Precise and accurate automated assessment of skeletal age from hand radiographs using deep learning

Joshua Bram, Ayoosh Pareek, Samuel A Beber, Ruth Hendry Jones, Olivia Christina Tracey, Daniel William Green, Peter D Fabricant

INTRODUCTION:

Surgeons treating skeletally immature patients rely on assessment of physeal status to determine appropriate/safe surgical strategies. The gold standard method of skeletal maturity estimation is the Greulich and Pyle (GP) atlas using hand radiographs. Unfortunately, variability of bone age estimates using the GP atlas from radiologist-determined consensus bone age (i.e. the ground truth) approaches 6 months. Additionally, this method is notably time-consuming. These concerns have prompted development of other shorthand methods for skeletal age determination, though these are limited by focus on one or two radiologic parameters for estimation.

As skeletal age in adolescent patients can greatly influence determination of appropriate orthopedic intervention (e.g. ACL reconstruction strategy), more reliable, automated methods of determining skeletal age are needed. Deep learning (DL) has previously been used for radiographic assessment but has been limited by low accuracy and mixed results. This study sought to develop a highly accurate and reliable DL model for determination of accurate skeletal age based on hand radiographs for routine use by orthopedic practitioners.

METHODS:

This retrospective study evaluated bone age using hand X-rays. Three publicly available hand radiograph datasets for model development and validation were utilized: the Radiological Society of North America (RNSA) dataset (12,611 training, 1,425 validation, and 200 test hand radiographs), the Radiological Hand Pose Estimation (RHPE) dataset (5,491 training, 713 validation, and 79 test hand radiographs), and the digital hand atlas (DHA, 1,390 hand films). All three datasets report corresponding sex and skeletal age, and the RHPE and DHA datasets also contain chronological age. Additionally, one institutional hand radiograph dataset (200 radiographs) was utilized for external validation.

Image pre-processing included a pre-trained U-net for hand mask prediction, a histogram equalization technique for image improvement, and alignment standardization. Two models were developed using the PyTorch Lightning framework to predict skeletal age from hand radiographs, with the ConvNeXt architecture serving as the common vision encoder. The model was first trained on the RSNA dataset using sex and skeletal age as inputs for 100 epochs with the Mean Absolute Error (MAE) loss function. The pre-trained model was subsequently fine-tuned on the RHPE dataset and validated on the DHA (**Table 1**) incorporating chronological age due to its clinical utility. Model performance was assessed using MAE and Root Mean Squared Error (RMSE).
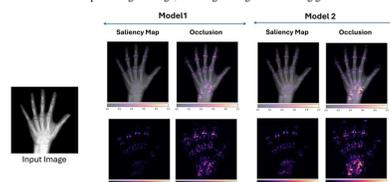
RESULTS:

Table 1 presents the comprehensive results of both models. First, model 1's MAE on the RSNA test set was 3.68 months, lower than all previously reported errors in the literature (**Table 2**). Inclusion of chronological age in model 2 improved performance evaluation metrics for the DHA, as MAE decreased from 5.66 months in model 1 to 4.65 months in model 2. On the institutional dataset, the MAE from model 1 was 4.82 months. Interpretability analysis in the form of gradient-saliency maps (**Figure 1**) revealed that both models focused on areas radiologists typically emphasize (digital physes/epiphyses, carpal bones, and distal radius and ulna physes/epiphyses).

DISCUSSION AND CONCLUSION:

Technological innovation in the field of artificial intelligence, specifically computer vision and DL, has allowed rapid improvement in the ability of models to reliably and accurately characterize radiographic images. With potential as valuable diagnostic and classification assistants, they necessitate large, accurate data inputs to achieve excellent performance. Leveraging newer DL technologies trained on nearly 20,000 hand radiographs, the models developed in this study achieved state-of-the-art performance across all three included hand radiograph datasets with high performance on an institutional dataset as external validation. The collective impact of these results is substantial and sets a new standard for pediatric bone age assessment. By developing a highly accurate hand radiograph model, this model enhances diagnostic precision, aiding in the timely and appropriate intervention for orthopedic conditions.

Figure 1. Gradient-based saliency maps. The heatmaps highlight the areas of the image that contribute most to predicting bone age, with brighter regions indicating greater contribution.

Table 1. Models results on test sets.

Dataset	Model 1 MAE	Model 1 RMSE	Model 2 MAE	Model 2 RMSE
RSNA Test set	3.68 [3.24 - 4.14]	4.92 [4.23 - 5.73]	-	-
RHPE	6.87	-	5.06	-
DHA	5.66 [5.37 - 5.95]	7.96 [7.34 - 8.73]	4.65 [4.42 - 4.87]	6.38 [5.96 - 6.81]
Institutional	4.82 [4.25 - 5.43]	6.42 [5.70 - 7.21]	5.31 [4.68 - 5.98]	7.07 [6.22 - 7.98]

MAE = Mean Absolute Error. RMSE = Root Mean Squared Error.
Brackets show the 95% confidence interval.

Table 2. Comparison of the MAE of models 1 and 2 against the state-of-the-art methods on the RNSA test set, DHA, and RHPE test sets

References	Year	RSNA Test set	DHA	RHPE
Halabi[5]	2017	4.2, 4.4, 4.5 (top 3)	-	-
Escobar[21]	2019	3.85	-	6.86
Liu[21]	2019	4.38	-	-
Koitka[17]	2020	4.56	11.93	-
Mao[23]	2023	4.58	-	-
Deeplasia[27]	2024	3.9	5.8	-
Model 1	2025	*3.68	5.66	6.87
Model 2	2025	-	*4.65	*5.06

* Indicates the smallest reported Mean Absolute Error (MAE)