medical Informatics Association, vol. 4, no. 3, pp. 199–212, 1997.

[23] C. M. Gaskin, S. L. Kahn, J. C. Bertozzi, and P. M. Bunch, “Background,” in *Skeletal development of the hand and wrist a radiographic atlas and Digital Bone Age companion*, New York, New York: Oxford University Press, 2011, p. 1.

[24] A. Zhang, J. W. Sayre, L. Vachon, B. J. Liu, and H. K. Huang, “Racial differences in growth patterns of children assessed on the basis of bone age,” *Radiology*, vol. 250, no. 1, pp. 228–235, Jan. 2009.

[25] “Physis,” *16bit*. [Online]. Available: https://www.16bit.ai/physis. [Accessed: 01-Dec-2022].

[26] K. He, X. Zhang, S. Ren, and J. Sun, *Example network architectures for ImageNe*. arXiv, 2015.

[27] M. ul Hassan, *VGG16 architecture*. Neurohive, 2018.