Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical transfer learning and region of interest standardization for bone age assessment in Thai children

Pitchaya Wiratchotisatian
Faculty of Science, Khon Kaen University
Khwankaow Tangprasert
Faculty of Engineering, Khon Kaen University
Sarun Paisarnsrisomsuk
Faculty of Engineering, Khon Kaen University
Panawit Hanpinitsak
panaha@kku.ac.th
Faculty of Engineering, Khon Kaen University
Wichuda Chaisiwamongkol
Faculty of Science, Khon Kaen University
Ratikorn Chaisiwamongkol
Faculty of Medicine, Khon Kaen University
Nipaporn Tewattanarat
Faculty of Medicine, Khon Kaen University
Chanakarn Poonpol
Faculty of Science, Khon Kaen University
Chatparin Pansukrada
Faculty of Science, Khon Kaen University
Wilairat Thaowandee
Faculty of Medicine, Khon Kaen University

Research Article

Keywords: Bone age assessment, Deep learning, Artificial intelligence, Pediatric radiology, Hand X-ray, Transfer learning

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11101710/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hierarchical transfer learning and region of interest standardization for bone age assessment in Thai children

Pitchaya Wiratchotisatian[1], Khwankaow Tangprasert[2], Sarun Paisarnsrisomsuk[2], Panawit Hanpinitsak[2*], Wichuda Chaisiwamongkol[1], Ratikorn Chaisiwamongkol[3], Nipaporn Tewattanarat[4], Chanakarn Poonpol[1], Chatparin Pansukrada[5], Wilairat Thaowandee[3]

[1]Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand.
[2]Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, Thailand.
[3]Department of Pediatrics, Faculty of Medicine, Khon Kaen University, Khon Kaen, 40002, Thailand.
[4]Department of Radiology, Faculty of Medicine, Khon Kaen University, Khon Kaen, 40002, Thailand.
[5]Information Technology Workgroups, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand.

*Corresponding author(s). E-mail(s): panaha@kku.ac.th;

Abstract

Bone age assessment (BAA) from hand radiographs is routinely used to evaluate child growth and endocrine disorders, but manual atlas-based reading is slow and reader-dependent, while deep learning models trained on large international datasets may lose accuracy in other populations. We developed a population-specific AI framework for BAA in Thai children. It combines ROI standardization using contrast enhancement, automatic hand detection, and segmentation with an EfficientNetB7 regression model trained through hierarchical transfer learning. The model first learned skeletal-maturity representations from the public RSNA and DHA datasets and was then fine-tuned on Thai hospital data. On a held-out Thai test set excluded from BAA training and validation, hierarchical

adaptation achieved a mean absolute error (MAE) of 4.64 months (RMSE, 5.85 months; $R^2 = 0.98$), compared with 5.73 months for pooled training and 7.60 months for the public-only baseline. The full preprocessing pipeline also reduced downstream MAE relative to raw radiographs. It met the predefined detection-and-segmentation processing criterion for all evaluated RSNA and DHA images; however, this finding represents processing completeness rather than independent quantitative validation of segmentation accuracy. Among GP, TW, and a hybrid strategy using TW when applicable and GP outside the validated TW3-RUS range, the hybrid produced the lowest point-estimate MAE, although its advantage over TW was small. Grad-CAM provided exploratory evidence of attention to anatomically relevant regions but also identified localized or non-skeletal attention. Improved RSNA hand ROI annotations are released to support future research.

Keywords: Bone age assessment, Deep learning, Artificial intelligence, Pediatric radiology, Hand X-ray, Transfer learning

1 Introduction

Bone age assessment (BAA) is a fundamental diagnostic tool in pediatric medicine for evaluating growth disorders, endocrine abnormalities, and conditions such as precocious puberty [6, 38]. By comparing skeletal maturity from hand and wrist radiographs to chronological age, clinicians can assess developmental progression and guide treatment decisions. Traditional methods, including the Greulich–Pyle (GP) atlas [13] and the Tanner–Whitehouse (TW) scoring system [49], remain widely used in clinical practice. However, these approaches are time-consuming, observer-dependent, and susceptible to inter- and intra-rater variability [6, 38], which may lead to inconsistent diagnoses and delayed clinical decision-making [7, 52].

To address these limitations, automated bone age assessment systems based on artificial intelligence (AI) and deep learning have been increasingly explored to provide faster, more consistent, and reproducible skeletal maturity estimation. The emergence of large public datasets, particularly the 2017 RSNA Pediatric Bone Age Challenge dataset [15, 39] and the Digital Hand Atlas (DHA) [12], catalyzed rapid progress in deep learning-based BAA. Early studies employed convolutional neural networks (CNNs) with transfer learning from natural image datasets [22, 23, 46]. In particular, Larson et al. [22] reported performance comparable to expert radiologists in a head-to-head evaluation. Subsequent research explored more advanced modeling strategies including ensemble learning [35], region-aware pipelines [4, 20], and segmentation-assisted preprocessing [31].

Commercial systems such as BoneXpert have demonstrated high consistency and practical utility in automated skeletal maturity estimation [53, 54]. Independent clinical studies further showed that automated BoneXpert analysis can achieve high agreement with expert reference ratings while substantially reducing interpretation

This study was funded by the Digital Economy and Society Fund, Thailand (Grant number: PR66010158).
Corresponding author: Panawit Hanpinitsak (e-mail: panaha@kku.ac.th).

time in routine practice [3]. More broadly, AI-assisted workflows have been shown to improve radiologist performance and reduce interpretation time while decreasing inter-rater variability [9, 47, 56].

Despite these advances, increasing evidence indicates that high benchmark accuracy does not guarantee reliable real-world deployment. One important reason is that skeletal maturation patterns vary across ethnic and demographic groups [34, 45], meaning that models trained on radiographs from one population may not generalize reliably to others. Consistent with this concern, models trained primarily on the RSNA dataset have been shown to exhibit demographic bias and clinically significant error variation when evaluated on external cohorts [1]. More broadly, deep learning systems in medical imaging may inadvertently exploit site-specific signals present in the training data, resulting in strong internal performance but degraded cross-institution generalization [61].

In response, several strategies have been proposed to mitigate cross-population domain shift. Post-hoc calibration techniques can partially correct systematic prediction bias when models are applied to geographically distinct cohorts without retraining the full network [42]. Multi-dataset training that incorporates local cohorts has also been explored to improve population-specific accuracy [19, 28, 30]. Nevertheless, systematic evaluation of hierarchical transfer strategies—progressively adapting representations learned from large public datasets to underrepresented local populations—remains limited, particularly in Southeast Asian pediatric cohorts.

In parallel, radiograph heterogeneity presents an additional deployment barrier. Clinical hand radiographs frequently include elongated forearm regions, embedded markers, and background artifacts that may inadvertently influence model predictions. While detection- or segmentation-based region-of-interest (ROI) extraction is commonly incorporated in multi-stage pipelines [4, 11, 16, 20, 31, 58], preprocessing choices are typically evaluated through limited ablation or treated as implementation details rather than systematically studied as contributors to prediction robustness. Differences in mask coverage, region-of-interest framing, and artifact suppression therefore remain insufficiently characterized. In particular, the contribution of preprocessing design choices—such as anatomical region detection before segmentation—to ROI quality and downstream bone age estimation has rarely been evaluated within a unified framework.

Finally, most deep learning studies adopt a single reference standard, typically GP-based ratings, although alternative standard-specific automated systems have also been reported, including TW3-based and RUS-CHN-based frameworks [59, 62]. A smaller but important line of work has also questioned the default reliance on GP-derived supervision. For example, Pan et al. [36] trained a deep learning model using chronological age from pediatric trauma radiographs and showed better performance than a GP-based deep learning comparator in an external U.S. trauma cohort, while Kim et al. [19] reported that a Korean model trained with chronological-age labels outperformed a GP-based model in Korean pediatric cohorts. In parallel, studies comparing reference standards directly have shown that standard choice itself can be clinically important. For example, Wang et al. [57] compared GP and TW3 assessment in Taiwanese children and found that TW3 was more accurate overall, whereas

GP retained an advantage in later adolescence. Given known differences between GP and TW scoring philosophies [38], and documented population-dependent variation in skeletal maturation [34, 45], supervision choice may materially affect learned representations and generalization performance. However, systematic comparison of GP-, TW-, and hybrid-label supervision within a unified population-specific adaptation framework remains scarce.

Motivated by these gaps, we evaluate a practical BAA framework for Thai pediatric radiographs. Rather than presenting the individual components as novel algorithms, the study focuses on their integration and evaluation under three applied questions: target-population adaptation, ROI standardization, and sensitivity to the reference-label strategy. The contributions of this study are summarized as follows:

1. **Population-Specific Hierarchical Transfer Learning:** We propose a two-stage transfer strategy that first leverages large public datasets (RSNA and DHA) for representation learning and subsequently fine-tunes on Thai clinical data. The approach is evaluated under a domain-shift-aware protocol using a strictly held-out Thai test set and compared against public-only and pooled-training baselines.
2. **ROI Standardization Pipeline Evaluated for Aggregate Effect and Cross-Dataset Processing Completeness:** We evaluate a preprocessing pipeline consisting of CLAHE enhancement followed by hand detection and hand segmentation, designed to reduce long-forearm inclusion and background artifacts, and quantify its aggregate impact on downstream MAE relative to raw input. We additionally report a predefined processing-completeness criterion across the RSNA and DHA datasets. This full-dataset completeness analysis is treated as an operational check and not as a substitute for quantitative segmentation validation against independent manual ground truth.
3. **Reference-Label Sensitivity Analysis:** We compare GP, TW, and a hybrid reference strategy within the Thai cohort. The hybrid target is defined as a rule-based substitution strategy rather than a statistical combination of two independent labels: TW is used when applicable, and GP is substituted for cases outside the validated TW3-RUS range.
4. **Reproducible Resource Contribution:** We release improved RSNA hand ROI annotations derived from our pipeline, positioned relative to existing public mask resources to support reproducible preprocessing in future BAA research, along with open-source implementations of both the preprocessing framework and the bone age prediction model.

By integrating population-specific adaptation, ROI standardization, and reference-label sensitivity analysis within a single framework, this study provides an applied evaluation of BAA in an underrepresented target population while explicitly separating predictive performance from preprocessing-completeness claims.

2 Literature Review and Related Works

Automated bone age assessment has evolved through several methodological paradigms, ranging from rule-based skeletal modeling to modern deep learning frameworks. In this section, we categorize prior work according to modeling strategy, ROI handling, and adaptation methodology, and position the present study within this landscape.

2.1 Rule-Based and Early Automated Systems

Before deep learning, automated systems such as BoneXpert relied on explicit bone modeling and geometric feature extraction to estimate skeletal maturity [33, 53, 54]. These systems achieved clinically acceptable performance and were validated in real-world settings, but required carefully engineered bone segmentation and predefined scoring logic. While robust and clinically validated, such systems rely on engineered features and explicit bone modeling, which may limit their flexibility when adapting to new populations or imaging conditions.

2.2 End-to-End Deep Learning Models

The availability of large annotated datasets, such as the Digital Hand Atlas (DHA) [12] and the RSNA Pediatric Bone Age Challenge dataset [15], enabled the rapid adoption of deep learning for automated bone age assessment. Early convolutional neural network (CNN) approaches demonstrated that transfer learning from natural image datasets could achieve competitive performance in predicting skeletal maturity from hand radiographs.

Spampinato et al. [46] proposed the BoNet framework, which combines transfer learning with task-specific feature learning by initializing early layers with ImageNet-pretrained weights and training higher-level layers on radiographic images. Similarly, Larson et al. [22] trained a deep residual network using clinical radiographs and demonstrated performance comparable to expert radiologists in estimating bone age.

Beyond individual architectures, Pan et al. [35] investigated ensemble learning strategies using models submitted to the RSNA Pediatric Bone Age Challenge and showed that combining heterogeneous models can further reduce prediction error compared with single-model systems.

Despite these advances, most end-to-end approaches treat the radiograph as a global input and directly regress bone age without explicitly standardizing the anatomical region of interest. While effective under curated benchmark settings, such designs assume similar imaging conditions between training and test data, which may limit their robustness in heterogeneous clinical environments.

2.3 ROI-Driven and Multi-Stage Pipelines

To better mimic the radiologist workflow, several studies introduced region-aware modeling or explicit preprocessing prior to bone age prediction. Early attempts incorporated spatial attention mechanisms within convolutional networks to highlight anatomically informative regions without explicit cropping. For example, Ren et

al. [43] proposed an attention-guided regression CNN that learns spatial importance maps during training, enabling the network to emphasize relevant skeletal regions while still processing the radiograph as a global input.

Subsequent work moved toward explicit anatomical ROI localization before age estimation. Bui et al. [4] incorporated Faster R-CNN to detect TW3 anatomical bones (e.g., phalanges and radius), followed by CNN-based maturity classification for each detected ROI. Similarly, Koitka et al. [20] localized multiple anatomical regions using Faster R-CNN and predicted bone age for each ROI using dedicated CNN regressors, with the final estimate obtained by aggregating predictions across regions. Detector-based pipelines have since been extended to finer-grained anatomical localization and multi-stage feature reasoning frameworks [8, 14, 32, 58].

Other approaches aim to reduce the need for explicit ROI annotation while preserving interpretability. For example, Liu et al. [29] explored multi-scale feature fusion from whole-hand radiographs, while Liu et al. [27] proposed a self-supervised attention mechanism that automatically discovers informative bone parts using weak image-level supervision. Similarly, Li et al. [26] introduced an annotation-free cascaded framework that sequentially localizes discriminative bone regions before age estimation. Additional studies have explored graph-based aggregation or prior-guided feature extraction to exploit anatomical structure more explicitly [17, 24, 25].

A complementary line of work focuses on standardizing the whole-hand ROI through segmentation and geometric normalization. Lee et al. [23] employed a patch-based pixel classification approach to generate hand masks and suppress background artifacts in radiographs. Iglovikov et al. [16] proposed a pipeline combining U-Net–based hand segmentation, keypoint-driven image registration, and region-specific CNN models, demonstrating that different hand regions contribute differently to skeletal maturity prediction. Gao et al. [11] likewise incorporated segmentation explicitly into a deep CNN-based BAA pipeline, while Liu et al. [31] applied Mask R-CNN to segment the hand prior to regression using an Xception-based network. More recently, Rassmann et al. [41] introduced Deeplasia, an open-source bone age assessment system that integrates background masking with RSNA-trained deep learning models and evaluates generalizability on external cohorts.

Although many of these approaches demonstrate that preprocessing can improve prediction performance, preprocessing pipelines are typically treated as fixed components or evaluated only through limited ablation experiments. Consequently, the impact of ROI framing, mask coverage, and artifact suppression on downstream bone age estimation remains insufficiently characterized. In particular, the potential benefit of combining anatomical region detection to standardize the field of view with subsequent segmentation to refine hand boundaries has not been systematically investigated.

2.4 Generalization, Bias, and Population Adaptation

Recent work emphasizes that strong benchmark performance does not guarantee robustness across institutions or demographic groups. Studies have reported demographic bias and clinically meaningful error disparities when models trained on

RSNA data are evaluated on external cohorts [1]. More broadly, the medical AI literature highlights the importance of transparent evaluation of bias and cross-site generalization for safe clinical deployment [21, 61].

Several strategies have been proposed to mitigate domain shift. Post-hoc population-specific calibration can reduce systematic prediction bias without retraining the base network [42]. Multi-dataset training incorporating local cohorts has also been explored to improve performance in specific ethnic populations [28]. Related population-specific studies have further shown that local demographic context and supervision choice can affect automated BAA performance, supporting the need for validation beyond the benchmark population [19]. However, hierarchical transfer strategies that progressively adapt representations from large public datasets to underrepresented clinical cohorts remain relatively unexplored.

2.5 Position of the Present Study

To contextualize the methodological contributions of this work, Table 1 summarizes representative deep learning approaches with respect to three aspects: (1) whether explicit ROI processing is employed, (2) whether cross-dataset domain shift is evaluated, and (3) supervision strategy used. In summary, prior work has demonstrated: (i) strong benchmark performance of deep CNNs, (ii) effectiveness of ROI-driven pipelines, and (iii) growing awareness of domain shift and demographic bias.

A smaller but important line of work has additionally shown that supervision design itself can vary, including chronological-age supervision and standard-specific automated systems based on TW3 or RUS-CHN scoring frameworks [19, 36, 59, 62]. Population-specific comparisons between GP- and TW3-based assessment suggests that supervision strategy deserves more explicit study in clinically deployed BAA systems [57]. However, to the best of our knowledge, a unified framework that simultaneously (1) quantifies cross-population domain shift using a held-out target cohort, (2) quantifies the aggregate contribution of ROI standardization through MAE-based ablation, and (3) compares multiple supervision standards within a hierarchical adaptation strategy has not been comprehensively investigated. The present study addresses these gaps in the context of Thai pediatric radiographs.

Table 1 Positioning of the proposed framework relative to representative prior bone age assessment studies, compared across ROI processing, domain-shift evaluation, and supervision strategy.

Study	Dataset	ROI Processing	Domain Shift Eval.	Supervision Strategy
Foundational deep learning approaches				
[46]	DHA	×	×	GP/TW framing
[23]	Institutional cohort	Partial (mask preprocessing)	×	GP
[22]	Clinical cohort + DHA test	×	Limited	GP
[43]	RSNA+Private	×	×	GP
[35]	RSNA	Not specified (ensemble models)	×	GP
ROI-focused approaches				
[4]	DHA	✓(TW3 anatomical ROIs)	×	TW3
[20]	RSNA	✓(TW3 anatomical ROIs)	×	GP
[27]	RSNA	✓(weakly annotated region attention)	×	GP
[26]	RSNA + private	✓(annotation-free critical bone extraction)	×	GP
[31]	Local dataset	✓(hand masking)	×	Expert-average
[14]	Chinese multicenter (100)	✓(epiphyseal region)	×	Chinese 05 standard
Domain-shift and generalization studies				
[1]	RSNA + DHA	Not specified	✓(bias analysis)	GP
[41]	RSNA+external cohorts	✓(hand masking)	✓(external cohort)	GP
[42]	RSNA + local cohort	✓(hand masking)	✓(calibration)	GP
[28]	RSNA + RHPE + Tibetan	✓(hand/wrist detect + keypoints)	✓(external Tibetan)	GP
Supervision comparison studies				
[36]	CHONY + HCH test	Partial (hand crop/standardize)	×	Chrono. age vs GP
[19]	Korean + 2 ext. cohorts	×	Limited (ext. institutional)	Chrono. age vs GP model
[57]	Taiwanese children	Not specified (BoneXpert)	×	GP / TW3
This Work	RSNA + DHA + Thai	**✓(CLAHE, detect, then segment)**	**✓(held-out Thai test)**	**GP / TW3 / Hybrid**

3 Materials and Methods

3.1 Study Overview

This study develops and evaluates a BAA framework for Thai pediatric hand radiographs, structured around three components that are introduced in this section and then evaluated in Section 4. First, Section 3.3 describes an image preprocessing and ROI standardization pipeline that combines contrast enhancement (CLAHE), hand detection, and hand segmentation to produce a consistently framed, artifact-suppressed input image. Second, Section 3.4 describes a bone age prediction model

built on an EfficientNetB7 backbone, trained using the hierarchical transfer learning strategy introduced in Section 3.5, which progressively adapts from the RSNA and DHA public datasets to the Thai clinical cohort. Third, Section 3.6 describes an experimental design that evaluates this framework along three axes: a domain-shift-aware comparison against public-only and pooled-training baselines, an ablation of the preprocessing pipeline's contribution, and a sensitivity analysis across three supervision standards, namely GP, TW, and hybrid TW+GP. Datasets used throughout are described immediately below, and the evaluation metrics used to report results in Section 4 are defined in Section 3.7.

3.2 Datasets

A total of 17,170 radiographs were initially available for study-specific screening from three sources: 1) the RSNA Pediatric Bone Age dataset, 2) the Digital Hand Atlas (DHA), and 3) Srinagarind Hospital, Faculty of Medicine, Khon Kaen University (KKU), Thailand. After study-specific exclusions, 17,120 radiographs were included in the analyses summarized in Table 2. The Thai cohort was acquired between September 2008 and December 2022. Public datasets (RSNA and DHA) were used for large-scale representation learning, whereas the Thai cohort was used for local adaptation and held-out target-domain testing. The RSNA Pediatric Bone Age Challenge dataset [15, 39] was originally released with 12,611 training images, 1,425 validation images, and 200 test images (14,236 total). Because the official test partition does not include publicly released bone-age annotations, those 200 images were not eligible for the present supervised analysis, leaving 14,036 labeled training and validation images. A further 28 RSNA images without a full hand were excluded, yielding 14,008 RSNA images. Of the 1,390 DHA images and 1,744 KKU radiographs initially available for study screening, 2 and 20 images, respectively, were excluded during study-specific screening, leaving 1,388 DHA and 1,724 KKU images included in the analysis.

Table 2 Summary of hand radiograph datasets used in this study.

Data source	**Initially available**	**Included**
RSNA	14,036	14,008
DHA	1,390	1,388
KKU	1,744	1,724
Total	**17,170**	**17,120**

For the Thai cohort, bone age assessment was performed by 12 physicians, comprising six pediatric radiologists and six pediatric endocrinologists. The readers were randomly assigned to four mixed-specialty groups of three, with each group reviewing 431 or 432 radiographs. Patient-identifying information was concealed, except for sex. Within each group, all three experts independently and blindly assessed each radiograph using both the Greulich–Pyle (GP) and Tanner–Whitehouse 3 radius–ulna–short bones (TW3-RUS) methods. Agreement was accepted when the bone age

estimates of any two readers were within six months, where the final numerical label was calculated as the mean of at least two concordant readings. Cases not meeting this criterion underwent a second and, when necessary, a third blinded independent reassessment, following washout periods of five months and three weeks, respectively. Any remaining disagreement after the third assessment was resolved through within-group expert consensus to establish the final reference bone age [51].

As a result, multiple supervision standards were defined from these readings, which are GP, TW3-RUS, and a hybrid TW+GP label. The hybrid label was not an arithmetic or statistical combination of GP and TW3-RUS estimates. TW served as the primary reference when TW3-RUS was applicable, whereas GP was substituted for cases outside the validated TW3-RUS age range. Accordingly, the hybrid analysis should be interpreted as evaluation of a rule-based reference strategy rather than as comparison with an independent third scoring standard.

The reliability of this expert-labeling process was evaluated in a companion analysis of the same Thai cohort [51]. Inter-rater agreement was excellent for both labeling approaches: the intraclass correlation coefficient for single-rater agreement, ICC(3,1), was 0.980 for GP and 0.980 for TW3-RUS, with mean absolute differences between readers of under 7 months for both. Intra-rater agreement across repeated readings was likewise excellent, with single-rater ICC values ranging from 0.955 to 0.974 for GP and 0.958 to 0.967 for TW3-RUS.

The image-quality screening leaves 1,724 radiographs in the Thai analysis cohort. This cohort included 805 male patients (46.7%) and 919 female patients (53.3%), with chronological age ranging from 0.8 months to 19.6 years. The age and sex distribution across the three included data sources is shown in Figure 1 (N=17,120; see Table 2). The largest number of cases is concentrated in the pre-adolescent to early-adolescent range, approximately 8 to 14 years, consistent with the age composition of the RSNA and DHA datasets that make up the majority of the combined sample. Detailed demographic distributions of the public datasets have been reported in their original publications [12, 15].

The bone age prediction model was implemented in Keras 2.9.0, whereas the detection and segmentation models were trained using the Ultralytics YOLOv8 framework. All experiments were performed on an NVIDIA A100 GPU.

3.3 Image Preprocessing and ROI Standardization Pipeline

Hand radiographs acquired in real-world clinical environments often exhibit substantial variability in contrast, orientation, field of view, and the presence of non-anatomical artifacts such as labels, markers, or operator hands. These factors can adversely affect the performance and generalization ability of deep learning models. To address these challenges, a multi-stage image preprocessing and hand region extraction pipeline was developed to standardize image inputs and enhance diagnostically relevant skeletal features prior to bone age prediction.[-1] The overview of the image preprocessing is shown in Figure 2.

[-1]Implementation available at https://github.com/Khao0/HandXRay-Preprocessing, which also links to the preprocessed RSNA and DHA images in its README.

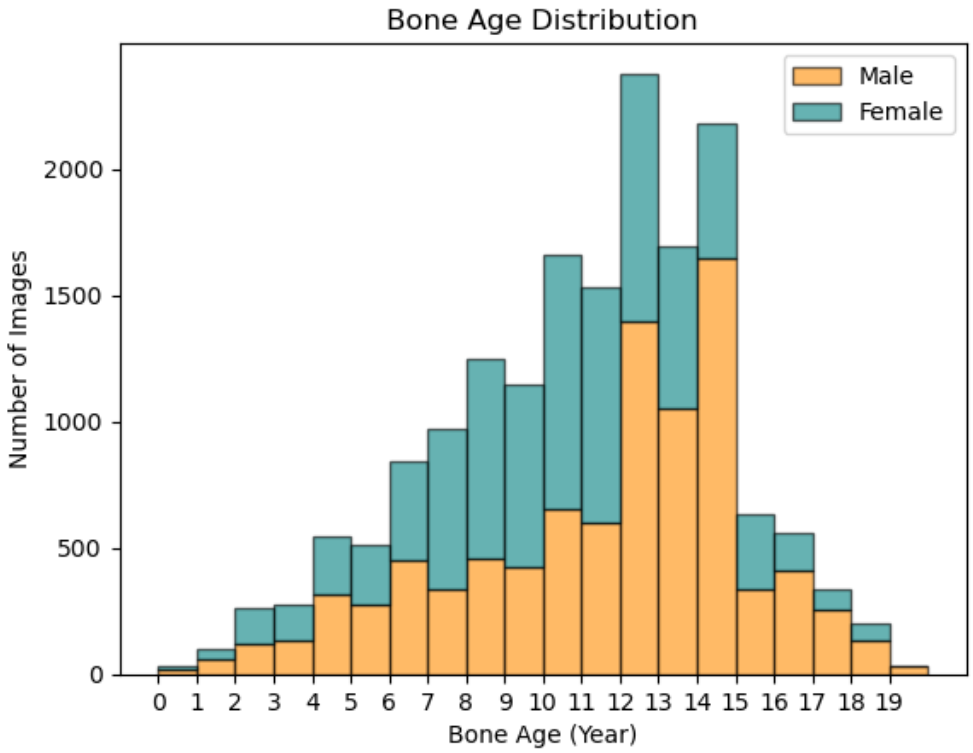


Fig. 1 Age and sex distribution of the analyzed dataset from all sources (Table 2).

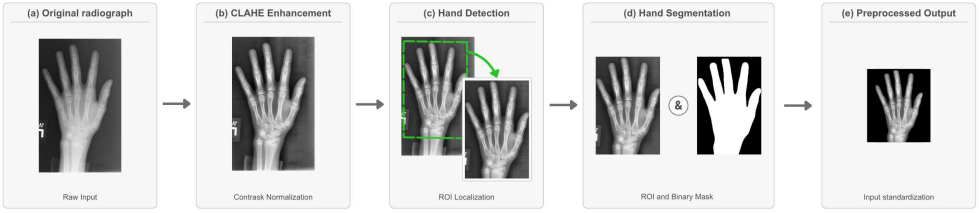


Fig. 2 Overview of the image preprocessing pipeline.

3.3.1 Contrast Enhancement using CLAHE

All radiographs were first enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) [37]. This technique improves local contrast while preventing excessive noise amplification by limiting histogram clipping. CLAHE was applied to emphasize fine skeletal structures, particularly growth plates and epiphyseal regions, which are critical for accurate bone age assessment across different age groups. The clip limit and tile grid size are 3 and 8×8, respectively.

3.3.2 Hand Detection

Following contrast enhancement, a hand detection model based on the YOLOv8-L architecture [18] was employed to localize the hand and wrist region. The model was trained using manually annotated bounding boxes to identify the full anatomical extent of the hand. The bounding-box ground truth used to train and validate this model comprised 600 images in total, drawn using the same stratified sampling approach described in Section 4.1 to preserve the distribution of sex and age: 300 images from RSNA, 200 images from DHA, and 100 images from the KKU cohort. Hand detection enabled automatic cropping of the region of interest, thereby eliminating irrelevant background content such as radiographic labels and imaging equipment that could introduce bias into the learning process.

The YOLOv8-L architecture was selected for both the hand detection and hand segmentation models because it provided satisfactory mAP performance while remaining compatible with the available hardware resources. The detection model was initialized from the `yolov8l.pt` checkpoint (≈43.7 million parameters), pretrained on the COCO dataset; initializing from COCO-pretrained weights, rather than training from scratch, allowed the model to learn low-level visual features efficiently and improved both training efficiency and final accuracy. The model was trained for 100 epochs at an input resolution of 640 × 640, with a batch size of 32, early stopping with patience of 20 epochs, and a dropout rate of 0.4 applied to mitigate overfitting. Training was supervised using a combination of box loss, computed as Complete Intersection over Union (CIoU) loss, which minimizes the discrepancy between predicted and ground-truth bounding boxes, and distribution focal loss (DFL), which refines the predicted distribution of box boundaries to improve localization precision. Detection performance was evaluated using mean Average Precision (mAP), computed on the validation set at the end of each training epoch.

3.3.3 Hand Segmentation

To further refine the region of interest, a dedicated hand segmentation model was developed to separate the hand from the background at the pixel level. This model was trained using manually annotated binary masks: 528 publicly available RSNA hand masks [40] were used as the initial ground truth, supplemented by an additional 95 hand masks manually annotated using Label-Studio [55] (20 from RSNA, 53 from DHA, and 22 from the KKU cohort) to broaden coverage of hand appearance and age range within the training set, for a combined total of 623 annotated masks. The model leveraged a YOLOv8-L segmentation architecture [18], initialized from the pretrained `yolov8l-seg.pt` checkpoint (≈46 million parameters), likewise pretrained on the COCO dataset. The segmentation model was trained using the same configuration as the detection model described above. In addition to the box loss (CIoU) and distribution focal loss (DFL) terms used for detection, the segmentation model was supervised with a segmentation loss, which quantifies the discrepancy between the predicted segmentation mask and the ground-truth mask. As with the detection model, mAP was computed on the validation set at the end of each training epoch to monitor

performance. The combined use of detection followed by segmentation improved segmentation accuracy and computational efficiency, particularly in challenging cases such as infant hand radiographs or images with left/right labels or with partial occlusion by the hand of the person who took the photos. For the qualitative comparison reported in Section 4.2.2 (Figure 5), two alternative segmentation architectures, Efficient-UNet and TensorMask, adapted from the comparison reported by Rassmann et al. [41], were trained on this same 623-image annotated mask set and evaluated across all three data sources (RSNA, DHA, and KKU) to provide a like-for-like comparison against the proposed pipeline.

3.3.4 Data Augmentation for Detection and Segmentation

Due to the limited size of labeled data, data augmentation is applied to increase labeled images in training set. We utilized the Python library Albumentations [5] to combine various transformations for comprehensive practical use. In real-world applications, X-ray images may be affected by rotation, flipping, blur, noise, or lighting artifacts such as sun flare and shadow. Thus, this study applied rotation, flip, brightness adjustment, shadow, sun flare, noise, and blur transformations, as illustrated in Figure 3(d)–(k).

Finally, random combinations of transformations were applied to the images using parameters from Table 3. The resulting augmented images, as illustrated in Figure 3(a)–(c), are currently prepared for training detection and segmentation models. This augmentation process increases the training set by a factor of 20, resulting in a training set comprising 9,600 images for detection and 9,960 images for segmentation.

3.4 Bone Age Prediction Model

3.4.1 Backbone Network and Regression Head

Figure 4 illustrates the bone age prediction model architecture.[0] Following hand segmentation, the resulting hand region was fit with a black edge and resized to 600×600 to produce a fixed-size input, which was then used as input to the bone age prediction model. The EfficientNetB7 [48] architecture, pretrained on ImageNet with global average pooling and without its original classification head, was selected as the backbone CNN for its strong accuracy-to-parameter ratio, achieved by jointly scaling network depth, width, and resolution rather than adjusting each independently as in earlier CNN architectures. EfficientNetB7, the largest variant in this family, was chosen to maximize representational capacity for the fine-grained skeletal features relevant to bone age assessment, such as growth plate morphology, given the computational resources available for this study. The extracted image features were passed through a fully connected layer of 256 units with ReLU activation. In parallel, patient sex was provided as an auxiliary scalar input and encoded through a separate dense branch of 256 units with ReLU activation, giving the network an explicit sex-conditioning signal alongside the radiograph. The image and sex feature representations were concatenated and passed through a further dense layer of 128 units with ReLU activation and L_2 kernel regularization ($\lambda = 1 \times 10^{-4}$), followed by a dropout layer (rate 0.3)

[0]Implementation available at https://github.com/Khao0/HiDA-BAA.

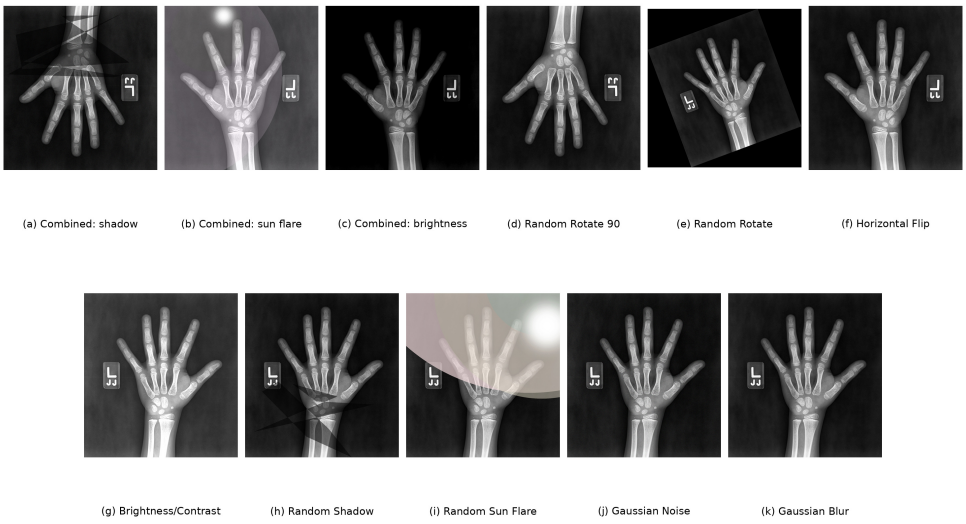


Fig. 3 Examples of all augmentation transformations used in this study, created using Albumentations. (a)–(c) show combined transformations (e.g., simultaneous rotation, brightness adjustment, and lighting effects) applied to produce the final training images, while (d)–(k) show the effect of each individual transformation in isolation: (d) random 90° rotation, (e) random-angle rotation, (f) horizontal flip, (g) brightness/contrast adjustment, (h) simulated shadow, (i) simulated sun flare, (j) Gaussian noise, and (k) Gaussian blur.

to reduce overfitting. A final linear-activation dense layer with a single output unit produced the predicted bone age in months. The model was compiled with the Adam optimizer and mean absolute error (MAE) as the training loss, with MAE also tracked as the reported metric.

3.4.2 Training Strategy

The same core training procedure was used across all three modeling strategies compared in this study (Section 3.6). The EfficientNetB7 backbone was initialized from ImageNet-pretrained weights and trained in two phases: a warm-up phase of 10 epochs at a learning rate of 1×10^{-4}, during which the backbone was kept frozen so that only the newly added dense layers were updated, followed by a training phase of up to 100 epochs at a learning rate of 1×10^{-5}, during which the top 100 layers of the backbone were unfrozen for fine-tuning. All three configurations used the Adam optimizer, a batch size of 32, a weight decay of 1×10^{-4}, early stopping monitored on validation loss with a patience corresponding to 10% of the total training epochs, and a reduce-on-plateau learning-rate schedule with patience of 3 epochs, decay factor of 0.2, and minimum learning rate of 1×10^{-8}.

The three configurations differed only in their training data and evaluation protocol. The public fine-tune model was trained on the public datasets using an 80:20

Table 3 Parameters for matrix transformations using the Albumentations library for augmenting the training set in the detection and segmentation model. Each parameter's range or options are listed.

Transformation	**Type**	**Parameters**
Rotation (Detection)	RandomRotate90	p=1
Rotation (Segmentation)	OneOf	p=1
	- Affine	p=0.75, rotate=(-60, 60), fit_output=True
	- RandomRotate90	p=0.75
Horizontal Flip	HorizontalFlip	p=0.25
Brightness and Contrast	RandomBrightnessContrast	p=0.75
Lighting Effects	OneOf	p=0.75
	- RandomSunFlare	p=1, src_radius=200, angle_lower=0, angle_upper=1, num_flare_circles_lower=3, num_flare_circles_upper=5
	- RandomShadow	p=1
Noise and Blur	OneOf	p=0.75
	- GaussNoise	p=1, var_limit=(10.0, 50.0)
	- GaussianBlur	p=1, blur_limit=(3, 7)

train-validation split, in a single run without cross-validation; consequently, no cross-validation MAE, RMSE, or R^2 is reported for this configuration in Section 4, and its test-set performance is reported. The pooled training model applied the same procedure to RSNA, DHA, and the Thai training subset combined, using the stratified 10-fold cross-validation protocol described in Section 4. The hierarchical adaptation model applied the same procedure in two sequential stages: the first stage matched the public fine-tune model exactly, and the second stage repeated the full warm-up-and-training procedure on the Thai training subset alone, initialized from the first stage's resulting weights rather than from ImageNet directly, and likewise evaluated using stratified 10-fold cross-validation.

3.4.3 Data Augmentation for Bone Age Prediction

To improve generalization to radiographs not seen during training, data augmentation was also applied when training the bone age prediction model on the Thai training subset, separate from the detection- and segmentation-specific augmentation described in Section 3.3.4. These transformations allow the model to learn from variation in hand positioning and orientation and improve robustness to differences in lighting conditions and image quality encountered in real-world settings. Two categories of transformations were applied using Albumentations [5]: geometric augmentations, comprising horizontal flip (probability 0.25), vertical flip (probability 0.25), and one of either a negative rotation ($-45°$ to $-15°$) or a positive rotation ($15°$ to $45°$) selected with

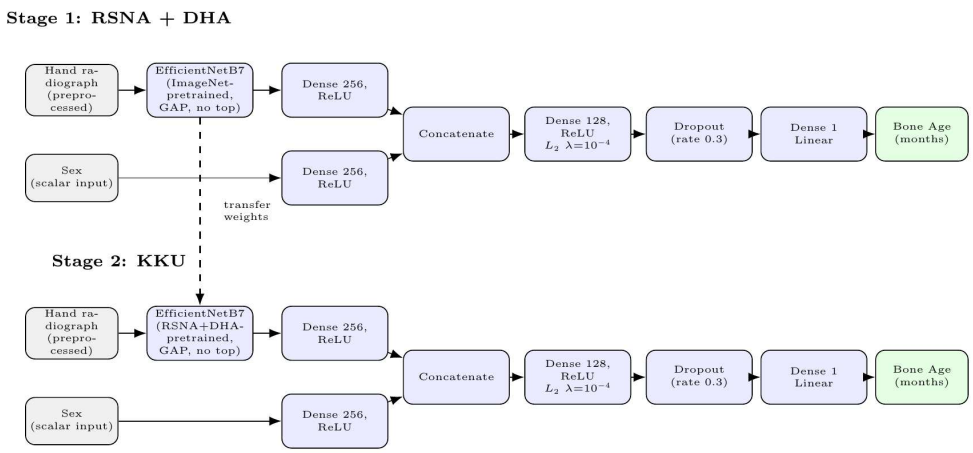


Fig. 4 Model architecture.

probability 0.5; and photometric augmentations, comprising random brightness and contrast adjustment (probability 0.5, maximum change of 20%) and Gaussian blur (probability 0.25). The resulting augmented training set is three times the size of the original training set.

3.5 Hierarchical Transfer Learning Strategy

The hierarchical adaptation strategy is motivated by three findings from the transfer learning literature. First, domain adaptation theory shows that pooling source and target data into a single training objective does not remove the underlying distributional gap between them, especially when the target domain is a small fraction of the combined data [2]. Second, earlier CNN layers tend to learn general-purpose visual features while later layers become more task- and dataset-specific [60], favoring a strategy that first learns general radiographic features before specializing to the target population. Third, training jointly on markedly different distributions lets the more numerous samples dominate optimization, interfering with the smaller distribution's representation in a manner reminiscent of catastrophic forgetting in sequential learning [10]. Together, these results motivate treating public-dataset pretraining and Thai-cohort fine-tuning as separate, sequential stages rather than pooling all data into a single training run.

To improve population-specific performance under cross-dataset domain shift, this study employed a hierarchical transfer learning strategy in which model parameters were adapted progressively from general visual representations to public bone age datasets and finally to the target Thai clinical cohort. The motivation for this design is that hand radiographs from different sources may differ in demographic composition, acquisition protocol, image appearance, and labeling context. Direct training on a limited local cohort may therefore underutilize information available in large public datasets, whereas relying only on public datasets may reduce performance when the model is applied to a demographically distinct target population.

The proposed strategy consisted of two transfer stages after initialization from ImageNet-pretrained weights. In the first stage, the model was trained on the public RSNA and DHA datasets to learn general radiographic and skeletal maturity representations from relatively large-scale data. In the second stage, the pretrained model was further fine-tuned using the Thai training subset so that the learned representation could adapt to target-domain characteristics, including local image appearance, population-specific maturation patterns, and annotation context. The final adapted model was then used for inference on the held-out Thai test set.

This progressive adaptation strategy was designed to preserve the representation strength obtained from large public datasets while allowing subsequent specialization to the local clinical domain. In contrast to a single-stage training process, hierarchical transfer learning explicitly treats the public datasets and the Thai cohort as sequential learning sources with different roles: the former provides broad feature initialization for bone age estimation, whereas the latter provides population-specific refinement.

Unless otherwise stated, the same network architecture, loss function, and optimization settings were used across transfer stages, with only the training dataset and stage-specific initialization differing. The comparative evaluation of this hierarchical strategy against alternative training configurations is described in Section 3.6.1.

3.6 Experimental Design

3.6.1 Domain-Shift-Aware Evaluation

To assess generalization across populations, we designed a domain-shift-aware evaluation protocol in which public datasets and the Thai clinical cohort were assigned different roles in model development and testing. The RSNA and DHA datasets were used as large-scale source-domain datasets for representation learning, whereas the Thai cohort served as the target-domain dataset for population-specific adaptation and evaluation. Specifically, three modeling strategies were developed and compared based on the EfficientNetB7 backbone.

1. **Public fine-tune model:** The EfficientNetB7 backbone, pretrained on ImageNet, was fine-tuned using only the public datasets, providing an out-of-domain baseline. This model was trained in a single run using an 80:20 train-validation split of the public data, rather than cross-validation.
2. **Pooled training model:** The EfficientNetB7 backbone, pretrained on ImageNet, was fine-tuned using RSNA, DHA, and the Thai training subset combined into a single training run, using stratified 10-fold cross-validation.
3. **Hierarchical adaptation model:** This model employed a two-stage fine-tuning strategy. In the first stage, the backbone was fine-tuned using the RSNA and DHA datasets, following the same recipe as the public fine-tune model. In the second stage, the resulting model was further fine-tuned on the Thai training subset using stratified 10-fold cross-validation, following the warm-up and fine-tuning schedule described in Section 3.4.

All three configurations were evaluated on the same held-out Thai test set using hybrid TW+GP labels, in which TW served as the primary reference standard and

GP was substituted only for cases falling outside the validated TW3-RUS age range at the low and high ends, where TW scoring could not be applied. This design allows direct comparison between conventional public-data training, joint multi-dataset training, and progressive population-specific adaptation under the same target-domain test condition.

3.6.2 Preprocessing Ablation Study

To quantify the aggregate contribution of ROI standardization, an ablation study was conducted by comparing two preprocessing configurations prior to bone age prediction, evaluated using EfficientNetB7 with frozen ImageNet-pretrained weights and no bone-age-specific pretraining or fine-tuning of the backbone. This configuration is deliberately simpler than, and distinct from, the three training configurations compared in Section 4.3; it is used here solely to isolate the aggregate effect of the preprocessing pipeline on downstream prediction error, independent of any particular pretraining or fine-tuning strategy. The first configuration used the full proposed pipeline, consisting of CLAHE enhancement followed by hand detection and hand segmentation. The second configuration applied no ROI preprocessing, and the original radiograph was used directly as model input.

3.6.3 Supervision Sensitivity Study

To investigate the effect of labeling standard on model behavior, we compared three supervision settings within the Thai cohort: Greulich–Pyle (GP), Tanner–Whitehouse (TW), and hybrid TW+GP. For each supervision setting, the same model architecture and training configuration (Hierarchical adaptation model) were used, and only the target labels were varied. The GP and TW settings followed the respective conventional reference standards, while the hybrid TW+GP setting was included to examine whether combining information from both standards could improve robustness.

This experiment was designed as a sensitivity analysis of the reference-label strategy under a unified adaptation framework. Because the hybrid target substitutes GP only where TW3-RUS is not applicable, it is not statistically independent of the TW target and its performance difference should not be interpreted as evidence that averaging two scoring systems reduces labeling noise.

3.7 Evaluation Metrics

Model performance was evaluated at multiple stages of the proposed framework, including hand detection and segmentation and the final bone age prediction model. Appropriate quantitative metrics were selected for each task to ensure comprehensive and clinically meaningful evaluation.

3.7.1 Evaluation of Hand Detection and Segmentation Models

The performance of the hand detection and hand segmentation models was evaluated using standard image analysis metrics, each computed by comparing model predictions against manually annotated ground truth.

Mean Intersection over Union (mIoU) quantifies the spatial overlap between a predicted region P and its ground-truth region G as

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \tag{1}$$

where $|\cdot|$ denotes the pixel area of a region. IoU ranges from 0 (no overlap) to 1 (perfect overlap); mIoU is the average IoU across all evaluated images. As a general guide for segmentation tasks, mIoU above approximately 0.90 is considered excellent, 0.70–0.90 good, and below 0.50 indicates substantial boundary error.

Dice Coefficient similarly measures overlap but weights correctly predicted (true-positive) pixels more heavily:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\,\text{IoU}}{1 + \text{IoU}}, \tag{2}$$

also ranging from 0 to 1, with values closer to 1 indicating better agreement. Because of this weighting, Dice is mathematically always greater than or equal to IoU for the same segmentation, which is why it is typically reported alongside IoU in medical image analysis.

Mean Average Precision (mAP) was used to evaluate the hand detection model. A predicted bounding box is counted as a true positive if its IoU with a ground-truth box exceeds a specified threshold; precision (the proportion of predicted boxes that are correct) and recall (the proportion of ground-truth boxes that are successfully detected) are then computed across confidence levels, and Average Precision (AP) is the area under the resulting precision–recall curve. In the notation used in Section 4, mAP@90 and mAP@95 denote AP computed at IoU thresholds of 0.90 and 0.95, respectively, with a higher threshold imposing a stricter localization requirement, while mAP@50:95 averages AP across IoU thresholds from 0.50 to 0.95 in steps of 0.05, following the standard COCO evaluation protocol, and is generally treated as the most stringent single summary of detection quality. All mAP variants range from 0 to 1 (or 0% to 100%), with values above 0.90 (90%) typically considered strong performance for detecting a well-defined, single-class object such as a hand.

3.7.2 Evaluation of Bone Age Prediction Performance

The performance of the bone age prediction models was evaluated on an independent test set, comparing predicted bone age values, $\hat{y}_i$, against expert-assessed reference values, y_i, across n test cases.

Mean Absolute Error (MAE),

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{3}$$

was the primary metric, reported in months, with lower values indicating closer agreement. As a clinical benchmark, expert readers in this cohort disagreed with each other,

and with their own repeated readings, by under 7 months on average [51]; a model MAE at or below this level is comparable to typical human inter-reader variability. Prior automated BAA systems typically report MAE or MAD of approximately 5–11 months on external cohorts (Table 8).

Root Mean Square Error (RMSE),

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

also reported in months, penalizes large errors more heavily than MAE; a much larger RMSE than MAE indicates a subset of cases with unusually large errors.

Coefficient of Determination (R^2),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

where $\bar{y}$ is the mean reference bone age, has a maximum of 1 for perfect prediction, equals 0 when the model performs no better than predicting the mean, and is negative for worse performance; values above approximately 0.90 indicate the model captures most of the variation in skeletal maturity.

These metrics were reported for both validation and test sets to assess model generalization and robustness.

4 Results and Discussion

4.1 Dataset Statistics and Experimental Setup Verification

Model development in this study consisted of three major components: hand detection, hand segmentation, and bone age prediction. Each component required a distinct data splitting strategy.

For the hand detection and hand segmentation models, ground truth annotations in the form of bounding boxes and segmentation masks were required. A subset of the full dataset containing these annotations was divided into training, validation, and test sets using an 80:10:10 ratio. Stratified sampling was applied to preserve the distribution of sex and age across all subsets.

For BAA target-domain evaluation, the 1,724-case KKU cohort was separated by stratified sampling into 1,550 development cases and 174 held-out test cases (approximately 90:10), preserving the distributions of sex and age. The 174 held-out cases were not used for BAA model training, cross-validation, or hyperparameter selection and formed the fixed Thai test set used to report test MAE throughout Section 4. RSNA and DHA were used as source-domain datasets for public-data training and were not part of this held-out Thai evaluation. Because KKU images were also included in the separately developed preprocessing annotation subsets, the term "held-out" in this manuscript refers specifically to the BAA prediction-model evaluation unless

end-to-end patient-level separation across preprocessing development is independently verified.

Within the training set, model performance was further assessed using stratified 10-fold cross-validation (CV) for the pooled training model and the hierarchical adaptation model. This procedure enabled optimal hyperparameter tuning, reduced the risk of overfitting, and improved the generalization ability of the final model. Each of the 10 fold-trained models was subsequently evaluated on the same independent Thai test set. In the tables that follow, *validation MAE* (together with the corresponding RMSE and R^2) refers to performance averaged across the 10 cross-validation folds' own held-out validation subsets, while *test MAE* (together with the corresponding RMSE and R^2) refers to the mean, and where applicable the standard deviation (SD), of the 10 fold-trained models' performance when each is evaluated on the same independent, unseen Thai test set. The public fine-tune model was instead trained as a single run over an 80:20 train-validation split of the public dataset, without cross-validation; consequently, no validation MAE, RMSE, or R^2 is reported for this configuration, and its test-set performance is reported as a single point estimate rather than a mean and SD across multiple models.

For the BAA prediction model, the fixed Thai test set was kept separate from BAA training and validation. We therefore avoid extending the no-leakage claim to the independently developed detection and segmentation components without a patient-identifier audit across their annotation subsets.

4.2 Effect of ROI Preprocessing Pipeline

4.2.1 Quantitative Ablation Study

Two complementary ablations were conducted: the first (Table 4) isolates the contribution of running hand detection immediately before segmentation, and the second (Table 5) quantifies the aggregate effect of the full pipeline, applying CLAHE followed by detection and then segmentation, on downstream bone age prediction, relative to using raw radiographs directly.

Table 4 Comparison of segmentation model performance with and without the use of a detection model prior to segmentation, evaluated using IoU, Dice Coefficient, and mAP metrics.

Detection Model Used Before Segmentation	mIoU (%)	Dice (%)	mAP@90 (%)	mAP@95 (%)	mAP@50:95 (%)
No	92.63	94.99	90.25	38.03	87.58
Yes	94.82	97.33	100.00	26.69	92.67

As shown in Table 4, running detection immediately before segmentation improved mIoU (from 92.63% to 94.82%), Dice coefficient (from 94.99% to 97.33%), mAP@90 (from 90.25% to 100.00%), and mAP@50:95 (from 87.58% to 92.67%), consistent with the intuition that first cropping to the hand region removes background clutter and

long-forearm content that would otherwise compete with the hand boundary during segmentation. The mAP@95, however, decreased when detection was used (from 38.03% to 26.69%), likely because detection-based cropping occasionally shifts predicted mask boundaries by a small number of pixels, an effect that only becomes visible at this strictest IoU threshold and does not undermine the overall improvement seen in mIoU, Dice, and mAP@50:95. It is also worth noting that this comparison does not hold all other factors equal, since the detection stage crops the image before segmentation is performed, changing the effective input the segmentation model sees; nonetheless, our aim here is to quantify the practical value of the preprocessing pipeline as a whole for more precise downstream bone age prediction, rather than to isolate segmentation performance under identical inputs.

Table 5 Effect of the full ROI preprocessing pipeline relative to raw input, evaluated using EfficientNetB7 with frozen ImageNet-pretrained weights and no bone-age-specific fine-tuning, to isolate the preprocessing pipeline's contribution independent of the training configurations compared in Table 6. Single-stage ablations (CLAHE only, detection only, segmentation only) were not evaluated separately.

Configuration (Baseline Model)	CV MAE (months)	Test MAE (months)
No Preprocessing	7.53	7.42
Full Pipeline (CLAHE, detect, then segment)	7.01	6.79

Table 5 shows that, using EfficientNetB7 with frozen, non-fine-tuned ImageNet weights, applying the full preprocessing pipeline reduced test MAE from 7.42 to 6.79 months (a reduction of 0.63 months, ≈8.5%) and CV MAE from 7.53 to 7.01 months, relative to using raw radiographs directly. This confirms that ROI standardization contributes a meaningful, if more modest, improvement on top of the gains later attributed to hierarchical adaptation (Section 4.3), supporting the two mechanisms as complementary rather than redundant contributors to overall performance. These values are not expected to match the public fine-tune model reported in Table 6 (test MAE 7.6 months): the two use different backbone configurations, a frozen, non-fine-tuned backbone here versus a fine-tuned backbone in Table 6, so the ablation in this section isolates the preprocessing pipeline's contribution independently of any particular pretraining or fine-tuning strategy.

4.2.2 Cross-Dataset ROI Processing Completeness and Qualitative Results

Because the proposed pipeline and the publicly available RSNA hand masks [40] were constructed under different annotation conventions, direct pixel-level agreement metrics (e.g., Dice or IoU) between the two are not a fair basis for comparison: the public masks are convex-hull-style annotations that frequently leave small gaps between fingers unfilled or include extraneous regions beyond the hand, whereas the proposed

pipeline was designed to output a single, fully filled hand silhouette. We therefore evaluate cross-dataset robustness of the preprocessing pipeline using a completeness criterion rather than pixel-overlap metrics: an image is counted as *successfully processed* if (1) the detection stage localizes a single, complete hand region (radiographs where only a partial hand was visible or detectable were excluded from this count), and (2) the resulting segmentation mask forms one contiguous, fully filled hand silhouette without unfilled gaps or spurious extraneous regions. Figure 5 illustrates this qualitatively on two representative cases, comparing the original radiograph against the mask produced by the proposed pipeline and against masks produced by two alternative segmentation architectures (Efficient-UNet and TensorMask, adapted from Rassmann et al. [41] and trained on the same 623-image mask set described in Section 3.3.3), where the latter exhibit unfilled gaps and extraneous regions of the kind the completeness criterion above is designed to flag.

Using the predefined operational completeness criterion, the pipeline completed hand detection and segmentation processing for all evaluated RSNA, DHA, and KKU images. An image was counted as complete if (1) the detection stage localized a single, complete hand region and (2) the resulting segmentation mask formed one contiguous, fully filled hand silhouette without unfilled gaps or spurious extraneous regions. This result indicates that the pipeline could operationally process the evaluated image collections, but it does not establish pixel-level segmentation accuracy on those full datasets because independent manual ground-truth masks were not available for every image. Accordingly, we refer to this result as processing completeness rather than quantitative external segmentation validation or cross-population robustness.

Processing completeness across the three data sources does not imply either quantitative segmentation accuracy or transfer of downstream BAA performance across populations (Section 3.6.1 and Table 6). The preprocessing and prediction results are therefore interpreted as distinct evaluations.

4.3 Domain-Shift Evaluation Across Populations

Table 6 Comparison of validation and test metrics across training configurations. The public-only baseline is a single-run point estimate. For pooled training and hierarchical adaptation, validation metrics are averaged across cross-validation folds and test MAE is summarized across the 10 fold-trained models evaluated on the same fixed Thai test set. The reported fold-to-fold SD is not a patient-level confidence interval.

	Validation (months)			Test (months)		
Model	MAE	RMSE	R^2	MAE (SD)	RMSE	R^2
Pretrained on RSNA+DHA		-		7.6	10.58	0.94
Pooled training on all data	5.74	7.33	0.98	5.73 (0.13)	7.57	0.97
Hierarchical adaptation (proposed)	4.92	6.34	0.98	4.64 (0.09)	5.85	0.98

Consistent with the complete preprocessing achieved across all three data sources, the public fine-tune model — despite benefiting from an ROI-standardization pipeline

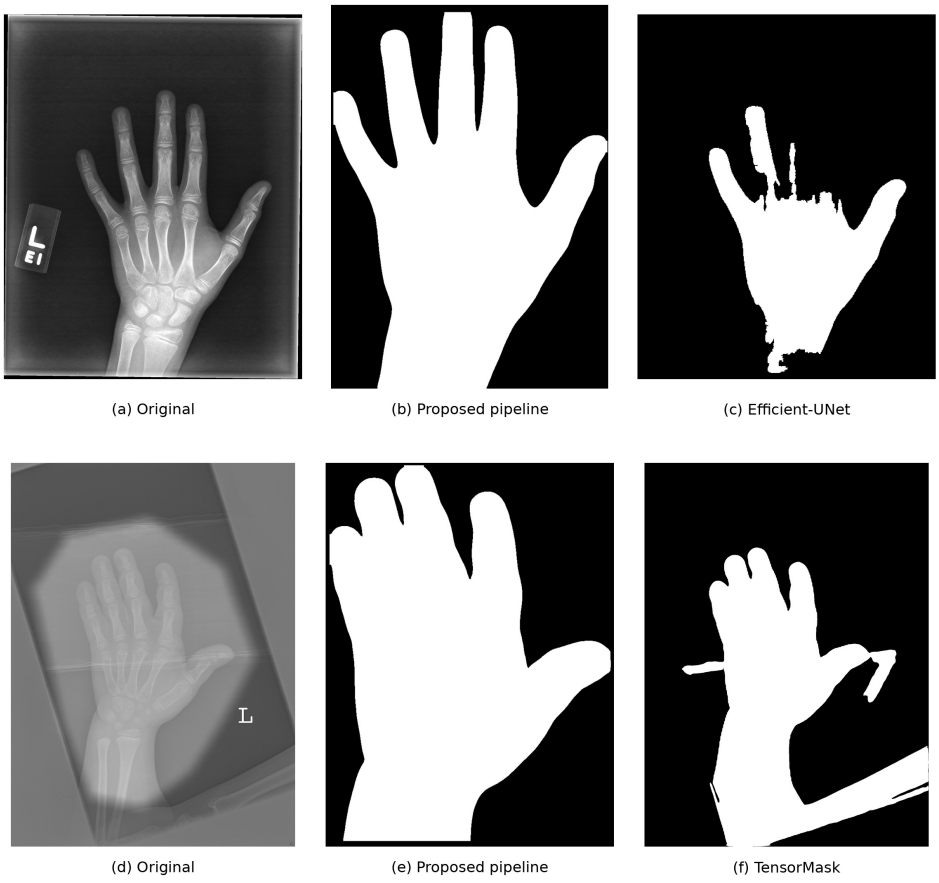


Fig. 5 Qualitative comparison of hand ROI standardization outputs in two representative cases. Panels (a)–(c) show the original radiograph, the output from the proposed pipeline, and the Efficient-UNet output for case 7493, respectively. Panels (d)–(f) show the original radiograph, the output from the proposed pipeline, and the TensorMask output for case 1583, respectively.

that processes external images completely — performs markedly worse when applied directly to the Thai cohort, with a test MAE of 7.6 months and an RMSE of 10.58 months. Its R^2 of 0.94, while still high in absolute terms, is nonetheless the lowest of the three configurations, against 0.97 for pooled training and 0.98 for hierarchical adaptation, indicating that the model captures the overall trend of skeletal maturation with chronological age but is less precise on individual Thai cases than either configuration that incorporates Thai training data. Because this baseline was trained as a single run without cross-validation, no corresponding validation metrics or standard deviation are available for it. Pooled training, which exposes the network to Thai images directly during a single joint training run, performed substantially better, with a test MAE of 5.73 months (SD 0.13) and an R^2 of 0.97, likely because even limited direct

exposure to Thai-specific appearance and maturation patterns during training is far more informative than none at all, as is the case for the public fine-tune model. Even so, pooled training remains outperformed by hierarchical adaptation, which achieved a test MAE of 4.64 months (SD 0.09), an RMSE of 5.85 months, and an R^2 of 0.98. This ordering is consistent with the theoretical considerations introduced in Section 3.5: pooling RSNA, DHA, and Thai data into a single training objective does not eliminate the divergence between the source and target distributions [2], and because Thai images represent a small fraction of the pooled dataset, gradient updates driven by the much larger volume of RSNA and DHA data likely dominate the joint optimization, limiting how far the model's representation can shift toward Thai-specific appearance and maturation patterns [10]. Hierarchical adaptation instead completes public-data representation learning first and then dedicates an uncontested fine-tuning phase to the Thai cohort, consistent with the view that later network layers benefit from being specialized to the target domain once general features are already in place [60]. These results indicate that successful ROI standardization alone does not close the cross-population performance gap: some degree of exposure to target-domain data is necessary, and explicitly staged adaptation of the predictive model provides a further, more reliable improvement beyond simply pooling the available data.

4.4 Supervision Strategy Comparison

Table 7 compares three supervision standards under the same hierarchical adaptation strategy. Hybrid TW+GP supervision achieved the lowest test MAE (4.64 months, SD 0.09), followed by TW (4.76 months, SD 0.15) and GP (4.94 months, SD 0.14). The corresponding RMSE and R^2 values follow the same ordering, with Hybrid TW+GP achieving the lowest test RMSE (5.85 months) and the highest test R^2 (0.98), against 5.97 months and 0.98 for TW and 6.28 months and 0.98 for GP. The hybrid reference strategy produced the lowest point-estimate test MAE, but its advantage over TW alone was modest (0.12 months). Because the hybrid label uses TW when applicable and substitutes GP outside the validated TW3-RUS range, the hybrid and TW targets are not independent reference standards. The observed difference therefore should not be interpreted as evidence that combining two independent scoring philosophies averages out labeling noise.

Table 7 Validation and test performance across reference-label strategies. Validation metrics are averaged across the 10 cross-validation folds; test MAE is reported as the mean (SD) of the same 10 fold-trained models evaluated on the fixed held-out Thai test set.

	Validation (months)			Test (months)		
Label method	MAE	RMSE	R^2	MAE (SD)	RMSE	R^2
GP	5.28	6.85	0.98	4.94 (0.14)	6.28	0.98
TW	5.02	6.59	0.98	4.76 (0.15)	5.97	0.98
Hybrid (TW+GP)	4.92	6.34	0.98	4.64 (0.09)	5.85	0.98

4.5 Model Interpretability

To provide qualitative insight into which anatomical regions the hierarchical adaptation model relies on when estimating bone age, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) [44] to the final convolutional layer of the EfficientNetB7 backbone. Grad-CAM produces a coarse localization heatmap that highlights the image regions most influential to the model's predicted output, allowing visual inspection of whether the network attends to clinically meaningful skeletal structures rather than to background or non-anatomical artifacts.

Figure 6 shows representative Grad-CAM overlays for four Thai test cases. In the illustrated cases (a) and (b), higher activation overlapped multiple skeletal regions that are also considered during conventional BAA, including the radius, ulna, carpal bones, metacarpals, phalanges, and growth plates. These visualizations are exploratory and do not establish that the model follows expert diagnostic reasoning. In other cases, exemplified by (c) and (d), attention was instead concentrated in more limited regions, particularly the carpal bones and the base of the thumb, and occasionally extended to areas outside the bone itself, with comparatively little emphasis on the other skeletal maturity indicators that experts typically consider. This divergence suggests that, in at least some cases, the model may derive predictive information from a narrower set of feature patterns than the broader, multi-region assessment performed by human experts, even though its overall predictive accuracy remains high (Section 4.3). This observation motivates future work that directly probes which specific bones, joints, or epiphyseal gaps the model relies on most heavily, for example through region-ablation or bone-segment masking experiments, to characterize whether these narrower attention patterns reflect genuinely informative shortcuts or a risk to be mitigated before broader clinical deployment; we return to this point in Section 4.7.

4.6 Comparison with Prior Work

Table 8 compares the proposed method with selected prior studies that are most relevant to the present work, namely studies addressing either supervision-target variation or cross-population generalization. To improve interpretability, only studies reporting an external or target-domain performance metric were included. Because prior studies differ in dataset composition, target population, annotation protocol, and evaluation target, the comparison should be interpreted as contextual rather than strictly head-to-head.

Among supervision-related studies, Pan et al. [36] reported an MAE of 11.1 months for a chronological-age–supervised model on an external U.S. trauma cohort, outperforming a GP-based comparator. Kim et al. [19] further showed that a Korean population-specific model trained with chronological-age labels achieved lower MAE than a GP-based model on two external Korean cohorts. Among domain-shift and generalization studies, Beheshtian et al. [1] reported a MAD of 6.9 months on the external DHA test set when evaluating the 16Bit RSNA challenge model, while Rassmann et al. [41] reported MAD values of 5.81 months on DHA and 5.96 months on a German dysplastic cohort for Deeplasia. Liu et al. [28] achieved 0.53 years MAD ($\approx$ 6.36 months MAD) on an external Tibetan test set, and Rassmann et al. [42] reported

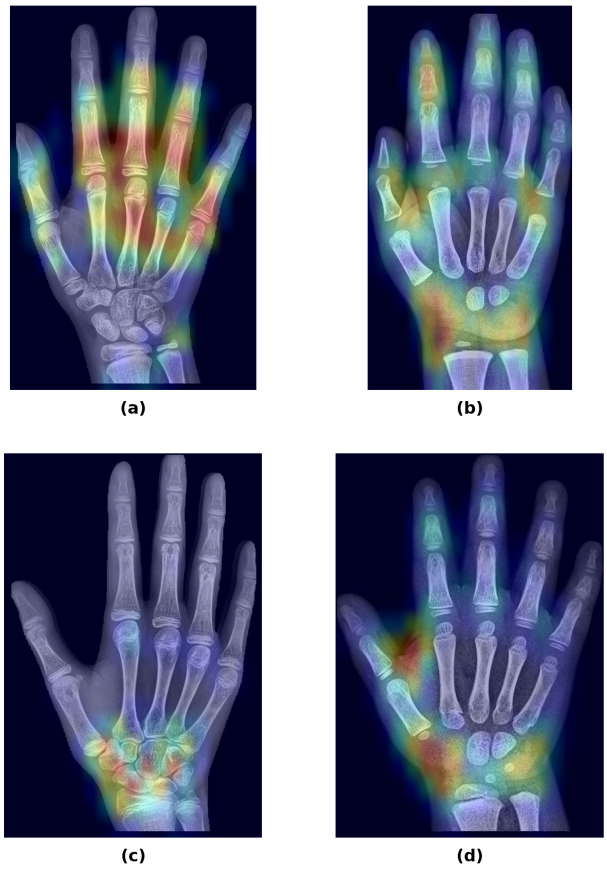


Fig. 6 Representative Grad-CAM visualizations for four Thai test cases. Warmer colors indicate regions with greater influence on the model output. Panels (a) and (b) show activation over multiple skeletal regions, whereas panels (c) and (d) show more localized activation around the wrist and thumb. These examples are qualitative and are not interpreted as evidence that the model reproduces expert reasoning.

5.7 months MAD after population-specific calibration on a Georgian cohort. More recently, Teodoro et al. [50] developed a population-specific GP-based model for a Brazilian pediatric cohort, reporting a MAD of 8.7 months overall and 7.2 months for ages 6–18, illustrating that the population-specific development challenge addressed in this study is an active concern beyond Southeast Asia. In our study, the proposed hierarchical transfer learning framework trained with hybrid TW+GP labels achieved an MAE of 4.64 months on the held-out Thai test set.

Within the evaluated Thai cohort, population-specific adaptation was associated with lower prediction error than the public-only and pooled-training configurations. Compared with prior work, the present study differs in three applied aspects: it evaluates a fixed held-out Thai target cohort for the BAA model, quantifies the aggregate

effect of ROI preprocessing on downstream MAE, and performs a sensitivity analysis across GP, TW, and a rule-based hybrid reference strategy. Because the target-domain evaluation remains single-center and the model comparisons do not include patient-level statistical uncertainty, these results should be interpreted as evidence within the present cohort rather than as proof of transportability or clinical readiness.

Table 8 Comparison with selected prior studies emphasizing supervision strategy and cross-population evaluation.

Study	Population / Evaluation Setting	Supervision	External / Target-Domain Error	Main Relevance to This Work
[36]	CHONY training, external HCH trauma cohort	Chronological age vs GP	11.1 months MAE	Alternative supervision target with external hospital evaluation
[19]	Korean development cohort, 2 external Korean cohorts	Chronological age vs GP-based model	8.2 months MAE, 9.5 months MAE	Population-specific supervision study with external validation
[1]	RSNA-trained 16Bit model, external DHA test set	GP	6.9 months MAD	External-cohort generalizability and bias analysis
[41]	RSNA-trained Deeplasia, external DHA and German dysplasia cohorts	GP	5.81 months MAD (DHA), 5.96 months MAD (GDBD)	External validation on heterogeneous and dysplastic cohorts
[28]	Han/Tibetan training, external Tibetan test set	GP	0.53 years MAD ($\approx$ 6.36 months MAD)	Cross-population domain-shift evaluation
[42]	RSNA-trained model calibrated on Georgian cohort	GP	5.7 months MAD	Population-specific calibration under domain shift
[50]	Brazilian pediatric cohort (n=423), 5-fold cross-validation	GP	8.7 months MAD (overall), 7.2 months MAD (ages 6–18)	Contemporaneous population-specific model for another underrepresented population
This work	RSNA + DHA to Thai adaptation, held-out Thai test set	**Hybrid TW+GP**	**4.64 months MAE**	**Hierarchical adaptation + ROI standardization + multi-standard supervision**

4.7 Limitations and Future Work

This study has several limitations that should be considered when interpreting its findings. First, the Thai radiographs were drawn from a single tertiary-care hospital, although the expert panel that produced the GP and TW3-RUS labels was drawn from multiple institutions and the reliability of these labels was evaluated in a companion analysis [51]. External validation on an independent Thai or other Southeast Asian cohort is therefore required before inferring transportability beyond this center. Second, the GP and TW3 reference standards were originally developed from predominantly North American and European populations; population-specific fine-tuning cannot by itself establish an independent biological ground truth. Third, full-dataset preprocessing success was assessed using an operational completeness criterion rather than independent manual ground truth for every image. It should therefore not be interpreted as quantitative external validation of detection or segmentation accuracy. Fourth, the hybrid target is a TW-primary/GP-substitution rule rather than an independent third reference standard, and the present analysis does not separately quantify the number and age distribution of GP-substituted cases or restrict comparison to cases in which both GP and TW are applicable. Fifth, the reported model-comparison uncertainty is based primarily on variation across fold-trained models rather than paired patient-level confidence intervals. Finally, the study is retrospective; prospective evaluation, computational-efficiency assessment, and independent external validation would be needed before clinical deployment.

5 Conclusion

This study evaluates a population-specific AI framework for automated bone age assessment in Thai children. The hierarchical adaptation model achieved an MAE of 4.64 months on the fixed held-out Thai BAA test set using the hybrid reference strategy, compared with 5.73 months for pooled training and 7.60 months for the public-only baseline.

The results also show why preprocessing and prediction performance should be reported separately. The ROI pipeline met the predefined processing-completeness criterion across the evaluated RSNA, DHA, and KKU collections, but this operational result is not equivalent to quantitative external segmentation validation. In contrast, BAA prediction error was lower after target-population adaptation than with public-only training. These findings support local re-evaluation of both preprocessing behavior and predictive performance when adapting AI systems to an underrepresented population, while avoiding claims of clinical readiness or cross-center generalizability that were not directly tested.

The reference-label sensitivity analysis further showed that the TW-primary/GP-substitution hybrid strategy had the lowest point-estimate MAE, although the difference from TW alone was small and was not tested with paired patient-level inference. Overall, the study provides evidence that population-specific adaptation and explicit reporting of preprocessing behavior are relevant considerations for AI-assisted BAA, while independent external and prospective validation remain necessary before clinical adoption.

Acknowledgements. The authors would like to thank all patients whose radiographic data were included in this study. We also express our sincere appreciation to the pediatric radiologists and endocrinologists who participated in independent image interpretation and expert annotation.

Declarations

Funding. This research was supported by the Digital Economy and Society Fund, Thailand (Grant No. PR66010158).

Ethics approval. This retrospective study was approved by the Research Ethics Committee of the Faculty of Medicine, Khon Kaen University, Thailand (protocol number HE674010). The study was conducted in accordance with institutional regulations and the Declaration of Helsinki. All patient data were anonymized before analysis to ensure confidentiality.

Consent to participate. The requirement to obtain informed consent from the patients and their parents or legal guardians was waived by the Research Ethics Committee of the Faculty of Medicine, Khon Kaen University, Thailand (protocol number HE674010) because of the retrospective nature of the study.

Clinical trial registration. Clinical trial number: not applicable.

Data Availability. The RSNA Pediatric Bone Age Challenge and Digital Hand Atlas datasets analyzed in this study are publicly available from the repositories cited in the manuscript. The Thai clinical radiographs are not publicly available because of patient confidentiality and the conditions of institutional ethics approval. The improved RSNA and DHA hand ROI annotations generated in this study are available at https://www.kaggle.com/datasets/kwankhaotangprasert/rsna-handmask and https://www.kaggle.com/datasets/kwankhaotangprasert/dha-handmask, respectively.

Code Availability. An open-source implementation of the image preprocessing and ROI standardization pipeline described in Section 3.3 is available at https://github.com/Khao0/HandXRay-Preprocessing, which also provides links to the preprocessed RSNA and DHA images in its README. The repository for the bone age prediction model described in Section 3.4 is available at https://github.com/Khao0/HiDA-BAA.

Consent for publication. Not applicable. The manuscript contains no identifiable individual participant information.

Competing interests. The authors declare no competing interests.

Declaration of Generative AI in Scientific Writing. During the preparation of this work, the authors used ChatGPT (OpenAI) in order to refine the language, improve the grammar, clarity, and structure of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Author contributions. Pitchaya Wiratchotisatian: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review and editing; Khwankaow Tangprasert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; Sarun Paisarnsrisomsuk: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision; Panawit Hanpinitsak: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review and editing; Wichuda Chaisiwamongkol: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft; Ratikorn Chaisiwamongkol: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision; Nipaporn Tewattanarat: Conceptualization, Data curation, Investigation, Methodology, Resources; Chanakarn Poonpol: Data curation, Investigation, Project administration, Visualization; Chatparin Pansukrada: Project administration, Software; Wilairat Thaowandee: Project administration, Supervision.

References

[1] Beheshtian E, Putman K, Santomartino SM, et al (2022) Generalizability and bias in a deep learning pediatric bone age prediction model using hand radiographs. Radiology 306(2):e220505

[2] Ben-David S, Blitzer J, Crammer K, et al (2010) A theory of learning from different domains. Machine Learning 79(1-2):151–175

[3] Booz C, Yel I, Wichmann JL, et al (2020) Artificial intelligence in bone age assessment: Accuracy and efficiency of a novel fully automated algorithm compared to the Greulich–Pyle method. European Radiology Experimental 4:6. https://doi.org/10.1186/s41747-019-0139-9

[4] Bui TD, Lee JJ, Shin J (2019) Incorporated region detection and classification using deep convolutional networks for bone age assessment. Artificial intelligence in medicine 97:1–8

[5] Buslaev A, Iglovikov VI, Khvedchenya E, et al (2020) Albumentations: Fast and flexible image augmentations. Information 11(2). https://doi.org/10.3390/info11020125, URL https://www.mdpi.com/2078-2489/11/2/125

[6] Cavallo F, Mohn A, Chiarelli F, et al (2021) Evaluation of bone age in children: a mini-review. Frontiers in Pediatrics 9:580314

[7] Chaisiwamongkol R, Tewattanarat N, Thawande W, et al (2026) Current status, problems, and needs assessment for diagnostic tools in bone age assessment and growth monitoring in thai children. Thammasat University Hospital Journal Online 11(1)

[8] Deshmukh S, Khaparde A (2022) Faster region-convolutional neural network oriented feature learning with optimal trained recurrent neural network for bone age assessment for pediatrics. Biomedical Signal Processing and Control 71:103016

[9] Eng DK, Khandwala NB, Long J, et al (2021) Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: A prospective multicenter randomized controlled trial. Radiology 301(3):692–699. https://doi.org/10.1148/radiol.2021204021

[10] French RM (1999) Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences 3(4):128–135

[11] Gao Y, Zhu T, Xu X (2020) Bone age assessment based on deep convolution neural network incorporated with segmentation. International journal of computer assisted radiology and surgery 15(12):1951–1962

[12] Gertych A, Zhang A, Sayre J, et al (2007) Bone age assessment of children using a digital hand atlas. Computerized medical imaging and graphics 31(4-5):322–331

[13] Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of the hand and wrist. The American Journal of the Medical Sciences 238(3):393

[14] Hai P, Bin Z, Kesheng L, et al (2025) Lightweight deep learning system for automated bone age assessment in chinese children: enhancing clinical efficiency and diagnostic accuracy. Frontiers in Endocrinology 16:1604133

[15] Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al (2019) The rsna pediatric bone age machine learning challenge. Radiology 290(2):498–503

[16] Iglovikov VI, Rakhlin A, Kalinin AA, et al (2018) Paediatric bone age assessment using deep convolutional neural networks. In: International Workshop on Deep Learning in Medical Image Analysis, Springer International Publishing Cham, pp 300–308

[17] Jia Y, Zhang X, Du H, et al (2022) Fine-grained precise-bone age assessment by integrating prior knowledge and recursive feature pyramid network. EURASIP Journal on Image and Video Processing 2022(1):12

[18] Jocher G, Chaurasia A, Qiu J (2023) Ultralytics YOLO. URL https://github.com/ultralytics/ultralytics

[19] Kim PH, Yoon HM, Kim JR, et al (2023) 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 Journal of Radiology 24(11):1151

[20] Koitka S, Kim MS, Qu M, et al (2020) Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. Medical Image Analysis 64:101743. https://doi.org/10.1016/j.media.2020.101743

[21] Larson DB (2022) Openness and transparency in the evaluation of bias in artificial intelligence

[22] Larson DB, Chen MC, Lungren MP, et al (2018) Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 287(1):313–322

[23] Lee H, Tajmir S, Lee J, et al (2017) Fully automated deep learning system for bone age assessment. Journal of digital imaging 30(4):427–441

[24] Li S, Liu B, Li S, et al (2022) A deep learning-based computer-aided diagnosis method of x-ray images for bone age assessment. Complex & Intelligent Systems 8(3):1929–1939

[25] Li X, Jiang Y, Liu Y, et al (2022) Ragcn: Region aggregation graph convolutional network for bone age assessment from x-ray images. IEEE Transactions on Instrumentation and Measurement 71:1–12

[26] Li Z, Chen W, Ju Y, et al (2023) Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. Frontiers in Artificial Intelligence 6:1142895. https://doi.org/10.3389/frai.2023.1142895

[27] Liu C, et al (2021) Self-supervised attention mechanism for pediatric bone age assessment with efficient weak annotation. IEEE Transactions on Medical Imaging 40(10):2685–2697. https://doi.org/10.1109/TMI.2020.3046672

[28] Liu Q, Wang H, Wangjiu C, et al (2024) An artificial intelligence-based bone age assessment model for han and tibetan children. Frontiers in Physiology 15:1329145

[29] Liu Y, Zhang C, Cheng J, et al (2019) A multi-scale data fusion framework for bone age assessment with convolutional neural networks. Computers in biology and medicine 108:161–173

[30] Liu Y, Ouyang L, Wu W, et al (2024) Validation of an established tw3 artificial intelligence bone age assessment system: a prospective, multicenter, confirmatory study. Quantitative Imaging in Medicine and Surgery 14(1):144–159

[31] Liu ZQ, Hu ZJ, Wu TQ, et al (2023) Bone age recognition based on Mask R-CNN using Xception regression model. Frontiers in Physiology 14:1062034. https://doi.org/10.3389/fphys.2023.1062034

[32] Mao X, Hui Q, Zhu S, et al (2023) Automated skeletal bone age assessment with two-stage convolutional transformer network based on x-ray images. Diagnostics

13(11):1837

[33] Martin DD, Calder AD, Ranke MB, et al (2022) Accuracy and self-validation of automated bone age determination. Scientific Reports 12:6388. https://doi.org/10.1038/s41598-022-10292-y

[34] Ontell F, Ivanovic M, Ablin DS, et al (1996) Bone age in children of diverse ethnicity. AJR American journal of roentgenology 167(6):1395–1398

[35] Pan I, Thodberg HH, Halabi SS, et al (2019) Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. Radiology: Artificial Intelligence 1(6):e190053. https://doi.org/10.1148/ryai.2019190053

[36] Pan I, Baird GL, Mutasa S, et al (2020) Rethinking Greulich and Pyle: A deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. Radiology: Artificial Intelligence 2(4):e190198. https://doi.org/10.1148/ryai.2020190198

[37] Pizer SM, Amburn EP, Austin JD, et al (1987) Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 39(3):355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

[38] Prokop-Piotrkowska M, Marszałek-Dziuba K, Moszczyńska E, et al (2021) Traditional and new methods of bone age assessment-an overview. Journal of clinical research in pediatric endocrinology 13(3):251

[39] Radiological Society of North America (2017) RSNA pediatric bone age challenge. https://www.kaggle.com/datasets/kmader/rsna-bone-age, accessed: 2026-02-04

[40] Rassmann S, Hustinx A, Krawitz PM, et al (2023) Hand mask for the RSNA bone age dataset. Zenodo Dataset, https://doi.org/10.5281/zenodo.7611677

[41] Rassmann S, Keller A, Skaf K, et al (2024) Deeplasia: deep learning for bone age assessment validated on skeletal dysplasias. Pediatric Radiology 54(1):82–95

[42] Rassmann S, Abashishvili L, Melikidze E, et al (2025) Population-specific calibration and validation of an open-source bone age AI. Scientific Reports 15:32673. https://doi.org/10.1038/s41598-025-20148-w

[43] Ren X, Li T, Yang X, et al (2018) Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE journal of biomedical and health informatics 23(5):2030–2038

[44] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 618–626

[45] Soegiharto BM, Cunningham SJ, Moles DR (2008) Skeletal maturation in indonesian and white children assessed with hand-wrist and cervical vertebrae methods. American journal of orthodontics and dentofacial orthopedics 134(2):217–226

[46] Spampinato C, Palazzo S, Giordano D, et al (2017) Deep learning for automated skeletal bone age assessment in x-ray images. Medical image analysis 36:41–51

[47] Tajmir SH, Lee H, Shailam R, et al (2019) Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiology 48(2):275–283. https://doi.org/10.1007/s00256-018-3033-2

[48] Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR, pp 6105–6114

[49] Tanner JM, Healy MJR, Goldstein H, et al (2001) Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method), 3rd edn. W.B. Saunders, open Library Work ID: OL18431112W (accessed 2026-03-03)

[50] Teodoro RdA, de Lima FV, Popa M, et al (2026) Automatic bone age assessment: a deep learning case study on the brazilian population with a supporting mobile application prototype. Scientific Reports 16:4488. https://doi.org/10.1038/s41598-025-34651-7

[51] Tewattanarat N, Chaisiwamongkol W, Wiratchotisatian P, et al (2026) Bone age assessment: comparative analysis of greulich-pyle and tanner-whitehouse 3 by pediatric radiologists and endocrinologists. BMC Medical Imaging 26:210. https://doi.org/10.1186/s12880-026-02350-y

[52] Thodberg HH (2009) An automated method for determination of bone age. The Journal of Clinical Endocrinology & Metabolism 94(7):2239–2244

[53] Thodberg HH, Kreiborg S, Juul A, et al (2009) The bonexpert method for automated determination of skeletal maturity. IEEE transactions on medical imaging 28(1):52–66

[54] Thodberg HH, Thodberg B, Ahlkvist J, et al (2022) Autonomous artificial intelligence in pediatric radiology: the use and perception of bonexpert for bone age assessment. Pediatric Radiology 52(7):1338–1346

[55] Tkachenko M, Malyuk M, Holmanyuk A, et al (2020–2025) Label Studio: Data labeling software. URL https://github.com/HumanSignal/label-studio, open source software available from https://github.com/HumanSignal/label-studio

[56] Wang X, Zhou B, Gong P, et al (2022) Artificial intelligence–assisted bone age assessment to improve the accuracy and consistency of physicians with different levels of experience. Frontiers in pediatrics 10:818061

[57] Wang YM, Tsai TH, Hsu JS, et al (2020) Automatic assessment of bone age in taiwanese children: a comparison of the greulich and pyle method and the tanner and whitehouse 3 method. The Kaohsiung Journal of Medical Sciences 36(11):937–943

[58] Xu X, Xu H, Li Z (2022) Automated bone age assessment: a new three-stage assessment method from coarse to fine. In: Healthcare, MDPI, p 2170

[59] Yang C, Dai W, Qin B, et al (2023) A real-time automated bone age assessment system based on the rus-chn method. Frontiers in Endocrinology 14:1073219

[60] Yosinski J, Clune J, Bengio Y, et al (2014) How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems (NeurIPS), pp 3320–3328

[61] Zech JR, Badgeley MA, Liu M, et al (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine 15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

[62] Zhou XL, Wang EG, Lin Q, et al (2020) Diagnostic performance of convolutional neural network-based tanner-whitehouse 3 bone age assessment system. Quantitative imaging in medicine and surgery 10(3):657