Advancing Handbone Maturity Estimation: Comparative Analysis and Novel Approaches Using CNNs and Decision Fusion

Berat Furkan Kocak †, Onur Bacaksiz‡

***Abstract*—Convolutional neural network (CNN) models have demonstrated the ability to estimate skeletal maturity from hand radiographs with accuracy comparable to that of expert radiologists and state-of-the-art automated models. This field has gained attention due to its potential for improving diagnostic accuracy in pediatric radiology, as exemplified by the RSNA Pediatric Bone Age Machine Learning Challenge, which highlighted the value of coordinated approaches in medical imaging.**

Despite significant progress, challenges remain in transitioning from image classification to regression tasks while achieving expert-level prediction quality. Additionally, the performance and efficiency of state-of-the-art architectures require further evaluation, especially concerning the role of external and domain-specific knowledge in improving prediction accuracy.

Building on the RSNA challenge's best model, our research advances the state of the art by experimenting with novel approaches, including implementing a decision fusion mechanism using carpal bone regions in order to improve prediction accuracy, creating a dynamic random augmentation method to address data imbalance, implementing and comparing multiple models (Inception-v4, CNN with CBAM attention, and ResNet50 with transfer learning), and assessing the importance of gender information in model performance.

Our goal is to provide a comprehensive comparative analysis of existing and novel approaches, offering future researchers insights for sound analysis and implementation.

This research also highlights the impact of data balancing through augmentation on model performance and demonstrates how decision-fusion mechanisms, using carpal bone regions, further enhance the precision of the state-of-the-art models.

***Index Terms*—Decision Fusion, Convolutional Neural Networks, Attention, Inception Network, Image Regression**

I. Introduction

Automated skeletal maturity estimation using hand radiographs has emerged as a pivotal topic in medical imaging, particularly in pediatric radiology, due to its potential to enhance diagnostic accuracy. The advent of convolutional neural networks (CNNs) has transformed this field, enabling models to achieve performance comparable to expert radiologists. Building on this foundation, our research integrates several state-of-the-art approaches from similar domains to explore their impact on novel methods for skeletal maturity assessment. Specifically, we compare the performance and efficiency of existing architectures, examine the role of domain-specific knowledge (carpal bones and gender information), and propose improvements in age prediction using hand radiographs.

A critical challenge in this domain is the high computational cost and prolonged training times associated with deep image processing architectures. Our study aims to mitigate these issues by leveraging simpler architectures enriched with external domain knowledge, achieving comparable accuracy. A key innovation in our work is the implementation of a decision fusion mechanism that incorporates carpal bone information to enhance model performance.

Our proposal is highly relevant to the current state-of-art, especially to inspire the medical professionals how their domain knowledge can improve the accuracy of even the simplest models comparable to the deep learning models especially when limited by the hardware constraints. We also provide a comparative analysis of different models specific to this domain, mitigating this effort for new researchers for them to have a fast start point aligning with their goals.

- **Integration of State-of-the-Art Approaches** We provide a comprehensive evaluation of state-of-the-art architectures, including Inception-v4, CNN with CBAM attention, and ResNet50 with transfer learning, highlighting their performance and efficiency in skeletal maturity estimation from hand radiographs.
- **Incorporation of Domain Knowledge** Leveraging insights from carpal bone regions and gender information, we implement a novel decision fusion mechanism that significantly improves model accuracy for age prediction, particularly addressing gaps in prior methodologies.
- **Custom Dynamic Augmentation for Data Balancing** To address dataset imbalance, we propose and implement a custom dynamic augmentation method that enhances model training stability and generalization, inspired by proven augmentation strategies in related research domains.
- **Relevance to Resource-Constrained Settings** By demonstrating how external domain knowledge can empower simpler architectures to achieve accuracy comparable to complex deep learning models, we propose solutions particularly beneficial for environments with limited hardware resources.

This paper is structured as follows. In Section II we describe the state of the art, the system and data models are respectively presented in Sections III and IV. The proposed signal process-

† Department of Mathematics, University of Padova, email: beratfurkan.kocak@studenti.unipd.it

‡ Department of Mathematics, University of Padova, email: onur.bacaksiz@studenti.unipd.it

ing technique is detailed in Section V and its performance evaluation is carried out in Section VI. Concluding remarks are provided in Section VII.

II. Related Work

This research is inspired by several foundational studies. For example, The paper *Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs* [1] demonstrated the efficacy of deep learning for bone age estimation, achieving an RMS error of 0.67 years using a ResNet50 architecture, stating a better result compared to the mean of human professional reviewers 0.82, showcasing deep learning capabilities. However, the model required 6 - 8 hours of training on a single K80 GPU using transfer learning. Training such a deep network without pre-trained weights would result in prohibitive training times, particularly on hardware with limited capabilities. Building on these insights, our research explores attention mechanisms and decision fusion to achieve comparable accuracy with lighter and more efficient models.

Another paper *The RSNA Pediatric Bone Age Challenge* [2] highlighted the utility of gender information in model architectures. Though the paper showcased several approaches to achieve a good accuracy in prediction, we see the need of updating the winning model's usage of Inception-v3 to Inception-v4.

Data balancing through augmentation has been previously shown to improve model generalization (Hamida et al.,2023) [3]. However, we couldn't find a study explicitly stating an augmentation mechanism to handle data imbalance in the dataset we are working on. Hence, we developed a custom dynamic augmentation method to uniformize our dataset distribution, improving training stability.

We also drew inspiration from augmentation strategies discussed in previous studies such as *Deep Neural Networks for Chronological Age Estimation From OPG Images* [4], to conduct an extensive grid search to identify optimal parameters for our dataset. Additionally, the paper *Automated age estimation from MRI volumes of the hand* [5] emphasized the critical role of carpal bones in age detection, particularly for individuals aged between 13 to 25. Even though this paper utilizes the carpal bone knowledge, it did not incorporate it into a decision fusion framework with lightweight CNN models on grayscale images. This limitation inspired our novel approach, which integrates carpal bone information through decision fusion, resulting in enhanced accuracy and computational efficiency.

III. Processing Pipeline

Our work is clearly structured in two Python notebooks. The first notebook, `DatasetPreprocessing`, focuses on data exploration and preprocessing of the zipped data obtained online, using the methodologies mentioned in part 4. The operations performed in the data processing notebook save three folders on the drive, namely `raw`, `clahe`, and `clahe_masked`, containing the train, validation, and test datasets. These datasets are obtained from the following: no CLAHE and mask applied, only CLAHE applied, and CLAHE + mask applied.

The main notebook, `Models_and_Training`, handles the creation of a TensorFlow dataset using the obtained PNG images, combined with ground truth and sex information obtained during preprocessing. This is done using the `from_tensor_slices` operation. TensorFlow datasets are useful because the amount of training data exceeds our hardware limits when it comes to loading the dataset into RAM for training. This requires an efficient batch approach, which is well-handled by TensorFlow datasets. We set a batch size of 32. It also enables us to use prefetching for increased performance and to shuffle the dataset easily. We structured the creation of the TensorFlow dataset in such a way that it first applies the necessary preprocessing, followed by dynamic augmentation logic (such as random flips) based on the training dataset distribution, as detailed in part 4.

We created seven different models in order to obtain a detailed comparative analysis of the results from each architecture, considering both performance and efficiency.

These models are as follows:

1) **Baseline CNN**: A CNN with 4 layers, combined with gender information through dense layers.
2) **InceptionV4**: An InceptionV4 block combined with gender information through dense layers.
3) **Attention-CNN**: Utilizes Convolutional Block Attention Modules (CBAM) as an additional block in the baseline architecture. This architecture enables both spatial and channel attentions.
4) **Decision Fusion Mechanism using Carpal Bones**: Combines the regression result from a model trained on full images with the baseline cnn architecture trained only on the carpal bone region of the hand. The best combination weights are obtained through grid search.
5) **ResNet50 using Transfer Learning**: This pre-trained model is used solely for performance comparison with other models. The last layer of the model is extracted, and a dense layer is incorporated to perform regression. This model does not incorporate gender information.
6) **ResNet50 combined with gender using Transfer Learning**: Similar to model 5, but with the addition of gender information to assess its impact on prediction accuracy.
7) **ResNet50 Combined with gender using Transfer Learning + Trainable First Layers**: Following common practices, given that we are dealing with a relatively small and non-ImageNet-like dataset, we made the first 10 layers of ResNet50 trainable to assess their effect on model performance.

For each model, we used root mean squared error (RMSE), which provides interpretable loss values as they directly relate to actual regression errors.

After these models are defined, separate training is performed for each model, utilizing the same training-validation datasets.

We used a maximum of 20 epochs, with an early stopping mechanism based on validation loss, with a patience of 3 epochs.

We recorded the time taken for training each model, their training loss, training mean absolute error (MAE), validation loss, and validation MAE for each epoch.

After training concludes, we saved the model parameters to a `.keras` file for later use in inference and performance statistics.

Test statistics are calculated on a separate test dataset containing 200 images. The Mean Absolute Error (MAE) for each model is calculated on this dataset. These numbers are then re-scaled to human-interpretable values representing ages in months.

Finally, a comparative table is generated for each model, including their inference time, inference accuracy, training and validation accuracies, training times, number of model parameters, and the size of the model.

This concludes our processing pipeline and clearly showcases the results of our work. Fig. 1 visualizes this process.

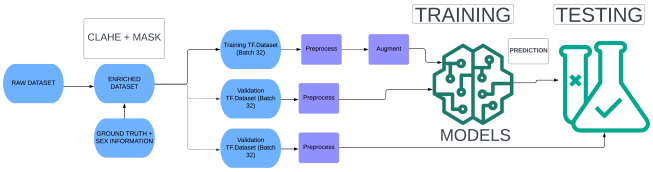

Fig. 1: Project Pipeline

IV. Signals and Features

A. Dataset

The study utilized a dataset of 14,036 clinical radiographs of the left hand from two institutions: Lucile Packard Children's Hospital at Stanford University (2,983 images) and Children's Hospital Colorado (11,053 images), for bone age assessment [1]. The data was divided as shown in Table 1:

Dataset	Male	Female
Train	6833	5778
Validation	773	652
Test	100	100

TABLE 1: Dataset-Gender Counts

The paper by Halabi et. al. [2] mentions the methodology of obtained ground truth labels as follows: All images were labeled with skeletal age estimates and sex based on clinical radiology reports at the time of imaging. The ground truth skeletal age estimates for all datasets were determined using six independent reviews per image, including:

- Clinical radiology reports.
- Four independent pediatric radiologist reviews (two per institution).
- A second review by one of the radiologists after one year.

The *Greulich and Pyle standard* was used for bone age estimation. The final ground truth for the test set was calculated in two steps:

1) A preliminary ground truth was obtained as the simple mean of the six estimates.
2) Reviewer estimates were corrected for bias, weighted by the inverse of their mean absolute difference (MAD), and combined to produce a weighted mean.

The weighted mean formed the final ground truth, providing a robust foundation for evaluating algorithm performance.

B. Exploratory Data Analysis (EDA)

In order to understand the underlining structure of the dataset, we implemented an exploratory data analysis.

1) Fig. 2 shows that the image dimensions across the dataset are not standardized. Standardizing the image dimensions is important, as most image processing models require consistent input layer dimensions. Another important point is that high-resolution images occupy more memory, especially during augmentation operations. Based on this information, we decided to standardize the images to 256x256 in the upcoming pre-processing stage.

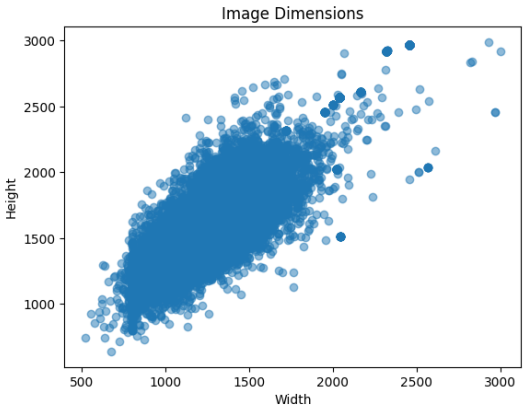

Fig. 2: Image Dimensions of the Training Dataset

2) Another analysis is performed to ensure that all images are grayscale, and this was confirmed for the dataset at hand.

3) Fig. 3 shows the mean image pixel intensities for each image in the training dataset to identify the presence of highly bright or dark images, which may potentially affect the model's learning process. To address this issue, we decided to utilize a contrast-enhancing method, such as CLAHE, in the pre-processing step.

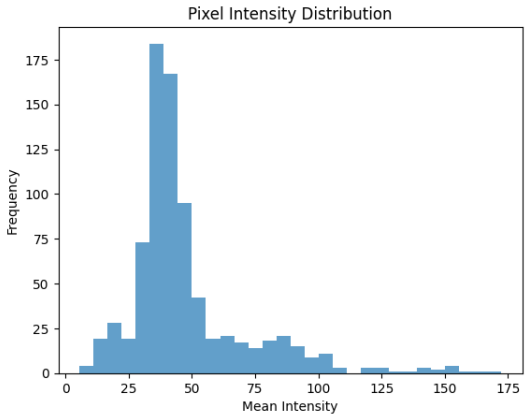

Fig. 3: Mean Pixel Intensity of the Training Dataset

4) Fig. 4 shows different nature of random collected images from the dataset, with different rotations, backgrounds etc.

Fig. 4: Three random selected images from the training dataset

C. Preprocessing

1) In order to enhance the visibility of our learning features (hand bones), the CLAHE method is applied. CLAHE (Contrast Limited Adaptive Histogram Equalization) is a contrast enhancement technique that improves the local contrast of an image by applying histogram equalization to small regions, or tiles, of the image. It limits the amplification of noise by clipping the histogram at a specified threshold. Fig. 5 clearly shows the resulting effect of CLAHE.

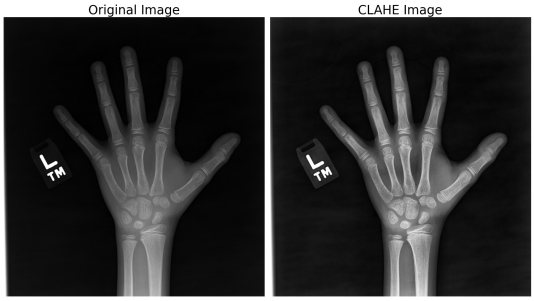

Fig. 5: CLAHE effect on a random image

2) A custom mask is applied to remove the labels present on the images. This mask is created using the histogram information of each image by detecting the peaks and valleys in the histogram and applying binary brightness masks with different thresholds based on these peaks and valleys. We obtained an acceptable result, as shown in Fig. **??**

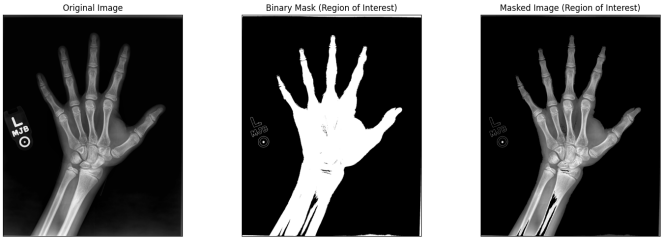

Fig. 6: Binary Mask to Remove Label

3) Thirdly, it is ensured that after the mask and CLAHE processing, the image pixel intensities remain within the 0-255 range. These values are then normalized to a 0-1 scale by dividing by 255.

4) Lastly, the training dataset is divided into bins based on their 0-1 scaled ground truth labels (ages in months). The histogram reveals the data imbalance. To create a more balanced dataset, the number of augmentations required for each image in every bin is calculated. Fig. 7 shows the provisional distribution of the training dataset after this dynamic augmentation is applied.

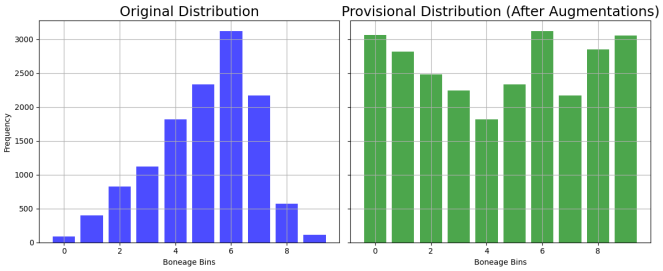

Fig. 7: Training Dataset Histogram with Dynamic Range Augmentation

V. Learning Framework

Below, the seven models used in this research for comparison are described in detail.

A. Baseline CNN Model

The *Baseline CNN Model* serves as the foundational architecture for this research, designed to predict age in months based on hand X-ray images and gender information. Below are the key components and details of its implementation:

Inputs:

- **Image Input:** Hand X-ray images are provided as a single-channel input with dimensions $256 \times 256 \times 1$.
- **Gender Input:** Gender information is provided as a scalar input.

Preprocessing: A preprocessing layer applies random rotations (+-10 degrees) to augment the image data. Missing pixels from rotations are filled using the nearest available pixel values.

Convolutional Neural Network (CNN) for Image Features:

- The CNN begins with an initial convolutional block, which includes:
  - A 3×3 convolutional layer with 32 filters.
  - Batch normalization for stable training.
  - ELU (Exponential Linear Unit) activation for non-linearity.
  - MaxPooling for spatial downsampling.
- Deeper convolutional layers are added with progressively increasing filters $(64, 128, 256)$. Each layer incorporates:
  - Residual-style convolutional blocks, which help mitigate the vanishing gradient problem by allowing shortcut connections.
  - MaxPooling for feature aggregation and dimensionality reduction.
- Global Average Pooling (GAP) is employed at the end to reduce the spatial dimensions while preserving learned features.

Features extracted from the CNN are passed through a fully connected layer with 512 units, ReLU activation, and $L2$ regularization. A dropout layer with a rate of 0.4 is applied to reduce overfitting.

The gender input is processed through a dense layer with 64 units and ReLU activation.

The extracted image features and gender features are concatenated to form a combined feature representation. The fused features are processed through two fully connected layers with 1024 units each, ReLU activation, $L2$ regularization, and a dropout rate of 0.3.

The final output layer uses a sigmoid activation function, predicting a single scalar value corresponding to the age in months.

A custom Root Mean Squared Error (RMSE) loss function is employed to measure the regression error, providing interpretable loss values.

The model is compiled with the Adam optimizer, using a learning rate of 0.0001 to facilitate stable and efficient training.

This baseline model integrates both image and gender information effectively while incorporating essential techniques such as residual connections, dropout, and augmentation to enhance its robustness and generalizability.

***Model Training Progress**:* The model training stopped at 14 epochs due to early stopping based on the validation loss. We obtained a `training MAE of 0.0570` and a `validation MAE of 0.0980`. The training took 34 minutes using free NVIDIA T4 GPUs available on Google Colab.

For the unbalanced dataset setup, the model training stopped at 18 epochs due to early stopping based on the validation loss. We obtained a `training MAE of 0.0599` and a `validation MAE of 0.0696`. The training took 21 minutes using free NVIDIA T4 GPUs available on Google Colab. Even though the number of epochs increased, the reduction in training time is explainable with the reduced amount of dataset, as we do not perform extra augmentations to balance the dataset.

Even though the resulting validation mae is better for the unbalanced dataset setup, we observed that the convergence of the training for the balanced dataset setup is smoother, and validation loss converged to its lowest value faster.

B. InceptionV4:

The architecture is highly similar to the baseline model with inputs of 256x256 grayscale images and gender. After the image input passes through the stacked Inception modules, a Global Average Pooling (GAP) layer is applied. This reduces the spatial dimensions while retaining global information, providing compact image features. These are concatenated with gender information and passed through additional dense layers to combine the information effectively. Dropout is employed to mitigate overfitting, using a rate of 0.3, as in the baseline model.

The difference lays in the following:

a) Inception Blocks: The Inception-v4 architecture employs a modular design with three primary types of blocks: **Inception-A**, **Inception-B**, and **Inception-C**, interspersed with **Reduction-A** and **Reduction-B** blocks for down-sampling. These blocks facilitate multi-scale feature extraction by utilizing parallel convolutional operations with varying kernel sizes.

The structure of the inception blocks are shown in Fig. 8

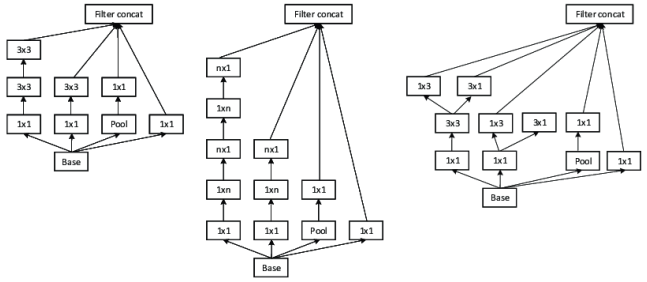

Fig. 8: Inception A-B-C blocks

b) Stem Block: The model begins with a **stem block**, which applies a series of convolutional and pooling operations to down-sample the input image while preserving spatial information.

***Model Training Progress**:* For the balanced dataset setup, the model training continued until reaching maximum number of epochs. We obtained a `training MAE of 0.0253` and a `validation MAE of 0.0426`. The training took 206 minutes (3 hours 24 minutes) using free NVIDIA T4 GPUs available on Google Colab.

C. Attention CNN with CBAM Integration

This subsection details the implementation of the Attention CNN model, which extends the baseline CNN by incorporating the Convolutional Block Attention Module (CBAM).

1) Architecture Overview: The general structure of this model is highly similar to the baseline CNN. The primary difference lies in the introduction of the CBAM module, which enhances feature representation by applying attention mechanisms at both the channel and spatial levels.

2) Convolutional Block Attention Module (CBAM): The CBAM module integrates two attention mechanisms:

- **Channel Attention:** This mechanism focuses on selecting important feature channels while suppressing irrelevant ones. It operates in two steps:
  1) Global average pooling and global max pooling are applied along the channel axis to generate two descriptors.
  2) These descriptors are processed by shared fully connected layers to reduce and restore channel dimensions. The outputs are combined using element-wise addition, followed by a sigmoid activation.

  The resulting channel attention map is applied to the input feature map using element-wise multiplication.
- **Spatial Attention:** This mechanism emphasizes significant spatial regions in the feature map. The steps include:
  1) Compute average and max pooling along the channel axis to generate two spatial descriptors.
  2) Concatenate these descriptors along the channel axis and pass them through a 7×7 convolutional layer with a sigmoid activation.

  The resulting spatial attention map is applied to the feature map using element-wise multiplication.

Fig. 9 showcases a visual representation of CBAM block [6]

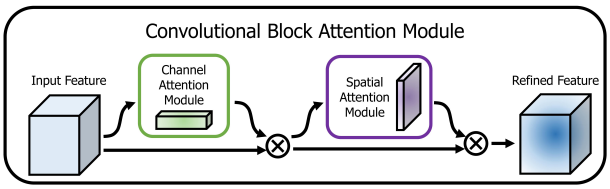

Fig. 9: CBAM block

The CBAM module is applied after each convolutional layer, as a drop-in enhancement to the baseline model's convolutional blocks. The rest of the model's structure, including preprocessing, dense layers, and gender integration, remains identical to the baseline CNN.

CBAM improves the model's ability to focus on relevant features by leveraging both channel and spatial attention, resulting in better representations of complex patterns in the data.

Model Training Progress: For the balanced dataset setup, the model training continued until reaching maximum number of epochs. We obtained a `training MAE of 0.0372` and a `validation MAE of 0.0483`. The training took 65 minutes (1 hour and 5 minutes) using free NVIDIA T4 GPUs available on Google Colab.

D. Decision Fusion Using Carpal Bones

In the decision fusion model, we trained baseline CNN model using cropped images that focus on the carpal bones in hand radiology images. These predictions are then combined with those from the different model's predictions that are trained on full images, using a weighted sum.

The optimal weights for this sum are determined through a grid search, which is tested on the test dataset to find the best combination for improved performance.

The model trained on carpal bone images receives input with dimensions $102 \times 102 \times 1$. This resolution is chosen because the images are cropped by removing 30% from both the left and right sides, 50% from the top, and 10% from the bottom, thereby isolating the carpal bone region of the hand, as illustrated in Fig. 10.

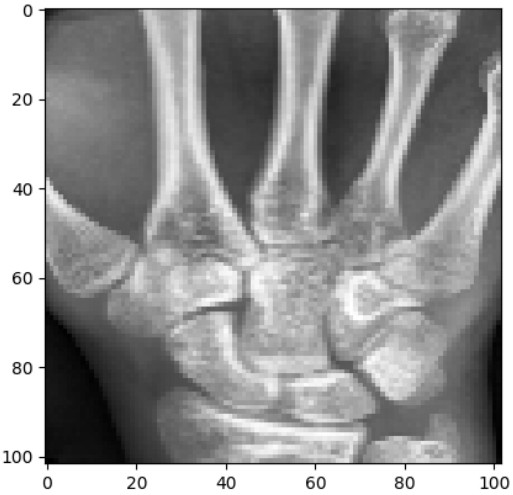

Fig. 10: Extracted Carpal Bones

Model Training Progress: The model solely trained on carpal bone images completed the training in 19 epochs due to early stopping based on the validation loss. We obtained a `training MAE of 0.0483` and a `validation MAE of 0.0607`. The training took 12 minutes using free NVIDIA T4 GPUs available on Google Colab.

E. ResNet50 Transfer Learning

The ResNet50 model, which is pre-trained on the ImageNet dataset, requires RGB images as input. To accommodate this requirement, the greyscale images are replicated across three channels, converting them to 256x256x3 images, which are then fed into the pre-trained ResNet50 model.

We remove the final layer of the ResNet50 architecture and connect it to dense layers, similar to our other models, to enable learning from our specific dataset.

The key difference in this model is that it does not incorporate gender information, making it suitable for performance comparison purposes.

The model training stopped at 12 epochs due to early stopping based on the validation loss. We obtained a `training`

MAE of 0.0564 and a validation MAE of 0.0680. The training took 49 minutes using free NVIDIA T4 GPUs available on Google Colab.

F. ResNet50 Transfer Learning with Gender

This model follows a similar logic to other architectures, using dense layers to combine gender information with the extracted image features from the ResNet50 model.

It is trained for performance comparison and to assess the impact of including gender information on the prediction accuracy.

Also, another training performed making the first 10 layers of the base model trainable, in an attempt to increase the prediction accuracy.

The model training stopped at 18 epochs due to early stopping based on the validation loss. We obtained a training MAE of 0.0515 and a validation MAE of 0.0626. The training took 49 minutes using free NVIDIA T4 GPUs available on Google Colab.

VI. Results

In this section, all the relevant test results obtained from the test dataset containing 200 images are presented.

A. Baseline CNN Model with Balanced Dataset

Prediction Mean Absolute Error: 18 months

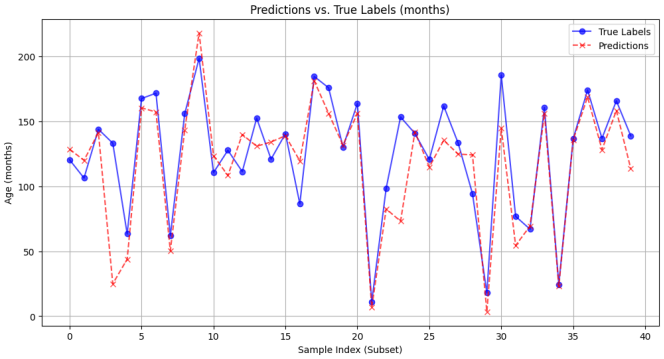

Fig. 11: Prediction Plot on 40 random selected test images

B. Baseline CNN Model with Unbalanced Dataset

Prediction Mean Absolute Error: 19.49 months

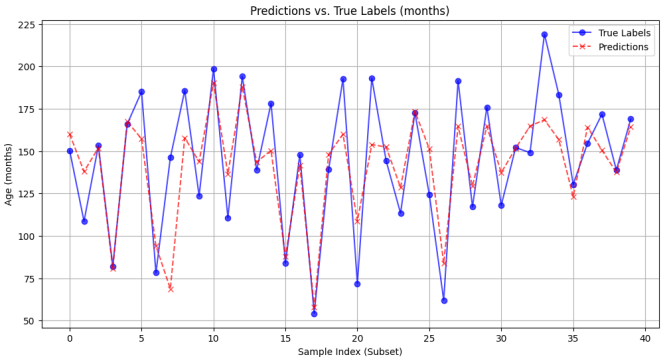

Fig. 12: Prediction Plot on 40 random selected test images

C. InceptionV4 Model with Balanced Dataset

Prediction Mean Absolute Error: 9.24 months

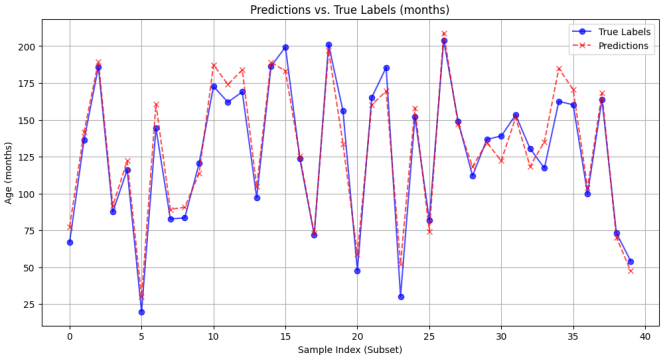

Fig. 13: Prediction Plot on 40 random selected test images

Fig. 14: Predictions for 3 random selected test images

D. Attention-CNN Model with Balanced Dataset

Prediction Mean Absolute Error: 10.96 months

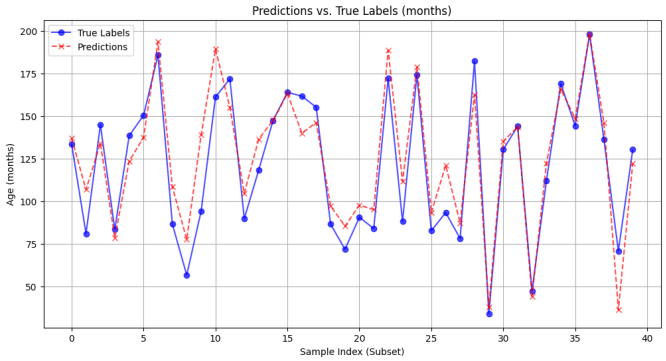

Fig. 15: Prediction Plot on 40 random selected test images

Fig. 16: Predictions for 3 random selected test images

E. Decision Fusion Model

Fig. 17 shows the prediction accuracy of baseline-cnn model trained with carpal bone regions.

`Prediction Mean Absolute Error: 15.28 months`

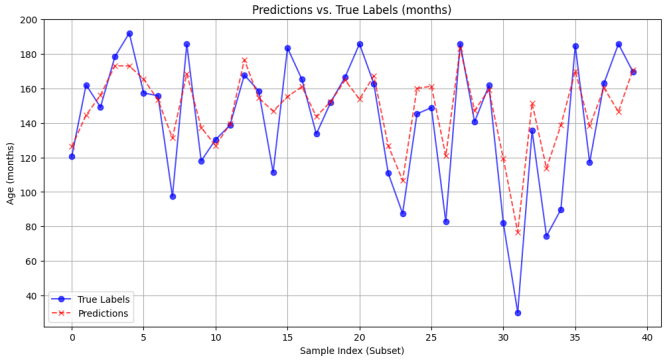

Fig. 17: Prediction Plot on 40 random selected test images

Fig. 18 shows the difference in prediction for 3 random images in test dataset, between inceptionv4 model and baseline-cnn model trained on carpal bones. It also showcases the combined prediction with best parameters threshold = 0.5 and carpal_weight=0.3, improving the prediction accurracy of the inception model.

Fig. 18: Predictions for 3 random selected test images

The decision fusion model obtained a `Scaled Prediction Mean Absolute Error: 0.04`, which is better than the one obtained using only inception-v4 model, which was 0.0415.

Decision fusion parameters are defined as follows:

- **threshold:** Stands for after which scaled age value, the cnn model prediction trained on carpal bones will be incorporated in decision.
- **carpal˙weight:** The coefficient of model prediction trained on carpal bones in weighted sum.

The best parameters are obtained through a grid search, utilizing all the combinations for 0.3, 0.4, 0.5, 0.6 and 0.7 for each parameter.

F. ResNet50 Transfer Learning Model (only images) with Balanced Dataset

`Prediction Mean Absolute Error: 10.37 months`

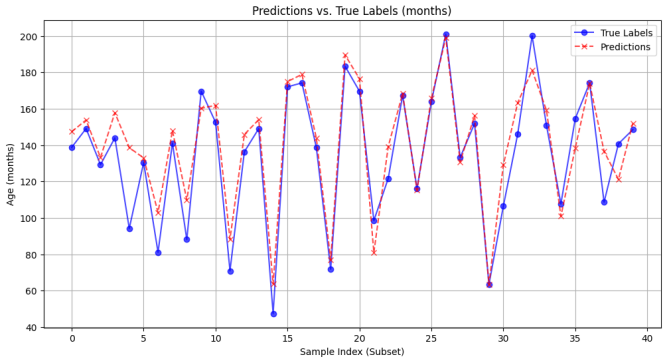

Fig. 19: Prediction Plot on 40 random selected test images

Fig. 20: Predictions for 3 random selected test images

G. ResNet50 Transfer Learning Model + Gender with Balanced Dataset

```
Prediction Mean Absolute Error: 14.76 months
```

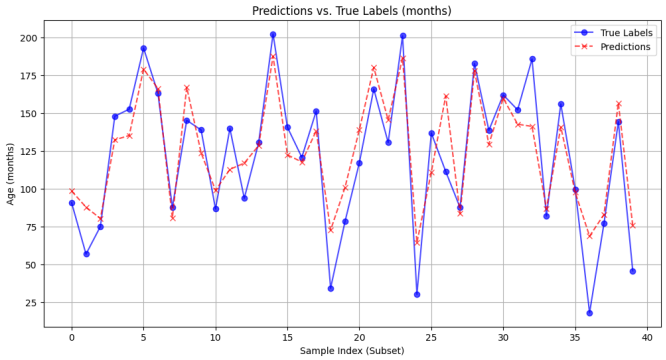

Fig. 21: Prediction Plot on 40 random selected test images

H. ResNet50 Transfer Learning Model + Gender + Trainable lower layers with Balanced Dataset

```
Prediction Mean Absolute Error: 15.29 months
```

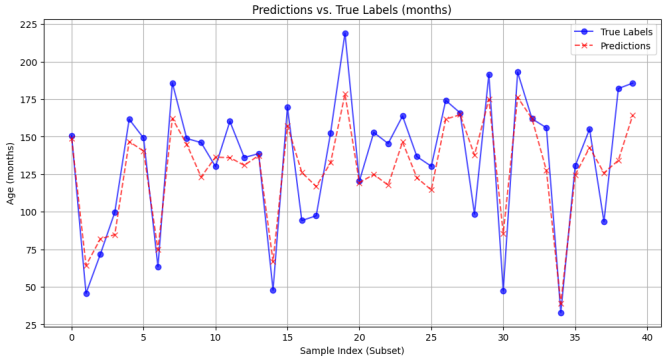

Fig. 22: Prediction Plot on 40 random selected test images

Fig. 23: Predictions for 3 random selected test images

I. Model Comparison

Model Name	Parameters	Model Size (MB)
Base CNN	2,162,945	24.85
Base CNN Unbalanced Dataset	2,162,945	24.85
Attention CNN	2,184,746	25.21
Inception V4	43,665,537	501.32
CNN Carpal	2,323,713	27.15
ResNet50 Images Only	24,637,313	282.31
ResNet50 Transfer Learning	26,278,401	121.41
ResNet50 Transfer Learning LL	26,278,401	121.52
Decision Fusion	45,828,482	526.17

Model Name	Mem. Use MB(Inference)
Base CNN	7.89
Base CNN Unbalanced Dataset	0.05
Attention CNN	1.32
Inception V4	3.88
CNN Carpal	-
ResNet50 Images Only	80.61
ResNet50 Transfer Learning	65.00
ResNet50 Transfer Learning LL	47.82
Decision Fusion	-

Model Name	Inference Time (s)
Base CNN	0.3581
Base CNN Unbalanced Dataset	0.2264
Attention CNN	0.3307
Inception V4	11.0596
CNN Carpal	0.4132
ResNet50 Images Only	4.0506
ResNet50 Transfer Learning	3.8197
ResNet50 Transfer Learning LL	3.2174
Decision Fusion	11.4177

Model Name	Test MAE
Base CNN	0.0793
Base CNN Unbalanced Dataset	0.0859
Attention CNN	0.0483
Attention CNN with Clahe+Masked Dataset	0.0830
Inception V4	0.0415
CNN Carpal	0.0502
ResNet50 Images Only	0.0457
ResNet50 Transfer Learning	0.0650
ResNet50 Transfer Learning LL	0.0674
Decision Fusion	0.0399

Fig. 24: Single image prediction, True Label: 0.6081

Model Name	Single Prediction
Base CNN	0.6055
Base CNN Unbalanced	0.6479
Attention CNN	0.6114
Inception V4	0.6944
CNN Carpal	0.5678
ResNet50 Images Only	0.6256
ResNet50 Transfer Learning	0.5931
ResNet50 Transfer Learning LL	0.5730
Decision Fusion	0.6719

VII. Concluding Remarks

In our research, we explored state-of-the-art architectures to address a regression problem focused on predicting skeletal age from hand bone radiographs. We implemented various preprocessing techniques, including masks and CLAHE, alongside diverse data augmentation strategies. Additionally, we examined the impact of balancing the dataset using dynamic augmentation rates on model performance. To further enhance accuracy, we integrated a decision fusion model, demonstrating its effectiveness in improving the predictive performance of existing models.

We achieved high accuracy in predicting skeletal age, comparable to results reported by Larson et al. (2018). Our best performance, using a decision fusion mechanism with the Inception-v4 model, yielded a mean absolute error (MAE) of 10.05 months (0.03987) on the test dataset. This is close to their MAE of 0.67 years (8.04 months) achieved with the ResNet50 architecture.

Key Findings from Comparative Analysis

1) **Decision Fusion Mechanism**: Combining a 4-layer CNN (with residual connections) trained on carpal bones and an Inception-v4 model trained on full images via a weighted sum resulted in the best test MAE of 0.03987.
2) **Pre-trained ResNet50**: Transfer learning with ResNet50 achieved the second-best test MAE of 0.415.
3) **Gender and Base Layers Impact**: Including gender information or unfreezing the base layers of ResNet50 degraded performance, increasing the test MAE to 0.0650 and 0.0674, respectively.
4) **Balanced Dataset Advantage**: Using data augmentation to balance the dataset improved the Baseline CNN's performance, reducing the test MAE compared to the unbalanced dataset.
5) **Attention Mechanisms**: Adding Convolutional Block Attention Modules (CBAMs) to the Baseline CNN significantly improved its test MAE from 0.0793 to 0.0483.
6) **Inference Time**: Inception-v4 had the longest inference time, requiring 11.05 seconds per image.

Our results highlight that a lightweight CNN architecture incorporating attention mechanisms can achieve high performance while maintaining efficiency, making it a strong candidate for applications requiring fast and accurate predictions.

Future Work

1) Due to constraints of Google Colab, we had to limit our learning process to a maximum of 20 epochs. We believe that slightly better accuracies can be obtained with longer training times, especially for deeper architectures.
2) The best parameters for the decision fusion mechanism were calculated using the Baseline CNN trained on carpal bones and the Inception-v4 network. Exploring different parameters across other network combinations may yield even better improvements.
3) Developing a more efficient preprocessing method to remove the labels on the images may enhance prediction accuracy further.

Difficulties Encountered

During the project, we encountered several challenges that provided valuable learning opportunities:

1) **Standardization of Target Labels and Images**: Initially, our model struggled to train effectively due to incorrect standardization of target labels and input images. This led to an "explosion" in our regression model's output, as the target labels ranged from 0 to 228. We resolved this issue by engaging in thorough discussions with our teaching assistants, who provided patient and effective guidance.
2) **Implementation of TensorFlow Datasets**: Our first attempt to implement TensorFlow datasets was ineffective due to improper use of the `flat_map()` function. This oversight resulted in inefficient batch processing, causing out-of-memory errors in our Colab session and preventing the full utilization of our dataset. By carefully revisiting the problem and adopting a meticulous approach, we developed a clean and efficient implementation. This not only resolved the issue but also allowed us to utilize the dataset as intended.

Reflections and Learnings

Through this project, we gained valuable insights into handling regression tasks with state-of-the-art architectures for skeletal age prediction. We learned the importance of proper preprocessing, as initial errors in label standardization significantly hindered model performance. Implementing TensorFlow datasets effectively was another key takeaway, where careful optimization resolved memory issues and ensured the full utilization of our data. Balancing datasets and exploring attention mechanisms like CBAMs highlighted the potential for enhancing model accuracy. Finally, the decision fusion mechanism demonstrated how combining models can yield superior results, emphasizing the value of creative model integration in machine learning projects.

REFERENCES

[1] D. B. Larson, M.-C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, pp. 313–322, April 2018. Epub 2017 Nov 2.

[2] 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, pp. 498–503, February 2019. Epub 2018 Nov 27.

[3] S. Hamida, O. El Gannour, A. Maafiri, Y. Lamalem, I. Haddou-Oumouloud, and B. Cherradi, "Data balancing through data augmentation to improve transfer learning performance for skin disease prediction," in *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1–7, 2023.

[4] N. Vila Blanco, M. Carreira, P. Varas-Quintana, C. Balsa-Castro, and I. TomÃ¡s Carmona, "Deep neural networks for chronological age estimation from opg images," *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 01 2020.

[5] D. Å tern, C. Payer, and M. Urschler, "Automated age estimation from mri volumes of the hand," *Medical Image Analysis*, vol. 58, p. 101538, 2019.

[6] S. Trivedi, "Understanding attention modules: Cbam and bam â€" a quick read," *Medium*, June 2020. Published on VisionWizard, Medium.