Bone Age Assessment from Hand-Wrist Radiographs: Neural Network Approaches

Daniela Di Labbio[†], Erica Marras[‡]

***Abstract*—Assessing bone age in infants and newborns is essential for identifying growth abnormalities and hormonal imbalances, enabling timely and effective intervention.**
Traditional methods like Greulich & Pyle (GP) and Tanner-Whitehouse (TW) have been widely used for decades. Recently, however, Machine Learning has emerged as a powerful tool to automate this biological age estimation. Most current approaches rely on pre-trained models and annotated images, sometimes including gender information when analyzing radiographs. This report examines three main model architectures applied to the RSNA dataset: a Convolutional Neural Network (CNN), and two Residual Neural Network (ResNet)-based architectures. The simple ResNet model undergoes further refinement through a Global-Local dual-branch configuration, allowing it to leverage previously learned features without relying on annotations. Additionally, the models are evaluated both with and without gender data, revealing the impact of gender information on improving the accuracy of bone age estimation.
The best performing model is the gender-incorporated ResNet, in one of the two versions, reaching an MAE of 8.01 and an accuracy within 1 year of 77.95%.

***Index Terms*—Hand Bone Age Assessment, Convolutional Neural Network, Residual Neural Network, Feature Extraction, Global - Local ResNet**

I. Introduction

In pediatric endocrinology, it is especially important to assess a child's growth and puberty in relation to biological age rather than chronological age. This is because various disturbances, primarily related to hormones, can cause a mismatch between chronological and biological age. Bone age (BA) is the only biological indicator of maturity available from birth to adulthood, making it a crucial tool for clinicians to accurately assess maturation rates in children and to detect and monitor growth issues. [1] Although no formal medical standard exists, for a few decades two methods based on non-dominant hand-wrist radiographs have been frequently employed for Bone Age Assessment: the Greulich & Pyle (GP) method and the Tanner-Whitehouse (TW) method. The former is an atlas matching method entirely based on comparisons, the latter is a scoring system which assess a point score to each of the 20 hand-wrist bones (ROIs, Region of Interest) that are then combined and mapped to an age value through a table.
These methods are time consuming, prone to human error and not well accurate due to the inter- and intra- clinicians variability, so the need to experiment Machine Learning approaches.

In Sec. II related works on Bone Age Assessment are presented, while in Sec. III the reader can find an high-level description of the experiments implemented.

An important step of this project is related to the way the images are preprocessed and subjected to augmentation techniques simulating natural variations and taking into account the characteristics of the dataset and the need for the model to generalize effectively across diverse conditions (Sec. IV). The proposed models (sec. V) are based on Convolutional Neural Networks (CNN) and include a baseline CNN, a ResNet-like architecture, and a Global-Local dual-branch variant. The primary objectives are to evaluate the performance of different architectures in terms of performance and resource efficiency (memory and time) (sec. VI), to demonstrate the impact of incorporating gender information, and to develop a method that allows the model to focus on important regions without relying on annotations.

The architectures were designed and implemented within the constraints of a single NVIDIA T4 GPU with 16GB of memory. These hardware specifications significantly influenced not only the experiments conducted and our architectural choices, but also the performance. This factor plays a crucial role and has been thoroughly addressed in the conclusions (sec. VII), where a comprehensive discussion on selecting the optimal configuration is provided.

II. Related Work

The release of the RSNA bone age assessment challenge in 2017 initiated a number of approaches to tackle bone age prediction [2]. Many of the top-ranking solutions in this competition shared commonalities, including the use of pre-trained models and an end-to-end framework. Notably, the 1st and 2nd place solutions incorporated gender information into their model architectures, while the 4th and 5th place models utilized annotations.

In [3], the authors aimed to compare expert evaluations with deep learning approach by fine-tuning a ResNet, handling it as a classification task, where each label is the combination of age in months and the gender. So, in this work, gender was treated as a separate prediction target rather than a feature to assist in model learning, resulting in a MAE of 7.56.

Koita et al. (2020) [4] emulated the Tanner-Whitehouse method, using CLAHE pre-processing for radiographs before inputting them into a two-stage model. The first stage detected six specific ossification regions, while the second stage used a

[†]Department of Mathematics, University of Padova, email: daniela.dilabbio@studenti.unipd.it

[‡]Department of Mathematics, University of Padova, email: erica.marras@studenti.unipd.it

gender-region-specific regression model to make predictions, resulting in a total of 12 models trained. This approach achieved a MAE of 4.56, the best performance known to date.

Keji Mao et al. (2022) [5] proposed a four-stage, end-to-end approach that includes a feature extractor, a ROI (Region of Interest) selector, a ROI guidance, and the final assessment. Their approach notably used features from a ResNet to create fine-grained images for more effective learning. Rather than integrating gender in the model, they trained separate models for each gender, yielding a MAE of 6.65.

Finally, *Li et al. (2023)* [6] proposed a gender-assisted, annotation-free model that performs cascaded extraction of critical bone regions. This model utilizes a pre-trained Inception V3 to identify the carpal area on a heat map, which is then masked. The masked image is fed into another Inception V3 extractor to focus on the metacarpal and phalanx regions. These refined layers, combined with gender information, generated the age prediction, achieving a MAE of 5.45.

These recent approaches have been highly influential and inspiring in our project. Following these methods, we explored the potential of a ResNet-based architecture with CLAHE-preprocessed images and examined feature extraction techniques to help guide the model toward more accurate learning.

III. Processing Pipeline

This work introduces three distinct architectures: a simple CNN (Sec. V-A) as the baseline model, and two ResNet-based architectures —one referred to as the Global-Local ResNet (Sec. V-B, V-C). Each architecture is tested both in its base form and with two different methods for integrating gender information during training (V-D).

The input to all models consists of RSNA hand-wrist radiograph images, which undergo a standardized preprocessing sequence outlined in Sec. IV-A. To enhance model generalization and simulate real-world variability, augmentation techniques (Sec. IV-B) are applied exclusively to the training set, also accounting for the limited batch size due to memory constraints.

An overview of these processing steps is provided in Fig. 1.

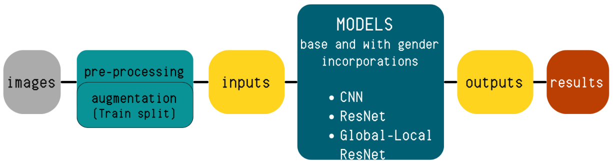

Fig. 1: Pipeline, an overview

The CNN model includes three convolutional layers with progressively larger filter sizes, serving as the initial baseline for subsequent models. Additionally, the ”gender #1” variant establishes a baseline for gender-integration experiments.

The ResNet models are composed of ten residual blocks, each featuring two convolutional layers with skip connections. This structure is chosen to mitigate the vanishing gradient problem and allow deeper layers to learn complex hierarchical features without performance degradation.

The Global-Local ResNet combines two processing paths: a global branch that employs a series of residual blocks to capture high-level features from the entire image, and a local branch that processes selected regions from the global branch with a refined set of filters. This dual approach enables the model to utilize both broad contextual and fine-grained information, enhancing predictive accuracy by focusing dynamically on the most informative areas.

All models are trained by minimizing the Mean Squared Error (MSE) Loss, with Early Stopping employed to prevent overfitting and optimize computational efficiency. The ADAM optimizer is used, paired with a learning rate scheduler for controlled convergence and stability. Following training, model performance is evaluated using the Mean Absolute Error (MAE) metric, and by calculating the Accuracy withing 1 year.

Comparing the models in terms of resource use and performance allows us to identify the optimal approach among those tested.

IV. Images and Features

Images in this project are sourced from the RSNA (Radiological Society of North America) dataset, used in the 2017 Pediatric Bone Age Challenge. This dataset includes 12,611 hand-wrist radiographs of children with bone ages ranging from 1 to 228 months (0-19 years), covering both male and female subjects. As shown in Tab. 1, the dataset is reasonably balanced, both in terms of gender distribution and age distribution in months. The images has been saved in a Google Drive folder connected to the Google Colab notebook, which avoids the need to re-upload the images in each session. Images are loaded in a single (gray) channel at a fixed resolution of 256 × 256. The Train/Validation/Test split is set at 80%, 10%, and 10% respectively.

	General	Male	Female
count	12611	6833	5778
mean	127.32	135.30	117.88
std	41.18	42.14	37.91
min	1	1	4
max	228	228	216

TABLE 1: Dataset distribution

A. Pre-Processing

To well prepare the data for model training, **normalization** is applied, scaling pixel values relative to the minimum and maximum intensity within each image to enhance contrast and standardize pixel values across the dataset.

To further improve contrast, especially in images with varying brightness levels, **CLAHE** (Contrast Limited Adaptive Histogram Equalization) is utilized. A dynamic clip limit is set based on image brightness, with darker images receiving a higher clip limit, thus avoiding over-enhancement of contrast.

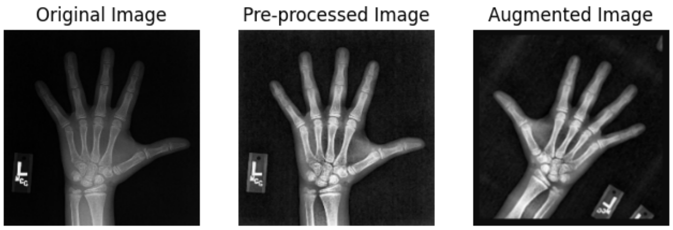

Fig. 2: Example of an image before and after pre-processing and augmentation.

B. Image Augmentation

The augmentation steps, applied solely to the training split, were selected to achieve multiple goals: mainly to increase dataset diversity and improve model robustness by mimicking real-world variations, and also to reduce overfitting risks given the small batch size (16), constrained by limited computational resources. All the transformations applied are:

- **Random Horizontal Flip**: The image is randomly flipped left-to-right, taking in consideration that the Bone Age Assessment is always based on non-dominant hands leading to an over–representation of left hands.
- **Random Rotation**: Using the Random Rotation layer, images are rotated up to approximately ± 7.2 degrees (0.125 radians), adding rotational diversity, simulating slightly different positions of hands that children could assume.
- **Random Brightness Adjustment**: Random brightness shifts (up to 10%) account for lighting changes, enhancing the model's robustness to different lighting conditions.
- **Random Zoom**: Images are randomly zoomed in or out within a range of 0.8 to 1.2 times their original size. This helps the model learn to recognize objects at various scales and distances, since the way the radiographs are taken could be subjects to variations.

After pre-processing (IV-A) and any applied augmentations (IV-B), images are resized to 224×224 — this resizing is done afterward to ensure all the transformations work with higher-quality images.

It is worth mentioning that to facilitate more effective model learning, the target bone ages in months have been normalized (e.g., a bone age of 228 months is scaled to 1). Normalization helps reduce the scale of target values, making it easier for the model to detect patterns and adjust predictions. For interpretability, the results are later denormalized to reflect the actual age in months, providing outputs in a meaningful and clinically relevant form.

V. Learning Framework

This section begins with an overview of essential components shared across all implemented models, aimed at enhancing training efficiency, managing learning dynamics, and ensuring robust performance. These core elements include the optimizer, learning rate scheduler, early stopping mechanism, and model weight saving strategy. Following this description, we delve into the specific architectures, detailing the unique aspects and implementation of each model (V-A, V-B, V-C).

- **Optimizer**: The training process leverages the Adam optimizer, which combines the advantages of both momentum and adaptive learning rates. By incorporating the principles of both techniques, Adam ensures stable convergence by dynamically adjusting the learning rates based on real-time gradient estimates for each parameter. This adaptive mechanism allows for efficient and reliable optimization, enhancing the overall performance of the model.
- **Learning Rate Scheduler**: A custom learning rate scheduler is implemented to modulate learning rates dynamically in response to validation performance. The scheduler operates within defined boundaries, with an initial learning rate of 0.001 and a minimum threshold of 0.00005. Specifically, if validation loss fails to improve for a set patience period of three epochs, the learning rate is reduced by 50%, allowing the model to make finer adjustments as it neared optimal convergence, but never falling below the minimum threshold. This approach mitigates issues of overshooting in earlier training stages while maintaining sufficient gradient magnitude through the imposed lower bound, enabling controlled convergence and promoting both model stability and generalizability.
- **Early Stopping:** To prevent overfitting and ensure optimal convergence, we implemented an early stopping mechanism that monitors validation loss during training. The system activates only after the learning rate has reached its minimum predetermined value, at which point it terminates the training process if no improvement in validation performance is observed over three consecutive epochs (patience threshold p = 3). This sequential approach ensures the model fully leverages the learning rate adaptation before triggering the early stopping condition.
- **Weight Saving:** Model persistence is ensured through a two-level weight preservation strategy: weights are saved at each epoch to capture the training trajectory, while a separate checkpoint stores the model state with the best validation performance. This setup supports training continuity, reproducibility, and post-training analysis, and also enables easy reloading from checkpoints in case of GPU interruptions.
- **Loss**: The model's performance is evaluated using Mean Squared Error (MSE) as the loss function, which quantifies the average squared difference between predicted

and actual bone age values. Formally, for a batch of N samples, the MSE is computed as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i represents the ground truth bone age and $\hat{y}_i$ the model's prediction for the i-th sample. This loss function penalizes larger prediction errors more heavily due to its quadratic nature, making it particularly suitable for regression tasks where accurate age prediction is crucial. Additionally, MSE's differentiability ensures stable gradient computation during the optimization process.

A. CNN

Convolutional Neural Networks (CNNs) are deep learning models highly effective in image processing, as they automatically learn spatial feature hierarchies and capture local dependencies with convolutional filters. In our study, we designed and implemented a custom CNN architecture which, as illustrated in Fig.3, processes the primary image input through a sequence of three interconnected convolutional blocks. Each block consists of a convolutional layer with progressively increasing filter sizes (32, 64, and 128 filters respectively), a 3x3 kernel, and the Rectified Linear Unit (ReLU) activation function. This layered structure enables the convolutional layers to extract increasingly complex hierarchical representations from the images. To optimize processing, a max-pooling layer follows each convolutional layer, reducing the spatial dimensions. To complete the base architecture, the final output of the convolutional process is flattened and passed through a fully connected (dense) layer of 64 units. Building upon this fundamental structure, our research focused on developing and analyzing three distinct variants of the CNN model, each characterized by a different approach to integrating gender information (Sec. V-D). The base model, serving as a comparative reference, exclusively uses the image input and produces the final prediction through a dense layer, without considering supplementary gender information.

B. ResNet

ResNet (Residual Neural Network) is a deep Convolutional Neural Network (CNN) designed to overcome issues like vanishing and exploding gradients that often occur in deep learning by using skip connections. These skip, or shortcut, connections allow information to bypass one or more layers, preserving gradient flow even in very deep networks.

The architecture implemented in this project, as illustrated in Fig. 4, begins with a 5 × 5 convolutional layer that has 32 filters and is followed by batch normalization, ReLU activation, and max pooling, setting the stage for subsequent layers to extract hierarchical features.

The core of this model consists of residual blocks with varying filter depths, which allow the model to capture increasingly complex patterns and to learn more abstract representations as the network deepens. In the initial section, four residual blocks with 32 filters process the input to learn basic features, followed by two blocks with 64 filters, two blocks with 128 filters, and two blocks with 256 filters. Each residual block contains two convolutional layers, both of which use 3×3 filters and maintain dimensional consistency through batch normalization and ReLU activation. The blocks are linked through skip connections, which allow the input of each block to be added directly to its output, enhancing gradient flow and enabling efficient learning.

After the residual blocks, the network includes a global average pooling layer to condense spatial information, resulting in a fixed-size vector suitable for regression. At this point, the model offers flexibility in processing gender information as an additional feature, if wanted (sec. V-D). In case the gender is integrated, that information is concatenated with the main image features before the final dense layer. This final dense layer outputs a single continuous value representing the predicted bone age (normalized as explained in sec. IV).

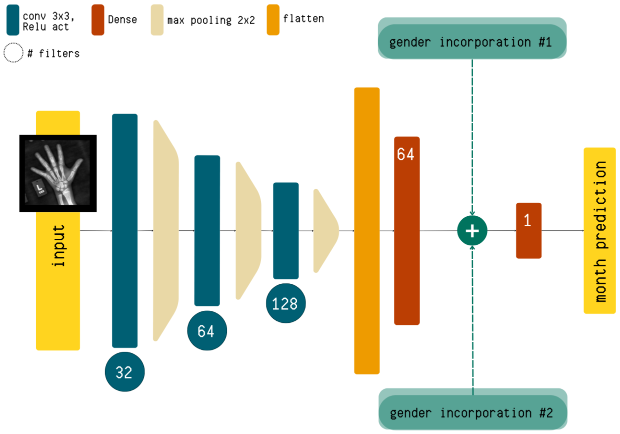

Fig. 3: CNN architecture overview with optional gender incorporation. Dashed arrows show two mutually exclusive configurations, both optional in the base version. (see Fig.4 for the specific details of gender incorporations)

C. Global-Local ResNet

The Global-Local ResNet (Fig. 5) is a dual-branch architecture designed to capture both global and localized features from radiographic images, enhancing the model's ability to recognize subtle patterns that might not otherwise be enough considered. The global branch starts with a 5×5 convolutional layer with 32 filters, followed by batch normalization, ReLU activation, and max pooling, preparing the input for subsequent residual blocks. These blocks, arranged with increasing filter depths (32, 64, and 128), progressively extract complex global features from the input.

For enhanced learning ability, the model includes a **local branch** that focuses on specific regions of interest within the image. This region is identified through an attention mask derived from the global feature map by extracting

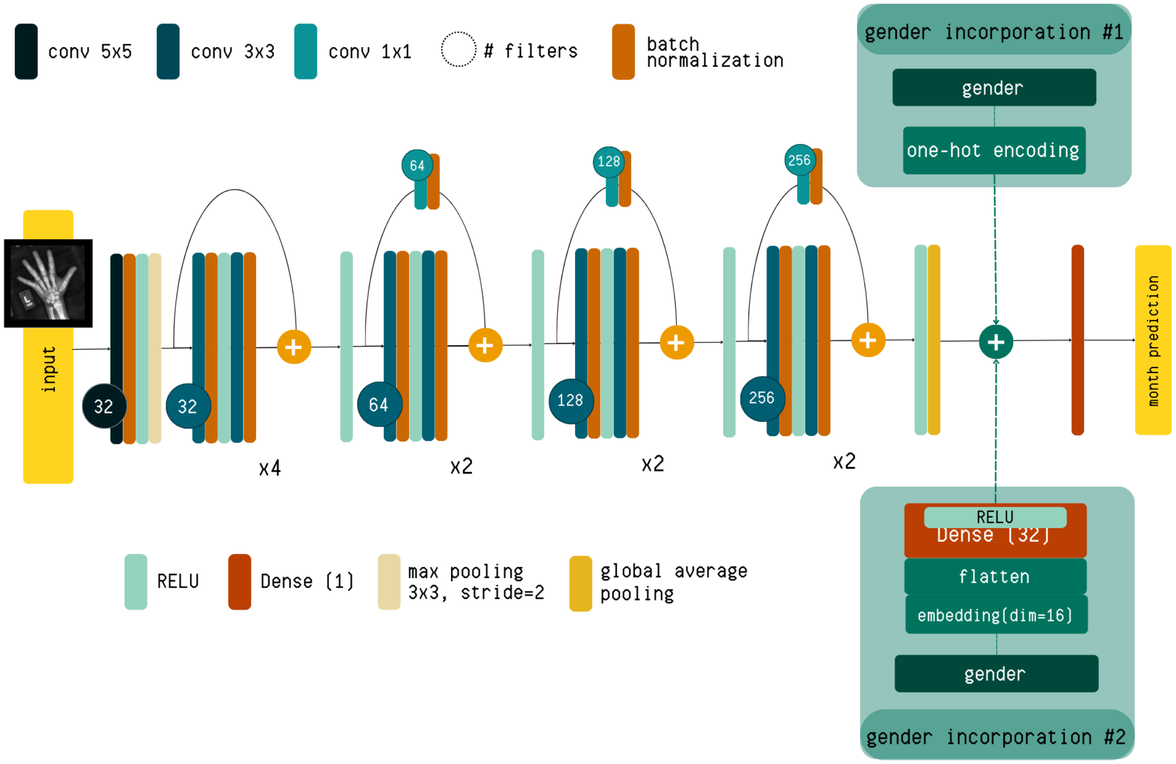

Fig. 4: ResNet architecture overview with optional gender incorporation. Dashed arrows show two mutually exclusive configurations (#1, #2), not considered in the basic version.

the maximum activation. After normalizing these activations, a binary mask is created using a threshold that highlights only the most salient regions (controlled by the parameter τ). Connected component analysis is applied to this mask, allowing the model to isolate the primary area of interest. To ensure consistency in input to the local branch, the resulting bounding box is adjusted to be square by expanding the smaller dimension to match the larger one, with any necessary padding applied symmetrically. This square region is then resized or padded to a fixed size, making it compatible with subsequent processing in the local branch.

The local branch processes this isolated region with additional convolutional layers and residual blocks (16, 32, and 64 filters), capturing fine-grained, localized features. After extracting global and local features, the two branches are concatenated, followed by global average pooling to produce a condensed feature vector. This vector can optionally incorporate gender information, either through one-hot encoding or embedding layers (sec. V-D), enabling a more context-aware prediction. In its simplest form (base variant), the model outputs a single regression value, without seeing the gender information.

D. Gender Incorporation

Gender information can be included in all the presented models in two distinct ways (see Fig.4 for a reference):

- **One-Hot Encoding of Gender (#1):**
  Gender information is represented as a one-hot encoded vector of dimension two. In the CNN model, this vector is concatenated directly with the output of the dense layer of dimension 64. In the two ResNet models, it is concatenated with the output of the global average pooling layer. This approach was chosen both to limit the usage of parameters and to see a first effect of incorporating gender information with minimal model complexity.
- **Gender Embedding (#2):**
  The gender information is used as a single integer and an embedding layer is used to learn a dense representation for each gender category, mapping it to a 16-dimensional vector. This embedded vector is then flattened and passed through a dense layer to expand it to a 32-dimensional representation, allowing the model to incorporate more complex gender-specific information. This approach was chosen to enable the model to capture subtler gender-related patterns through a richer, trainable embedding. Finally, this 32-dimensional gender feature is concatenated with the layer before the last dense one which outputs the month prediction.

VI. Results

This section presents a comprehensive analysis of the experimental findings with the three models implemented —CNN,

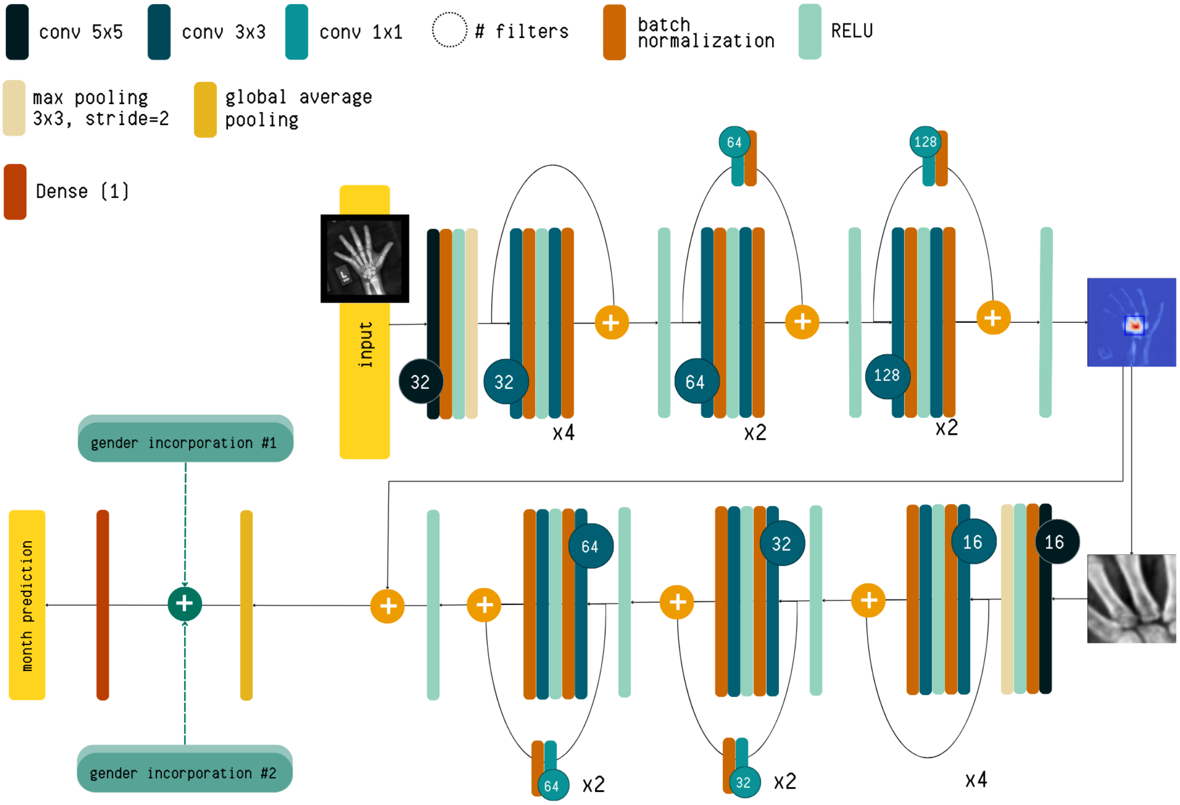

Fig. 5: Global-Local ResNet architecture overview with optional gender incorporation. Dashed arrows indicate two mutually exclusive gender integrations, not considered in the base version.

ResNet, and Global-Local ResNet— progressively building insights into the models' performance and resource efficiency. It is important to recall that the free Google Colab environment, along with the T4 GPU, is used for the experiments and that the dataset, consisting of 12,611 images, has been split into train, validation, and test sets, with percentages set at 80%, 10%, and 10%, respectively.

The resource requirements, in terms of parameter count and memory usage, are detailed in Tab. 2, and analyzing it offers valuable insights into the balance between model complexity and computational demands.

Specifically, the CNN architecture has the highest parameter count and memory usage, making it the most computationally intensive of the three models. On the other hand, the ResNet model has a more compact structure, reducing both the number of parameters and the memory consumption of the CNN. Finally, the global-local ResNet model is the most resource-efficient, with the lowest parameter count and memory requirements, providing a balanced trade-off for applications where memory and computation constraints are critical. Furthermore, note that including gender information doesn't significantly affect model complexity and computational requirements.

Moving on a more refined and detailed analysis, the neural network architectures are evaluated based on their computational efficiency in terms of time (Tab. 3), and their prediction performance (Tab. 4). A key factor influencing both aspects was the choice of the learning rate scheduler, which significantly impacted the number of epochs across all models and so the overall training time, while also enhancing the results. Although our primary objective was to maximize model performance, it's worth noting that adjusting the learning rate scheduler for efficiency by increasing its minimum threshold could still produce satisfactory results without excessively compromising accuracy, and saving time.

Model	**Variant**	**Parameters**	**Memory**
CNN	base	6,515,329	24.85 MB
	gender #1	6,515,331	24.85 MB
	gender #2	6,515,937	24.86 MB
ResNet	base	2,839,937	10.83 MB
	gender #1	2,839,939	10.83 MB
	gender #2	2,840,545	10.84 MB
Global-Local ResNet	base	922,113	3.52 MB
	gender #1	922,115	3.52 MB
	gender #2	922,721	3.52 MB

TABLE 2: Number of Parameters and Memory for each model

The CNN architecture, while offering a computational speed advantage over other models, with an average epoch training time of 800 s, delivers less impressive predictive performance.

In its best configuration, the model achieves an accuracy of only 57.02% within the one-year threshold. Comparatively, the ResNet architecture, although requiring around 1,400 s for each epoch, demonstrates substantially improved accuracy metrics. In particular, the gender #2 variant yields an MAE of **8.01** and reaches an accuracy of **77.95%** within the one-year threshold. These results establish ResNet as the best-performing model, striking the optimal balanced between time computational resources usage and prediction accuracy. Among the Global-Local ResNet variants, the gender #2 model achieves an accuracy of 73.75% and an MAE of 8.48. This performance, which meets annual targets, represents a significant improvement over the base version and a satisfyng result. Although the Global-Local ResNet has the lowest memory computational requirements, its average epoch time is the longest among all models, approximately 1,700 seconds.

For a better analysis, the plot shown in Fig. 6 demonstrates the relationship between predicted and actual bone ages (in months) referred to the ResNet model, gender #2 variant. This best model demonstrates robust performance across both genders, as evidenced by the tight clustering of points along the reference line. However, some notable deviations are observed at the extremes of the age spectrum, with increased dispersion of points for very young and elderly subjects. This suggests reduced prediction accuracy in these age ranges, a pattern consistent across both genders. This finding provides valuable insight into the model's operational boundaries.

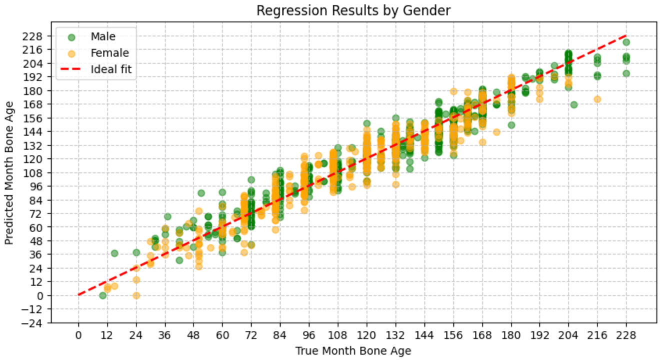

Fig. 6: Relation between actual and predicted bone ages (in months) for ResNet variant gender #2 model.

In conclusion, the experimental results, as shown in the Tab. 4, reveal a clear trend across the different models and variants, highlighting the efficacy of gender-integrated versions in terms of overall performance, both considering mean absolute error and accuracy within 1 year. Moreover, it is evident how the choice to use ResNet-based architectures improved the results, decreasing the overall MAE to be within 1 year.

VII. Concluding Remarks

The experiments conducted in this study provided a comparative analysis of different approaches to bone age assessment in children, including methods that integrate gender as a predictive factor. Key challenges involved using unannotated images, training the models from scratch, and evaluating the impact of gender on prediction accuracy in an effort to improve outcomes. Three different main architectures have been tested: a CNN, a ResNet with ten residual blocks and a Global-Local ResNet to let the model further focus on the most important local part detected from the Global branch. Consulting the results, we have clearly seen how the gender incorporation positively impacted the improvement of the models, resulting in the best variants among all for each architecture.

Model	**Variant**	**Training Time (s) and number of epochs**	**Test Time (s)**
CNN	base	30,609 (38 ep.)	81
	gender #1	36,223 (45 ep.)	81
	gender #2	35,000 (43 ep.)	81
ResNet	base	63,137 (45 ep.)	74
	gender #1	61,876 (43 ep.)	69
	gender #2	72,966 (51 ep.)	66
Global-Local ResNet	base	78,094 (48 ep.)	81
	gender #1	61,362 (36 ep.)	66
	gender #2	54,456 (33 ep.)	67

TABLE 3: Training and Test Time for each model

Moreover, while the Global-Local architecture delivered good and satisfactory results and stands out as the model with the fewest parameters, its longer training time and the notable gap in accuracy over one year compared to the best-performing model led us to select the ResNet gender #2 model as the overall best choice.

Although our MAE results (best one 8.01) did not surpass the state-of-the-art benchmarks reported in Sec. II (4.56), they offer a strong foundation for further enhancement and could potentially compete with existing solutions as refinements are made.

There is ample room for improvement. We believe the Global-Local ResNet architecture could be enhanced by deepening the Global Branch, which was limited by resource constraints. To address the time-intensive nature of training, we could also increase the threshold in the learning rate scheduler or change the percentage reduction itself, still ensuring the model continues to learn in a controlled manner.

Additionally, the architecture could be explored by utilizing features from shallower layers to capture high-level spatial features alongside the already taken deeper semantic feature information.

Future improvements could involve modifying the Global-Local ResNet architecture to allow the model to make separate predictions from the Global branch, the Local branch, and the combined features from both branches. This adjustment would involve calculating three MSE losses—one for each prediction—which would enable the model to update its weights by considering the contributions of each branch individually as well as their combined effect. This approach

Model	**Variant**	**Overall MAE**	**Male MAE**	**Female MAE**	**Overall Accuracy within 1 year**	**Male Accuracy within 1 year**	**Female Accuracy within 1 year**
CNN	base	15.28	14.98	15.30	47.83%	51.51%	43.72%
	gender #1	13.10	13.06	13.11	53.53%	54.82%	52.09%
	gender #2	12.33	12.37	12.33	57.02%	56.63%	57.45%
ResNet	base	11.96	12.74	11.90	58.45%	52.71%	64.82%
	gender #1	8.35	8.15	8.37	75.42%	78.31%	72.19%
	gender #2	**8.01**	**7.92**	**8.02**	**77.95%**	**79.22%**	**76.55%**
Global-Local ResNet	base	11.12	10.29	11.18	62.97%	71.39%	53.60%
	gender #1	8.55	8.39	8.57	73.51%	76.20%	70.52%
	gender #2	8.48	8.27	8.50	73.75%	76.05%	71.19%

TABLE 4: Performance

could help the model learn more efficiently by clarifying the role of each branch in the final prediction, though it would likely require additional memory resources.

We also find the methodology proposed by *Li et al. (2023)* [6] compelling; future studies could explore the use of masking techniques to guide the model's attention not only toward the primary influential area (which, as in our model, was identified as the carpal region) but also to the metacarpal and phalanx regions. The Global-Local ResNet allowed us to explore feature representations before the final prediction step, providing valuable insights into the model's inner workings. However, resource limitations presented challenges, slowing experimentation and restricting the range of configurations we could test. Initially, we used a small portion of the dataset for hyperparameter tuning and preliminary architecture tests, but for consistency with the final results based on the full dataset, we chose not to report those findings here.

References

[1] M. Prokop-Piotrkowska, K. Marszałek-Dziuba, E. Moszczyńska, M. Szalecki, and E. Jurkiewicz, "Traditional and new methods of bone age assessment-an overview," *Journal of Clinical Research in Pediatric Endocrinology*, vol. 13, no. 3, p. 251, 2021.

[2] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. Mamonov, A. Bilbily, M. D. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. P. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, "The rsna pediatric bone age machine learning challenge.," *Radiology*, vol. 290 2, pp. 498–503, 2019.

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

[4] S. Koitka, M. S. Kim, M. Qu, A. Fischer, C. M. Friedrich, and F. Nensa, "Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks," *Medical Image Analysis*, vol. 64, p. 101743, 2020.

[5] K. W. J. M. Keji Mao, Wei Lu and G. Dai, "Bone age assessment method based on fine-grained image classification using multiple regions of interest," *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 15–23, 2022.

[6] Z. Li, W. Chen, Y. Ju, Y. Chen, Z. Hou, X. Li, and Y. Jiang, "Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction," *Frontiers in Artificial Intelligence*, vol. 6, 2023.