Original Article

Artificial Intelligence in Bone Age Assessment of Pediatric Hand-Wrist X-Rays: Systematic Review and Meta-Analysis of Studies from 2019–2024

Sony Sutrisno[1]*, Ronny[2]

[1] Department of Radiology, Faculty of Medicine and Health Sciences, Krida Wacana Christian University, Jakarta-Indonesia
[2] Department of Radiology, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta-Indonesia

Article Info

Corresponding Author
Sony Sutrisno
Department of Radiology
Faculty of Medicine and Health Sciences, Krida Wacana Christian University, Jakarta-Indonesia
Email:sony.sutrisno@ukrida.ac.id

Date Received:
22nd September, 2024
Date Revised:
29th January, 2025
Date Accepted:
11th March, 2025

This article may be cited as:
Sutrisno S, Ronny. Artificial Intelligence in Bone Age Assessment of Pediatric Hand-Wrist X-Rays: Systematic Review and Meta-Analysis of Studies from 2019–2024. J Postgrad Med Inst. 2025;39(1):1-9. http://doi.org/10.54079/jpmi.39.1.3492

Abstract

Objective: To explore the existing literature on the use of artificial intelligence (AI) and machine learning (ML) in bone age assessment (BAA).

Methodology: This study was conducted in accordance with the PRISMA guidelines. Three databases (PubMed, Scopus, and Cochrane) were screened for studies published from 2019–2024. Two reviewers independently selected the studies. The modified Kitchenham-Charters' checklist was used to critically appraise the studies. Only studies reporting the mean absolute error (MAE) were included in the analysis. Data regarding the study characteristics, subject characteristics, ML technique, model ground truth, and the performance of the studies were extracted.

Results: The review included 33 studies, mostly from East Asia. Most studies used in-house datasets consisting of hand radiographs of their respective local population. Convolutional neural networks are the most popular AI algorithms used. Most studies used radiologist-annotated BA as ground truth rather than the chronological age. The meta-analysis revealed a weighted MAE of 7.54 months, an improvement compared to the previous study.

Conclusion: AI and ML models continue to demonstrate rapid advancements in their application for BAA. This study described the current trends in ML research and explored ongoing obstacles in BAA, as well as the prospective role of AI. While promising, further research is still required to address current limitations, such as validity issues. Subsequent studies should also be conducted with rigorous methodology and thorough reporting.

Keywords: Artificial intelligence, Bone age assessment, Machine learning, Mean absolute error, Pediatrics

Introduction

Bone age assessment, essentially a measurement of skeletal maturity, is typically performed to evaluate the growth of bones and diagnose disorders in pediatric population. It is frequently employed in the field of pediatric endocrinology, where it guides the evaluation of short or tall stature, growth disorders, and pubertal disorders. Adult height predictions can also be derived from BA through various proposed algorithms. Moreover, it can be found implemented in other disciplines as well, such as orthodontics, orthopedics, and even forensics.[1,2] Conventional methods of BAA typically involve the use of plain radiograph of the hand and wrist of the non-dominant hand (usually the left). These radiographs are then compared to standard plates in an atlas, with the two most popular methods being the Greulich-Pyle (GP) and the Tanner-Whitehouse (TW) method.[3]

The GP method involves comparing the subject's radiograph with images containing hand radiographs along with their corresponding age and explanation of the gradual age-related changes. The standards were sourced from 1,000 upper-middle-class male and female Caucasians living in Ohio, USA, from 1931–1942. Despite being popular due to its simple learning curve and ease of use, the GP method also exhibited subjectiveness and a high inter- and intra-observer variability.[1,3]

The TW method is reported to be a more objective method for BAA. It involves the maturity scoring of each bone segment compared to a standard, which is then calculated and translated to the final bone age. The original TW (TW1) data was collected from 2,600 average-class British children, in the 1950s–1960s and published in 1962. Subsequent updates were later issued to account for factors influencing the total bone maturity score and BA. Additionally, two additional methods of bone evaluation and an algorithm for height prediction were also introduced, with the TW2 (1983) and the TW3 (2001).[1,3] Although it shows greater reproducibility than the GP method, the TW method is considerably more time-consuming. One study reported the average time for BAA estimation with the TW method as 7.9 minutes, compared to 1.4 minutes with the GP method; a difference of 6.5 minutes, which is troublesome when handling a large volume of patients.

In response to the issues of variability and duration, automated techniques have been developed to optimize the BAA process. Artificial intelligence (AI) has shown promise in enhancing the accuracy and efficiency of BAA. In 2017, the Radiological Society of North America (RSNA) held a contest for the development of AI and machine learning (ML) models that could determine BA from a curated set of pediatric hand radiographs. The five best models achieved a mean absolute error (MAE) ranging from 4.2–4.5 months.[4] Since then, numerous models and techniques have been developed and refined.

Although a similar study has been published before,[5] the rapid growth of the body of knowledge warrants an updated review of the current evidence on the role of AI in BAA. This systematic review aim to describe literature published from 2019–2024 to offer an updated perspective on the role of AI on BAA.

Methodology

This systematic review was reported according to the Preferred Reporting Items for Systematic Review and Meta Analysis (PRISMA).[6]

Search strategy and study eligibility

The authors independently conducted literature search for studies attempting to find out the role of artificial intelligence in predicting bone age from pediatric hand X-rays. The search was performed on July 3, 2024, by both authors in three databases, i.e. PubMed, Cochrane, and Scopus. The specific queries used to find the relevant studies are given in Table 1. The study protocol can be found at https://osf.io/dz7m9. Studies were deemed eligible only if they satisfied the predetermined inclusion and exclusion criteria (Table 2). Any disagreements were discussed to a consensus. The records were screened and annotated in Microsoft Excel (RRID: SCR_016137).

Data extraction

The authors extracted the data and tabulated the results with Microsoft Excel. Each of the authors re checked the other author's extracted data for accuracy. The following data was extracted: (1) lead author and year of publication: (2) study country of origin: (3) subjects' country of origin in the testing data set: (4) sample size of testing data set: (5) proportion of sex and age of the testing data set: (6) machine learning algorithm used: (7) BAA technique; (8) the ground truth used for model testing; (9) the presence of other readers for comparison and the employed BAA technique:and (10) study outcomes. For uniformity, all extracted data expressed in years was converted to months by multiplying by twelve.

Critical appraisal

Critical appraisal of retrieved studies was done using a checklist based on the guidelines by Kitchenham and Charters7, which was then further modified by Dallora.[8] The appraisal was performed by both authors, and a consensus was reached in case of any discordance between the reviewers in case of any disagreement. The checklist evaluated studies based on three parameters: (1) study design and reporting (6 items); (2) data quality (5 items); and (3) techniques employed (3 items). Each item was scored as "0", "0.5", or "1", depending on whether the study did not fulfill, partially fulfilled,

or fully fulfilled the corresponding item, respectively. Studies were deemed eligible in terms of quality if they reached the threshold score of 8. The full checklist is displayed in Table 3.

Data analysis

The study outcome to be analyzed was the mean absolute error (MAE), which is defined as "the measure of errors between paired observations expressing the same phenomenon". In this case, it is the measure of errors between the AI model and the ground truth. It is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where y_i is model's predicted value for observation *i*, x_i is the ground truth value for observation *i*, and n is the number of observations. Only studies reporting the MAE with standard deviation or confidence intervals were further included in the meta-analysis.

The meta-analysis of the MAE was performed in Microsoft Excel by calculating the weighted mean using the inverse-variance weighting method. Studies reporting multiple MAE of the model were combined. Studies reporting the MAE with confidence intervals were converted to standard deviation for the purpose of weighting. A forest plot was generated with JASP 0.19.0.0 (RRID:SCR_015823) to illustrate the contribution of each included study along with the overall results.

Results

Search results and study characteristics

From the initial 194 records identified from the three databases, a total of 33 studies were included in this review.[9–41] The details of the literature search can be seen on the PRISMA flow diagram of this study (Fig. 1). Using the aforementioned checklist, sixty-one studies were assessed for eligibility. A total of nineteen studies were below the quality threshold and thus ineligible for inclusion. The description of the study aims, machine learning techniques, and study results were mostly clear from all the assessed studies. However, only seventeen studies performed assessment of statistical significance, making it the most frequently unfulfilled item by the studies. The appraisal results are available in Supplementary 1.

Of the 33 studies, most (n = 11) were published in 2023. Most of the studies originated from Asia, specifically China (n = 14) and South Korea (n = 7). There were also studies from Europe, North America, and South Africa. The datasets used to evaluate the models were sourced by various means. Many studies used the RSNA bone age dataset (n = 10).[42] Some also used the University of South California's Digital Hand Atlas for the test dataset (n = 5).[43] However, most studies utilized in-house datasets to measure the model's performance. Many studies do not clearly report the size, sex, and age of the samples used in the test set (Table 4).

A sheer majority of studies (n = 29) utilized a convolutional neural network (CNN) architecture in the ML models. Other algorithms used were support vector machine (SVM), deep neural network, and triplet loss algorithm. Conventional methods such as GP and TW were utilized by some models in the BAA itself. However, many studies also utilized their own novel methods. The ground truths used in the studies were a combination of in-house annotations and BAs labelled from the database. Five studies used the chronological age (CA). Many of the studies also compared the performance of their models with either human readers, other models, or different configurations of their own model (Table 5).

The performance of the models was mostly reported in MAE (also written as mean absolute deviation by some authors). There were also studies that did not report the MAE but reported correlation coefficients instead. The performance of the studied models is elaborated in Table 6.

After evaluation and conversion, 13 studies were included in the meta-analysis.[12–14,16,25–28,31,32,35,40] The studies displayed homogeneity ($I^2 = 0$) and had a weighted MAE of 7.54 months, with a 95% confidence interval of 7.31–7.77 (Figure 2).

Discussion

ML is a branch of AI in which a machine learns by identifying patterns and makes decisions based on its learning process with little human intervention. Important traits of ML models include their ability to adapt independently of human interference, to learn from past computations, and to produce valid and reliable results when introduced to new data. This is achieved by feeding the model data and letting it optimize its own parameters so that its performance improves.[44,45] Numerous ML algorithms have been developed, such as decision tree learning, clustering, Support Vector Machines (SVM), k-means nearest neighbor, restricted Boltzmann machines, and random forests. ML techniques require the extraction of discriminant features to work efficiently. These features are still largely unknown and are still being researched. However, CNN models are popular ML algorithm for image recognition as they are able to self-learn and extract the needed features to perform image interpretation.[46] They are also the most prevalent model found in the current study.

CNNs are a type of deep learning, a new and rapidly growing area of ML. They mainly consist of 3 components, which are: (1) the convolutional layer: responsible for creating feature maps summarizing discriminating features from the input by systematic application of

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

filters; (2): the pooling layer: handles feature maps from each node of the convolutional layer to form a new set of pooled feature maps; (3): fully-connected layer: receives pooled feature maps from the final pooling layer. This layer proceeds to interpret the image based on the pooled extracted features. The model then compares the output to the ground truth and computes its errors. It then determines the needed direction and magnitude of change and updates itself. The process repeats itself until there is not much improvement in the error. One important benefit of CNNs is how they are able to independently determine the most important discriminating features from the data.[44,45]

Deep neural networks are similar to CNN as both are classified as deep learning. They are called "deep" because of the number of layers involved, usually more than twenty layers. This is made possible by the increasingly powerful computing ability of today's computers.[45]

Another type of ML is the SVM, which transforms data into the widest plane (support vector) between two classes. While not new, interest in SVM has been revived by the addition of functions which are able to map classes to other dimensions (hyperspace) by the application of nonlinear functions. As such, the new classes are able to be separated by a plane (hyperplane), which was previously impossible.[45]

Triplet loss is a type of metric learning that learns by grouping items into three: an anchor (baseline), a positive image sample (belonging to the same category as the anchor), and a negative image sample (belonging to a different category from the anchor). Its aim is to minimize the distance between the anchor and positive sample and maximize the distance between the anchor and negative sample through the use of a loss function.[21] Compared to the results of the previous study, the performance of current models has improved. The previous study reported a pooled MAE of 9.96 months,[5] while the current study showed a lower MAE at 7.54 months. However, these results may not be fully comparable as the datasets used were different. The age range of the subjects in the dataset were also not uniformed in both previous study and the cur-

Table 1. Queries used on the databases for literature search

Database	Queries (searched on July 3, 2024)	Hits
PubMed	((“Artificial Intelligence”[Mesh] OR “Machine Learning”[Mesh] OR “Deep Learning”[Mesh] OR “Neural Networks, Computer”[Mesh] OR “Algorithms”[Mesh]) OR (“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network” OR “algorithm”)) AND ((“Bone Age”[Mesh] OR “Bone Development”[Mesh] OR “Epiphyses”[Mesh]) OR (“bone age” OR “skeletal age” OR “bone development” OR “epiphyseal” OR “epiphysis”)) AND ((“Radiography”[Mesh] OR “Hand”[Mesh] OR “Radiographic Image Interpretation, Computer-Assisted”[Mesh]) OR (“hand x-ray” OR “hand radiography” OR “hand imaging” OR “hand radiograph”)) AND ((“Child”[Mesh] OR “Adolescent”[Mesh]) OR (“child” OR “children” OR “pediatric” OR “adolescent”)) Filters: from 2019 – 2024	77
Cochrane	(“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network*” OR “algorithm*”) AND (“bone age” OR “skeletal age” OR “bone development” OR “epiphyseal” OR “epiphysis”) AND (“hand x-ray” OR “hand radiography” OR “hand imaging” OR “hand radiograph”) AND (“child*” OR “children” OR “pediatric*” OR “adolescent*”)	4
Scopus	TITLE-ABS-KEY (“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network*” OR “algorithm*”) AND TITLE-ABS-KEY (“bone age” OR “skeletal age” OR “bone development” OR “epiphyseal” OR “epiphysis”) AND TITLE-ABS-KEY (“hand x-ray” OR “hand radiography” OR “hand imaging” OR “hand radiograph”) AND TITLE-ABS-KEY (“child*” OR “children” OR “pediatric*” OR “adolescent*”) AND PUBYEAR > 2018 AND PUBYEAR < 2025	113

Table 2. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
• Published in English • Published between 2019 and 2024 • Studied subjects up to the age of 18 years • Developed and/or tested the performance of artificial intelligence models for bone age determination through hand-wrist x-rays • Assessed bone age in living individuals	• Utilized radiological modalities other than x-ray (plain radiograph) • Assessed bone age using bones other than those in the hands and wrists • Involved subjects with known conditions that may affect normal bone growth or the interpretation of bone age • Concerned with height prediction • Solely focused on the role of AI in enhancing radiologists' performance

Table 3. Modified Kitchenham and Charters checklist for critical appraisal

Parameter & Item	Score (0, 0.5, or 1)
Study Design and Reporting	
Are the aims of the study clearly stated?	
Does the study describe clearly the population being studied?	
Does the study clearly describe the inclusion/exclusion criteria?	
Are the limitations of the study reported either during the explanation of the study design or during the discussion of the study results?	
Were the findings clearly reported?	
Is there no lack of blinding that could introduce bias?	
Data Quality	
Are the data characteristics (variables, subjects' country of origin, etc.) used in the study clearly defined?	
Are the data characteristics (variables, subjects' country of origin, etc.) used in the study valid?	
Is the data collection method clearly described?	
Are the data characteristics related to age valid/verified in some way?	
Technique(s) Employed	
Is/are the ML technique(s) being employed clearly described?	
Is it clear how accuracy was measured?	
Was statistical significance assessed?	
Total score (threshold: 8 out of 13)	

0: item not fulfilled at all; 0.5: item partially fulfilled; 1: item fully fulfilled

Table 4. Study and characteristics of subjects in the test set

Study		Characteristic of Test Set			
Author (Year)	Country	Country (institution)	Size	Sex	Age
Escobar et al.[9] (2019)	Colombia	NM (author group's dataset) USA (RSNA)	Authors' group: 80 RSNA: 80 Total: 160	Whole dataset: 54% female and 46% male Authors' group Test set: NM RSNA Test set NM Whole dataset: 54% female and 46% male	Whole dataset: central tendency: 126 months Range: 0–240 months RHPE Test set NM RSNA NM
Liu et al.[10] (2019a)	China	USA (DHA)	NM	NM	NM
Liu et al.[11] (2019b)	China	USA (RSNA)	2000	NM	NM

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

Booz et al.[12] (2020)*	Germany	Germany (in-house: University Hospital Frankfurt)	514	Male: 252 Female: 262	Mean: 10.20 ± 4.85 years Range: 3–17
Dehghani et al.[13] (2020)*	Iran	USA (DHA)	442	Male: 222 Female: 220	Range: 0–18 years, central tendency NM
Guo et al.[14] (2020)*	China	USA (DHA)	NM	NM	NM
Koitka et al.[15] (2020)	Germany	Germany (in-house: University Hospital Essen) USA (RSNA and DHA)	Essen: 16,834 (total) RSNA: 200 DHA: 1,389 (total)	Essen (total = 16,834) Male: 8,927 Female: 7,907	RSNA Mean: 132 months DHA: NM
				RSNA (n = 200) Male: 100 Female: 100 DHA (total = 1,389) Male: 700 Female: 689	In-house: NM
Reddy et al.[16] (2020)*	USA	USA (RSNA)	200	Male: 100 Female: 100	NM for test set
Zulkifley et al.[17] (2020)	Malaysia	USA (RSNA)	200	NM	NM
Cheng et al.[18] (2021)	Taiwan	Taiwan (in-house: China Medical University Hospital)	820	Males: 312 Females: 508	Central tendency NM Range: 0 to 20 years
He et al.[19] (2021)	China	USA (RSNA)	200	NM	Central tendency NM Range: 0 to <19 years
Lee et al.[20] (2021)*	South Korea	South Korea (in-house: Korea University Anam Hospital)	102	Male: 51 Female: 51	Mean: 10.95 ± 2.37 years Range: 4.92–17.0 years
Madan et al.[21] (2021)	India	USA (DHA)	NM	NM	NM Range: 0 –3 years
Cheng et al.[22] (2022)	Taiwan	Taiwan (in-house: China Medical University Hospital & Asia University Hospital)	368	Male: 190 Female: 178	Mean: 8.97 ± 3.73 years Range: 2.06–15.99 years
Hui et al.[23] (2022)	China	China (in-house: Beijing Jishuitan Hospital)	1021	Male: 402 Female: 619	Range: 4–18 years, central tendency NM

Hwang et al.[24] (2022)	South Korea	South Korea (in-house: Asan Medical Center and Pusan National University Yangsan Hospital)	485	Male: 262 Female: 223	Mean: 10.0 ± 4.3 Range: 2–17 years
Lea et al.[25] (2022)*	South Korea	South Korea (in-house: Guro Hospital)	474	Male: 35 (7.4%) Female: 439 (92.6%)	Mean: 9.09 ± 1.68 years Range: 4–17 years
Zhang et al.[26] (2022)*	China	China (in-house: Beijing Jishuitan Hospital)	486	NM	Not mentioned explicitly (in histogram)
Zhao et al.[27] (2022)*	China	China (in-house: Zhangzhou Affiliated Hospital of Fujian Medical University)	400	NM	Range: 3–16 years, central tendency NM
Deng et al.[28] (2023)*	China	USA (RSNA)	97	NM	Range: 9–228 months, central tendency NM
Kasani et al.[29] (2023)	Iran	USA (RSNA)	200	NM	NM
Kim et al.[30] (2023a)	South Korea	South Korea (in-house: Pusan National University Yangsan Hospital & Dankook University Hospital)	Pusan/Yangsan: 343 Dankook: 321	Pusan/Yangsan Male: 183 Female: 160 Dankook Male: 164 Female: 157	Median: Pusan/Yangsan 10 (4-15) years Dankook 9 (5-14) years
Kim et al.[31] (2023b)*	South Korea	South Korea (in-house: Asan Medical Center)	227	NM	Range: 24–228 months, central tendency NM
Kim et al.[32] (2023c)*	South Korea	South Korea (in-house: unnamed institution)	453	Male: 148 Female: 305	Male: Mean: 11.10 years Median Age: 12.0 years Range: 2.0 - 17.0 years Female: Mean: 11.23 years Median Age: 11.5 years Range: 3.0 - 16.0 years
Li et al.[33] (2023)	China	USA (RSNA) China (in-house: Chongqing Jintongjia Children's Hospital)	RSNA: 200 In-house: NM	NM for both datasets	NM for both datasets

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

Liu et al.[34] (2023)	China	China (in-house: Guangzhou Twelfth People's Hospital)	50	Male: 25 Female: 25	NM Majority in the range of 6–15 years
Nguyen et al.[35] (2023)*	France	France (in-house: 4 unnamed French hospitals)	206	Male: 102 Female: 104	Males: 10.9 ± 3.7 years Females: 11 ± 3.7 years Range: 5–17 years
Wang et al.[36] (2023)	China	USA (RSNA)	200	Male: 100 Female: 100	NM
Yang et al.[37] (2023)	China	China (in-house: unnamed hospital in western China)	450	NM for test set	NM for test set
Zhang et al.[38] (2023)	China	China (in-house: Zhangzhou Affiliated Hospital of Fujian Medical University)	1,174	Male: 564 Female: 610	Range: 4–18 years, Central tendency NM
Liu et al.[39] (2024)	China	China (in-house: several unnamed hospitals)	744	Male: 378 Female: 366	NM
Nam et al.[40] (2024)*	South Korea	South Korea (in-house: Korea University Guro Hospital)	553	Male: 332 Female: 221	Mean: 9.8 ± 2.8 years Range: 4–17 years
Pape et al.[41] (2024)	Germany	Germany (in-house: unnamed German hospital)	1,253	NM explicitly Male: 55.7% Female: 44.3%	Median: 129.6 months Range: 100–155

DHA: University of South California Digital Hand Atlas; NM: not mentioned; RSNA: Radiological Society of North America bone age dataset; *included in the meta-analysis

Table 5. Technical aspects included studies

Study	Machine learning technique		Ground truth	Comparison	
Author (Year)	Algorithm (name)	BAA technique		Reader	BAA technique
Escobar et al.[9] (2019)	CNN (BoNet)	Novel method	Annotation by two expert radiologists using an unmentioned method	None	NA
Liu et al.[10] (2019a)	CNN (VGG-Net-16) + Non-Subsampled Contourlet Transform	Novel method	Mean bone age estimate by two expert radiologists from the DHA	Various models	Various techniques

Liu et al.[11] (2019b)	CNN (VGG-U-Net + cGAN)	Novel method	Labelling of bone age by experts	Different configurations of the model itself	Novel method
Booz et al.[12] (2020)*	CNN (BoneXpert v. 2.1)	Novel method	Independent analysis by two experienced pediatric radiologists	Three radiologists (1 board certified radiologist and two radiology residents)	GP
Dehghani et al.[13] (2020)*	Support vector machine	Local binary pattern	Mean of two radiologists' estimation from the USC hand atlas	None	NA
Guo et al.[14] (2020)*	CNN (BoNet+)	Novel method	Mean value of bone age estimation provided by two expert radiologists using an unmentioned method	Other models	Various techniques
Koitka et al.[15] (2020)	CNN (ResNet) + Regression network	Novel method	RSNA: according to RSNA DHA: according to DHA In-house: pediatric radiologist using GP method	None	NA
Reddy et al.[16] (2020)*	Region-based CNN (RetinaNet) for index finger identification CNN (Xception)	Novel method	Incorporation of RSNA and estimates by 4 pediatric radiologists using GP	Three pediatric radiologists, with one being a tiebreaker in cases of no consensus	GP
Zulkifley et al.[17] (2020)	CNN	Novel method	Weighted mean of an estimate from six medical practitioner, method not mentioned	Various deep learning networks	Various techniques
Cheng et al.[18] (2021)	CNN (incV2res-Net)	Novel method	The median bone age estimate of a panel of five professional pediatricians or radiologists using an unmentioned method	5-fold cross-validation	Not mentioned
He et al.[19] (2021)	CNN + regression (SE-ResNet)	Novel method	Labelled bone age from the RSNA dataset	Other models with various architecture	Various

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

Lee et al.[20] (2021)*	CNN (unnamed)	GP and modified TW3 hybrid	The average estimated GP bone age of a pediatric endocrinologist and two musculoskeletal radiologists	Two one-year fellow-ship-trained musculoskeletal radiologists	GP
Madan et al.[21] (2021)	Triplet loss (Triplet Network)	Novel method	Estimation by two pediatric radiologists using the GP method	None	NA
Cheng et al.[22] (2022)	Deep neural network (EFAI-BAA)	Novel method	None	Three qualified physicians from three different centers not associated with development of the model	GP
Hui et al.[23] (2022)	CNN (unnamed)	TW3	Chinese dataset: Average of two experts' readings independently using TW3	Other deep learning models (DenseNet, ResNet, VGGNet, BoNet, etc.)	Various techniques
Hwang et al.[24] (2022)	CNN (VUNO Med-BoneAge v. 1.1.0)	GP & novel method	CA	Independent estimation by two board-certi-fied pediatric radiologists	GP
Lea et al.[25] (2022)*	CNN ((VUNO Med-BoneAge v. 1.0.3)	GP	CA	A musculoskele-tal radiologist, a radiology resident, and a pediatric endocrinologist	GP
Zhang et al.[26] (2022)*	CNN (SMANet)	TW3-RUS and TW3-C	Agreement by 3 out of a panel of 5 experienced radiologists using TW3	Various deep learning models	Various techniques
Zhao et al.[27] (2022)*	CNN (Xception)	TW3-C	Consensus of three radiolo-gists and the maker of the China-05 Bone Age Standard when a consensus was not reached	Three radiolo-gists	TW3-C
Deng et al.[28] (2023)*	CNN (ResNet50, SENet, DenseN-et-121, EfficientNet-b4, and CSPNet)	Novel method	Labelled bone age from the RSNA dataset	Two other models (SIMBA and Chen et al.) and two radiologists with unmentioned method	Not mentioned

Kasani et al.[29] (2023)	CNN and regression-based method (MobileNetV2)	Novel method	Labelled bone age from the RSNA dataset	Other deep learning models	Various techniques
Kim et al.[30] (2023a)	CNN (no name)	Novel method	CA	Another model trained with Korean population (VUNO Med-BoneAge v. 1.1.0)	GP
Kim et al.[31] (2023b)*	CNN (MobileNetV2)	Novel method	Mean of GP bone age (estimated by two radiologists) and the chronological age	Different configurations of the model itself	Novel method
Kim et al.[32] (2023c)*	CNN (unnamed model, referred to as M1 in the study)	Novel method (GP-TW hybrid)	The mean of the bone age was independently estimated by three reviewers (a pediatric endocrinologist and two musculoskeletal radiologists) using GP method	Different configuration of the model (excluding carpal analysis, referred as M2 in the study)	Novel method (GP-TW hybrid)
Li et al.[33] (2023)	CNN (Inception V3 + Xception & ResNet50)	Novel method	RSNA: labeled bone age from the RSNA dataset CQJTJ: Bone age from clinical reports	Different configurations of the current model Other models from other studies	Various
Liu et al.[34] (2023)	CNN (Mask R-CNN & Xception)	Novel method	The average value of the predicted age of multiple experts	Different configurations of the model itself	Novel method
Nguyen et al.[35] (2023)*	CNN (ConvNeXt)	GP	Mean of GP bone age independently estimated by two board-certified pediatric radiologists	Senior general radiologist	GP
Wang et al.[36] (2023)	CNN (unnamed model)	Novel method	Labelled bone age from the RSNA dataset	Other models from other studies	Various
Yang et al.[37] (2023)	CNN (YOLOv5)	RUS-CHN	Three radiologists using RUS-CHN method independently	NanoDet PP-PicoDet (other models)	RUS-CHN

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

Zhang et al.[38] (2023)	CNN (PEARLS)	TW3-RUS	Average of the bone age estimates of the clinical radiology report and a pediatrician	Human assessment	TW3-RUS
Liu et al.[39] (2024)	CNN (unnamed feature pyramid objective detection CNN)	TW3-RUS and TW3-C	Average value of two to three radiologists for each picture	Thirty professional endocrinologists and radiologists, with each picture read by 3 reviewers, including at least one endocrinologist and one radiologist	TW3-RUS and TW3-C
Nam et al.[40] (2024)*	CNN (VUNO Med-BoneAge v. 1.0.3)	GP	CA	Pediatrician and musculoskeletal radiologist	GP
Pape et al.[41] (2024)	Not specified (IB Lab Panda v. 1.06)	GP	CA	None	NA

BAA: bone age assessment; CA: chronological age; CNN: convolutional neural network; GP: Greulich-Pyle; NA: not applicable; RUS-CHN: China 05 RUS-CHN; TW3-RUS: Tanner-Whitehouse 3 radius, ulna, short bones; TW3-C: Tanner-Whitehouse 3 carpal; *included in the meta-analysis

Table 6. Performance of included studies

Study	Performance
Escobar et al.[9] (2019)	MAE for RSNA = 3.85 months MAE for authors' group dataset = 6.86 months
Liu et al.[10] (2019a)	MAE 0.69 years ≈ 8.28 months
Liu et al.[11] (2019b)	MAE: 6.41 months
Booz et al.[12] (2020)*	AI vs ground truth MAE: 0.34 (95%CI 0.15–0.54) years ≈ 4.08 (1.8–6.48) months RMSE: 0.38 (95%CI 0.19–0.54) years ≈ 4.56 (2.28–6.48) months Reader vs. ground truth MAE: 0.79 (95%CI 0.61–0.96) years ≈ 9.48 (7.32–11.52) months RMSE: 0.89 (95%CI 0.59–1.23) years ≈ 10.68 (7.08–14.76) months ($p < 0.001$)
Dehghani et al.[13] (2020)*	MAE ± SD Male: 0.56 ± 0.49 years ≈ 6.72 ± 5.88 months Female: 0.55 ± 0.49 years ≈ 6.6 ± 5.88 months Combined mean: 6.66 ± 5.87 months
Guo et al.[14] (2020)*	MAE 0.762 ± 0.103 years ≈ 9.14 ± 1.24 months
Koitka et al.[15] (2020)	MAE RSNA: 4.56 months DHA: 11.58 months In-house: 7.55 months

Reddy et al.[16] (2020)*	MAE ± SD Whole hand Model: 4.7 ± 3.7 months Radiologist: 6.0 ± 5.5 months (p = 0.013) Index finger Model: 8.0 ± 7.5 months Radiologist: 5.1 ± 4.0 months (p = <0.0001)
Zulkifley et al.[17] (2020)	MAE: 8.2 months MSE: 121.902 months^2 RMSE: 11.04 months
Cheng et al.[18] (2021)	MAE 0.281 years ≈ 3.37 months MSE 0.203 years^2 ≈ 2.4 months^2
He et al.[19] (2021)	MAE 6.04 months
Lee et al.[20] (2021)*	Mean Bone Age (Model vs Ground truth) 11.35 ± 2.76 vs. 11.39 ± 2.74 years (p = 0.31) MAE 0.39 (95%CI 0.33–0.45) years ≈ 4.68 (95%CI 3.96–5.4) months
Madan et al.[21] (2021)	AUC Binary class (0–1, 1–2 years): 0.92 Multi-class (0–1, 1–2, 2–3 years): 0.82
Cheng et al.[22] (2022)	Concordance correlation coefficient between the model and readers from: - Kaohsiung Veterans General Hospital: 0.9828 (95% CI: 0.9790–0.9859, p = 0.6782 - Taichung Veterans General Hospital: 0.9739 (95% CI: 0.9681–0.9786, p = 0.0202 - Taipei Tzu Chi Hospital: 0.9592 (95% CI: 0.9501–0.9666, p = 0.4855)
Hui et al.[23] (2022)	MAE 0.444 years ≈ 5.328 months
Hwang et al.[24] (2022)	Modified model MAE: 11.31 months RMSE 14.48 months Radiologist 1 MAE: 13.09 months RMSE: 16.44 months Radiologist 2 MAE: 13.12 months RMSE: 16.54 months Radiologist 1 vs modified model MAE: p<0.001 Radiologist 2 vs modified model MAE: p<0.001
Lea et al.[25](2022)*	MAE Model vs CA: 11.06 months Model vs Reader 1: 7.21 months Model vs Reader 2: 7.88 months Model vs Reader 3: 10.06 months Combined mean (vs. Reader 1–3): 8.38 ± 1.22 months all p<0.025
Zhang et al.[26] (2022)*	MAE TW3-RUS: 0.43 ± 0.17 years ≈ 5.16 ± 2.04 months TW3-C: 0.45 ± 0.13 years ≈ 5.4 ± 1.56 months Combined mean: 5.28 ± 1.82 months
Zhao et al.[27](2022)*	MAE 0.2 ± 0.45 years ≈ 2.4 ± 5.4 months

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

Deng et al.[28] (2023)*	CSPNet (best performing CNN) Articular surface dataset MAE 7.34 ± 0.11 months RMSE 9.75 ±0.16 months Epiphysis dataset MAE 7.73 ± 0.13 months RMSE 9.95 ± 0.17months Combined mean (MAE): 7.54 ± 0.12 months
Kasani et al.[29] (2023)	MAE 3.84 months
Kim et al.[30] (2023a)	MAE Pusan/Yangsan: 8.2 months Dankook: 13.1 months Combined mean (MAE): not calculable due to lack of standard deviation
Kim et al.[31] (2023b)*	MAE 8.19 ± 6.85 months
Kim et al.[32] (2023c)*	MAE M1: 0.366 (95%CI 0.337–0.395) years ≈ 4.392 (95%CI 4.044–4.740) months M2: 0.388 (95%CI 0.358–0.418) years ≈ 4.656 (95%CI 4.296–5.016) months RMSE M1: 0.483 years ≈ 5.796 months M2: 0.505 years ≈ 6.06 months
Li et al.[33] (2023)	MAE RSNA: 5.45 months In-house: 3.34 months
Liu et al.[34] (2023)	MAE 5.4 months
Nguyen et al.[35] (2023)*	MAE AI vs comparator: 5.9 vs 8.7 months p<0.001
Wang et al.[36] (2023)	MAE 4.17 months p > 0.05
Yang et al.[37] (2023)	MAE: 0.35 years ≈ 4.2 months RMSE: 0.46 years ≈ 5.52 months RMS percentage error: 0.11 years ≈ 1.32 months
Zhang et al.[38] (2023)	Average MAE Female 0.46 years ≈ 5.52 months Male 0.45 yrs ≈ 5.4 months
Liu et al.[39] (2024)	AI vs reviewer TW3-RUS MAE: -0.072 vs 0.00 (p<0.001) RMSE: 0.52 vs 0.54 (p=0.02) Accuracy: 94.55 vs 92.34% TW3-C: MAE: -0.18 vs 0.00 (p<0.001) RMSE: 0.85 vs 0.78 (p<0.001) Accuracy: 80.38 vs 83.01%
Nam et al.[40] (2024)*	MAE Model: 8.5 ± 6.8 months Pediatrician: 2.6 ± 4.1 months Radiologist: 2.3 ± 3.4 months
Pape et al.[41] (2024)	MAE PA 11.1 months MAE Oblique 11.0 months

AI: artificial intelligence; AUC: area under the curve; MAE: mean absolute error; MSE: mean square error; RMSE; root mean square error; RSNA: Radiological Society of North America; SD: standard deviation; TW3-RUS: Tanner-Whitehouse 3 radius, ulna, short bones; TW3-C: Tanner-Whitehouse 3 carpal; *included in the meta-analysis

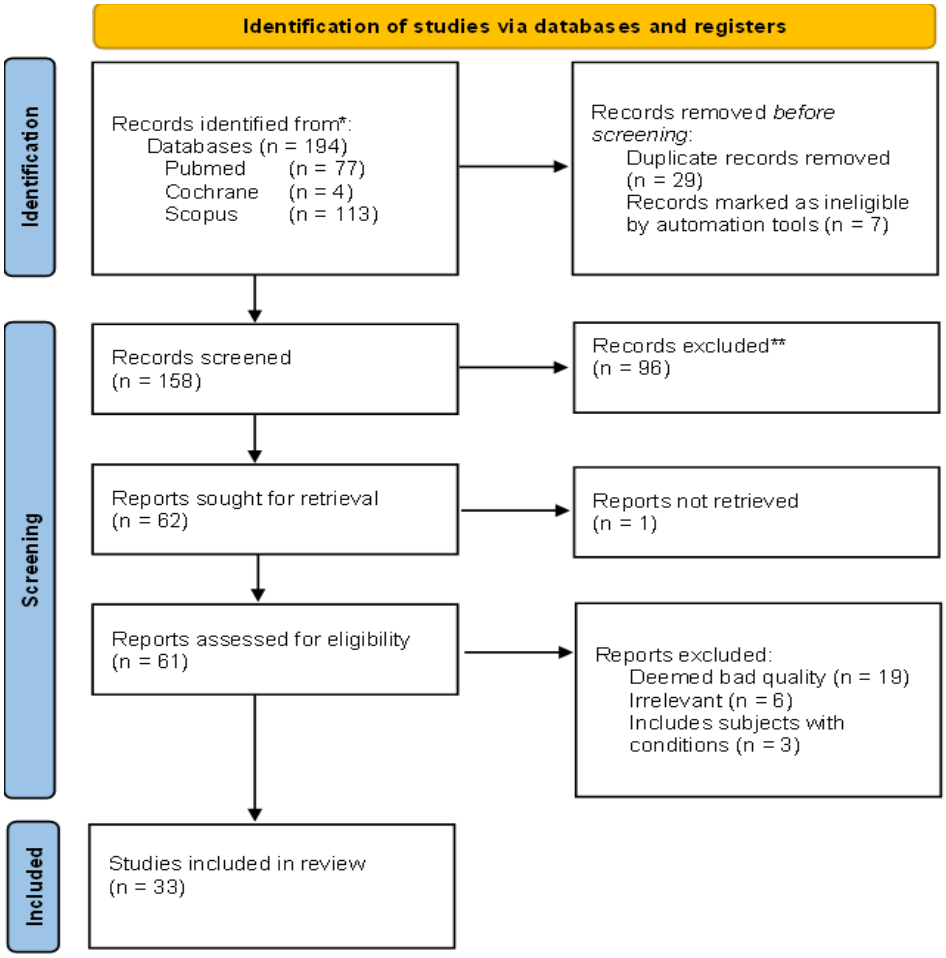

Figure 1: PRISMA flow diagram

rent study.

The previous study stated that the interests of AI and ML for the purpose of BAA is greatest in the United States and Western Europe, as evidenced by the creation of medical imaging databases and the RSNA pediatric bone age challenge.[5] The evidence gathered in this systematic review instead demonstrates a shift of trends with many studies being published by East Asian authors. While some studies employed a Western database as a testing set, most of the Asian studies used an in-house dataset comprising hand radiographs of local subjects. This would be beneficial to each of the respective authors' communities as it would increase their model's relevancy and validity in performing BAA. However, these datasets are probably not publicly available, unlike western datasets, as they belong to their respective institutions. Furthermore, the ethnic composition of the datasets is usually not reported. The development of a model based on the local population should not leave out the ethnic minorities also present in the community.

In the context of ML, the issue of multiethnicity in BAA is currently more important than accuracy. The capabilities of contemporary AI and ML BAA models are already accurate for clinical practice. The standard deviation for the BA of healthy children older than 3.5 years is greater than 5.4 months, and an error rate lower than this would not have a significant clinical impact. On the other hand, applying standards to subjects using models developed by people with a different ethnic background from different times would naturally influence the validity of the results.

Certainly, this challenge is also met by conventional

J Postgrad Med Inst 2025;39(1):242-46.
http://doi.org/10.54079/jpmi.39.1.3492

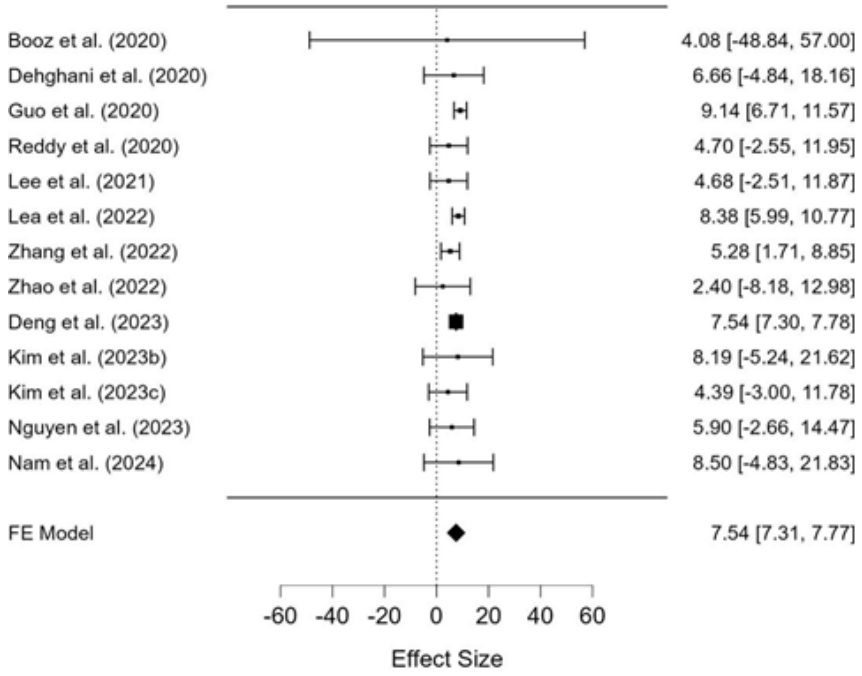

Figure 2: Forest plot of the meta-analysis

methods of BAA. A 2019 Meta-analysis highlighted the imprecision of GP atlas when used in the Asian male and African female population. This problem arises when bone age is used to determine the subject's CA, where the study showed that the BA of Asian males was delayed and advanced at certain ages. Meanwhile, the BA of African females were advanced compared to the GP atlas.[47] In another study comprising children from the Indian state of Uttar Pradesh, the GP-derived BA of male and female subjects was found to be retarded by around 9 months and 4 months than the chronological age, respectively.[48] Models trained using these standards along with human-annotated BA as ground truth would inadvertently "inherit" this issue.

As suggested by Rubin, AI models should no longer use BA as ground truth for training because it would only train the models to merely predicting the radiologist's estimation and not the best answer.[49] Instead, the CA may be more beneficial to be used as ground truth, possibly setting up new norms and advancements in the field of musculoskeletal radiology instead of limiting its potential. The radiologist's BA estimate is indeed crucial for the ground truth during the early days of BAA models. Nowadays, the BAA performance of AI models is even better than some radiologists. By using CA, AI models can go beyond mimicking radiologists and bring more clinical value beyond current capabilities, such as by predicting future health outcomes derived from radiographic findings.

Pan et al.[50] had previously demonstrated the ability of AI models in estimating CA from pediatric trauma hand radiographs. About 92–95% of the model's prediction was within 24 months of the CA and concordance was high between the AI models, two radiologist readers, and the chronological age. While their models demonstrated a systematic bias of overpredicting age for younger subjects, the study proved that CA estimation using AI models was feasible nevertheless.

Several of the included studies in this review also used CA for ground truth. Assuming all the subjects included in those studies have normal bone development, the models were then predicting the CA. The MAE reported from those studies ranged from 8.19–13.1 months.[24,25,30,40,41] The model which was tested by Nam et al.[40] on the Korean population had a MAE of 8.5 ± 6.8 months, which was quite far off from the radiologist's and the pediatrician's MAE of 2.3 ± 3.4 and 2.6 ± 4.1 months. It was also reported to have a low concordance rate of 58.8% with a 12-month cutoff. It is important to note that the model used the GP method for BAA.

Another issue is its laborious and time-consuming process. Even though the GP method is still considerably faster than the TW method, AI models are still more efficient in conducting BAA than human readers. Booz et al.[12] reported that their model analyzed 514 radiographs with a mean evaluation time of 21 seconds, compared to the average evaluation time of 165 seconds from three radiologists using the GP method. Additionally, the model was also faster than the mean reading time of the two radiologists whose estimates were used as the ground truth, which was 182 seconds. This translated to an 86.9% and 88.5% reduction in mean reading time, respectively. All the while having

a significantly lower MAE too (4.08 vs. 9.48) months.

Radiologists receiving assistance with GP BA estimates from AI exhibited lower MAE and lower interpretation time than their unassisted counterparts. The proportion of which the estimate had a difference of more than 12 months or more than 24 months from the ground truth was also lower in radiologists using AI.[51] The boost in MAE was also seen across radiologists of varying experience. However, radiologists may be prone to automation bias and sacrifice diagnostic accuracy for shorter evaluation time. This would be a challenge for a widespread implementation of AI assistance in clinical radiology practice. The overreliance on AI may lessened by controlling its mediators and implementing mitigation measure.[52] While occasional bias would be inevitable, its impact can be reduced by ensuring validity of the AI's BAA estimates.

An experiment by Yi et al. in 2021 intentionally fed both appropriate and inappropriate data inputs to a DCNN BAA model. The appropriate inputs were left-hand radiographs, while inappropriate inputs consisted of radiological (chest radiographs) and non-radiological (image of street numbers) images. The model previously won the 2017 RSNA pediatric bone age challenge with a 0.99 concordance rate with the ground truth. Interestingly, the model failed to distinguish between the inputs and even calculate bone age for inappropriate images.[53] This outcome underscores the current limitation of a fully automated total BAA system and highlights the importance of data supervision and verification to guarantee valid and reproducible results.

With the increasing number of studies from various countries, it would be ideal for a multinational-multicenter collaboration to be held. Subsequently, a digital database of radiographs from subjects of various ages, ethnicity, geography, and socioeconomic background can be generated in a relatively short amount of time. The pooled radiographs can then be used as a training dataset, producing a versatile and capable model.

Several limitations exist within the current study. The dissimilar age range and a variable data set size render an analysis of the model's performance in conducting BAA to be a challenge. Furthermore, studies also struggle with incomplete reporting of test set size and the proportion of their subjects' age and sex. The multi-ethnic nature and socioeconomic factors underlying the datasets also hindered clear conclusions to be drawn.

Conclusion

AI and ML models continue to demonstrate rapid advancements in their application for BAA. This study described the current trends in ML research and explored ongoing obstacles in BAA, as well as the prospective role of AI. The pooled MAE of models published in 2019–2024 was 7.54 months, which is lower than that reported in the previous study. While promising, further research is still required to address current limitations, such as validity issues. Subsequent studies should also be conducted with rigorous methodology and thorough reporting.

References

1. Cavallo F, Mohn A, Chiarelli F, Giannini C. Evaluation of Bone Age in Children: A Mini-Review. Front Pediatr. 2021 Mar 12;9.
2. Franklin D, Flavel A, Noble J, Swift L, Karkhanis S. Forensic age estimation in living individuals: methodological considerations in the context of medico-legal practice. Research and Reports in Forensic Medical Science. 2015 Oct;53.
3. Prokop-Piotrkowska M, Marszałek-Dziuba K, Moszczyńska E, Szalecki M, Jurkiewicz E. Traditional and New Methods of Bone Age Assessment-An Overview. J Clin Res Pediatr Endocrinol. 2021 Aug 23;13(3):251–62.
4. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology. 2019 Feb;290(2):498–503.
5. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. Stoean R, editor. PLoS One. 2019 Jul 25;14(7):e0220242.
6. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. The BMJ. 2021;372:n71.
7. Kitchenham B, Charters SM. Guidelines for performing systematic literature reviews in software engineering (Version 2.3). Software Engineering Technical Report EBSE-2007-0. 2007. Available from: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf.
8. Dallora AL. machine learning techniques in the automated age assessment of youthInvestigating the employment of. Karlskrona: Blekinge Institute of Technology; 2019. p. 6.
9. Escobar M, González C, Torres F, Daza L, Triana G, Arbeláez P. Hand Pose Estimation for Pediatric Bone Age Assessment. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Shenzhen: Springer; 2019. p. 531–9.
10. Liu Y, Zhang C, Cheng J, Chen X, Wang ZJ. A multi-scale data fusion framework for bone age assessment with convolutional neural networks. Comput Biol Med. 2019 May;108:161–73.
11. Liu B, Zhang Y, Chu M, Bai X, Zhou F. Bone Age Assessment Based on Rank-Monotonicity Enhanced Ranking CNN. IEEE Access. 2019;7:120976–83.
12. Booz C, Yel I, Wichmann JL, Boettger S, Al Kamali A, Albrecht MH, et al. Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated

algorithm compared to the Greulich-Pyle method. Eur Radiol Exp. 2020 Dec 28;4(1):6.

13. Dehghani F, Karimian A, Sirous M. Assessing the Bone Age of Children in an Automatic Manner Newborn to 18 Years Range. J Digit Imaging. 2020 Apr 6;33(2):399–407.
14. Guo J, Zhu J, Du H, Qiu B. A bone age assessment system for real-world X-ray images based on convolutional neural networks. Computers & Electrical Engineering. 2020 Jan;81:106529.
15. Koitka S, Kim MS, Qu M, Fischer A, Friedrich CM, Nensa F. Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. Med Image Anal. 2020 Aug;64:101743.
16. Reddy NE, Rayan JC, Annapragada A V., Mahmood NF, Scheslinger AE, Zhang W, et al. Bone age determination using only the index finger: a novel approach using a convolutional neural network compared with human radiologists. Pediatr Radiol. 2020 Apr 20;50(4):516–23.
17. Zulkifley MA, Abdani SR, Zulkifley NH. Automated Bone Age Assessment with Image Registration Using Hand X-ray Images. Applied Sciences. 2020 Oct 16;10(20):7233.
18. Cheng CF, Huang ET-C, Kuo J-T, Liao KY-K, Tsai F. Report of Clinical Bone Age Assessment using Deep Learning for an Asian population in Taiwan. Biomedicine (Taipei). 2021 Aug 25;11(3):50–8.
19. He J, Jiang D. Fully Automatic Model Based on SE-ResNet for Bone Age Assessment. IEEE Access. 2021;9:62460–6.
20. Lee K-C, Lee K-H, Kang CH, Ahn K-S, Chung LY, Lee J-J, et al. Clinical Validation of a Deep Learning-Based Hybrid (Greulich-Pyle and Modified Tanner-Whitehouse) Method for Bone Age Assessment. Korean J Radiol. 2021;22(12):2017.
21. Madan S, Gandhi T, Chaudhury S. Bone Age Assessment for Lower Age Groups Using Triplet Network in Small Dataset of Hand X-Rays. In: Singh M, Kang D-K, Lee J-H, Tiwary US, Singh D, Chung W-Y, editors. Intelligent Human Computer Interaction 2020. Daegu: Springer; 2021. p. 142–53.
22. Cheng C-F, Liao KY-K, Lee K-J, Tsai F-J. A Study to Evaluate Accuracy and Validity of the EFAI Computer-Aided Bone Age Diagnosis System Compared With Qualified Physicians. Front Pediatr. 2022 Apr 8;10.
23. Hui Q, Wang C, Weng J, Chen M, Kong D. A Global-Local Feature Fusion Convolutional Neural Network for Bone Age Assessment of Hand X-ray Images. Applied Sciences. 2022 Jul 18;12(14):7218.
24. Hwang J, Yoon HM, Hwang J-Y, Kim PH, Bak B, Bae BU, et al. Re-Assessment of Applicability of Greulich and Pyle-Based Bone Age to Korean Children Using Manual and Deep Learning-Based Automated Method. Yonsei Med J. 2022;63(7):683.
25. Lea WW, Hong S-J, Nam H-K, Kang W-Y, Yang Z-P, Noh E-J. External validation of deep learning-based bone-age software: a preliminary study with real world data. Sci Rep. 2022 Jan 24;12(1):1232.
26. Zhang Y, Zhu W, Li K, Yan D, Liu H, Bai J, et al. SMANet: multi-region ensemble of convolutional neural network model for skeletal maturity assessment. Quant Imaging Med Surg. 2022 Jul;12(7):3556–68.
27. Zhao X, Zhang M, Cheng M, Yue X, Li W, Li C. Construction of artificial intelligence system of carpal bone age for Chinese children based on China 05 standard. Med Phys. 2022 May;49(5):3223–32.
28. Deng Y, Chen Y, He Q, Wang X, Liao Y, Liu J, et al. Bone age assessment from articular surface and epiphysis using deep neural networks. Mathematical Biosciences and Engineering. 2023;20(7):13133–48.
29. Kasani AA, Sajedi H. Hand bone age estimation using divide and conquer strategy and lightweight convolutional neural networks. Eng Appl Artif Intell. 2023 Apr;120:105935.
30. Kim PH, Yoon HM, Kim JR, Hwang J-Y, Choi J-H, Hwang J, et al. Bone Age Assessment Using Artificial Intelligence in Korean Pediatric Population: A Comparison of Deep-Learning Models Trained With Healthy Chronological and Greulich-Pyle Ages as Labels. Korean J Radiol. 2023;24(11):1151.
31. Kim KD, Kyung S, Jang M, Ji S, Lee DH, Yoon HM, et al. Enhancement of Non-Linear Deep Learning Model by Adjusting Confounding Variables for Bone Age Estimation in Pediatric Hand X-rays. J Digit Imaging. 2023 Jun 2;36(5):2003–14.
32. Kim S-U, Oh S, Lee K-H, Kang CH, Ahn K-S. Improvement of Bone Age Assessment Using a Deep Learning Model in Young Children: Significance of Carpal Bone Analysis. Iranian Journal of Radiology. 2023 Jul 18;20(2).
33. Li Z, Chen W, Ju Y, Chen Y, Hou Z, Li X, et al. Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. Front Artif Intell. 2023 Mar 2;6.
34. Liu Z-Q, Hu Z-J, Wu T-Q, Ye G-X, Tang Y-L, Zeng Z-H, et al. Bone age recognition based on mask R-CNN using xception regression model. Front Physiol. 2023 Feb 14;14.
35. Nguyen T, Hermann A-L, Ventre J, Ducarouge A, Pourchot A, Marty V, et al. High performance for bone age estimation with an artificial intelligence solution. Diagn Interv Imaging. 2023 Jul;104(7–8):330–6.
36. Wang C, Wu Y, Wang C, Zhou X, Niu Y, Zhu Y, et al. Attention-based multiple-instance learning for Pediatric bone age assessment with efficient and interpretable. Biomed Signal Process Control. 2023 Jan;79:104028.
37. Yang C, Dai W, Qin B, He X, Zhao W. A real-time automated bone age assessment system based on the RUS-CHN method. Front Endocrinol (Lausanne). 2023 Mar 15;14.
38. Zhang D, Liu B, Huang Y, Yan Y, Li S, He J, et al. An Automated TW3-RUS Bone Age Assessment Method with Ordinal Regression-Based Determination of Skeletal Maturity. J Digit Imaging. 2023 Feb 22;36(3):1001–15.
39. Liu Y, Ouyang L, Wu W, Zhou X, Huang K, Wang Z, et al. Validation of an established TW3 artificial intelligence bone age assessment system: a prospective, multicenter, confirmatory study. Quant Imaging Med Surg. 2024 Jan;14(1):144–59.
40. Nam H-K, Lea WW-I, Yang Z, Noh E, Rhie Y-J, Lee K-H, et al. Clinical validation of a deep-learning-based bone age software in healthy Korean children. Ann Pediatr Endo-

crinol Metab. 2024 Apr 30;29(2):102–8.

41. Pape J, Hirsch FW, Deffaa OJ, DiFranco MD, Rosolowski M, Gräfe D. Applicability and robustness of an artificial intelligence-based assessment for Greulich and Pyle bone age in a German cohort. RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren. 2024 Jun 8;196(06):600–6.
42. Mader KS. RSNA Bone Age [Internet]. 2017 [cited 2024 Aug 10]. Available from: https://www.kaggle.com/datasets/kmader/rsna-bone-age
43. Image Processing and Informatics Lab. Digital Hand Atlas [Internet]. 2013 [cited 2024 Aug 10]. Available from: https://ipilab.usc.edu/research/baaweb/
44. Rana M, Bhushan M. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimed Tools Appl. 2023 Jul 24;82(17):26731–69.
45. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. RadioGraphics. 2017 Mar;37(2):505–15.
46. Sarvamangala DR, Kulkarni R V. Convolutional neural networks in medical image understanding: a survey. Evol Intell. 2022 Mar 3;15(1):1–22.
47. Alshamrani K, Messina F, Offiah AC. Is the Greulich and Pyle atlas applicable to all ethnicities? A systematic review and meta-analysis. Eur Radiol. 2019 Jun 7;29(6):2910–23.
48. Tiwari PK, Gupta M, Verma A, Pandey S, Nayak A. Applicability of the Greulich–Pyle Method in Assessing the Skeletal Maturity of Children in the Eastern Utter Pradesh (UP) Region: A Pilot Study. Cureus. 2020 Oct 10;12(10):e10880.
49. Rubin DA. Assessing Bone Age: A Paradigm for the Next Generation of Artificial Intelligence in Radiology. Radiology. 2021 Dec;301(3):700–1.
50. Pan I, Baird GL, Mutasa S, Merck D, Ruzal-Shapiro C, Swenson DW, et al. Rethinking Greulich and Pyle: A Deep Learning Approach to Pediatric Bone Age Assessment Using Pediatric Trauma Hand Radiographs. Radiol Artif Intell. 2020 Jul 1;2(4):e190198.
51. Eng DK, Khandwala NB, Long J, Fefferman NR, Lala S V., Strubel NA, et al. Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial. Radiology. 2021 Dec;301(3):692–9.
52. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. 2012 Jan 1;19(1):121–7.
53. Yi PH, Arun A, Hafezi-Nejad N, Choy G, Sair HI, Hui FK, et al. Can AI distinguish a bone radiograph from photos of flowers or cars? Evaluation of bone age deep learning model on inappropriate data inputs. Skeletal Radiol. 2022 Feb 5;51(2):401–6.

Authors' Contribution Statement

SS contributed to the conception, design, acquisition, analysis, interpretation of data, drafting of the manuscript, and final approval of the version to be published. R contributed to the design, acquisition, analysis, interpretation of data, drafting of the manuscript, and critical review of the manuscript. All authors are accountable for their work and ensure the accuracy and integrity of the study.

Confilct of Interest

Authors declared no conflict on interest

Grant Suppport and Financial Disclosure

None

Data Sharing Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.