Automated Segmentation and Length Measurement of Metacarpal and Phalangeal Bones for Hand Radiograph Evaluation

Philip Gutberlet
University of Bonn
Aron Kirchhoff
University of Bonn
Eike Bolmer
University of Bonn
Philipp Schmidt
Otto-von-Guericke University Magdeburg
Fabio Hellmann
University of Augsburg
Johannes Grün
University of Bonn
Alexander Hustinx
University of Bonn
Elisabeth André
University of Augsburg
Thomas Schultz
University of Bonn
Klaus Mohnike
Otto-von-Guericke University Magdeburg
Peter Krawitz
University of Bonn
Behnam Javanmardi
b.jav@uni-bonn.de

University of Bonn

Research Article

Keywords: pediatric hand radiograph, deep learning, segmentation, length measurement, automation, diagnosis

Posted Date: December 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-8427266/v1

Additional Declarations: Competing interest reported. P.G., A.K., E.B., P.S., and B.J. are co-inventors on a pending patent application containing parts of the methods described in this manuscript.

Automated Segmentation and Length Measurement of Metacarpal and Phalangeal Bones for Hand Radiograph Evaluation

Philip Gutberlet[1], Aron Kirchhoff[1], Eike Bolmer[1], Philipp Schmidt[2], Fabio Hellmann[3], Johannes Grün[4], Alexander Hustinx[1], Elisabeth André[3], Thomas Schultz[4,5], Klaus Mohnike[2], Peter Krawitz[1], Behnam Javanmardi[1*]

[1]Institute for Genomic Statistics and Bioinformatics, Universitätsklinikum Bonn, Baunscheidtstr. 17, Bonn, 53113, Germany.
[2]Medical Faculty, Otto-Von-Guericke-University Magdeburg, Street, Magdeburg, Germany.
[3]Human-Centered Artificial Intelligence, University of Augsburg, Universitaetsstrasse 6a, 86159, Augsburg, Bavaria, Germany.
[4]b-it and Computer Science Department, University of Bonn, Bonn, 53115, Germany.
[5]Lamarr Institute for Machine Learning and Artificial Intelligence.

*Corresponding author(s). E-mail(s): behnam.javanmardi@uni-bonn.de;

Abstract

Evaluating hand and wrist radiographs is essential in pediatric endocrinology and clinical genetics, particularly for the assessment of suspected skeletal anomalies. In this study, we present *Auto-Bone-Caliper*, an automated system for the segmentation and length measurement of metacarpal and phalangeal (M&P) bones, trained and evaluated on public datasets comprising both normal and dysmorphic cases. We first introduce InstanceSAM, a two-stage framework that detects and segments all 19 M&P bones in pediatric hand radiographs, achieving Dice scores of 98.7% for normal bones and 95.0% for dysmorphic bones. We further develop and evaluate three methods for bone-length estimation, identifying a ***k***-means–based approach as the most accurate, with relative errors of 2.2% for normal bones and 4.5% for dysmorphic bones. Our automated pipeline, *Auto-Bone-Caliper*, integrates InstanceSAM with the ***k***-means–based length estimation method. Additionally, we statistically compare measurements

obtained using *Auto-Bone-Caliper* on an independent dataset with a healthy reference catalog of normal bone morphologies, observing a high level of agreement (Wasserstein-1 distance = 0.012). Finally, we demonstrate a clinical use case of *Auto-Bone-Caliper* by obtaining M&P profiles for three genetic conditions, namely Turner syndrome, achondroplasia, and pseudohypoparathyroidism. Our results highlight the potential of the *Auto-Bone-Caliper* to streamline and standardize M&P length measurement, providing an objective and reproducible tool suitable for clinical application.

Keywords: pediatric hand radiograph, deep learning, segmentation, length measurement, automation, diagnosis

1 Introduction

The evaluation of hand and wrist X-rays is a fundamental component in pediatric endocrinology and clinical genetics, particularly for assessing children with suspected skeletal dysplasias and other bone anomalies [1]. Beyond their traditional role in estimating bone age [2–4], these radiographs provide critical insights into bone morphology, growth plate architecture, and ossification patterns that can help identify or differentiate a wide spectrum of skeletal disorders [5]. In children with disproportionate short stature, deformities of the long bones, or abnormal bone density, detailed analysis of the hand X-ray can reveal diagnostic features such as metaphyseal irregularities, epiphyseal dysgenesis, and abnormal carpal ossification sequences characteristic of specific dysplasias (e.g., achondroplasia, pseudoachondroplasia, or metaphyseal chondrodysplasia) [6]. Moreover, hand radiographs assist in distinguishing endocrine causes of growth disturbance from structural or genetic bone disorders, guiding clinicians toward appropriate hormonal, genetic, or metabolic testing. The importance of hand X-rays in diagnosing skeletal dysplasias arises from their ability to capture the fine detail of ossification centers and cortical bone modeling—making them indispensable for both initial evaluation and longitudinal monitoring of children with abnormal skeletal development [5]. In particular, the lengths and proportions of individual metacarpal and phalangeal bones (hereafter referred to as M&P bones) are frequently altered across a wide range of bone disorders, and their detailed assessment can greatly aid in diagnosis [7–9].

For example, M&P profile analysis (which is the analysis of length and proportions of the M&P bones on a hand X-ray) have been shown to be a useful tool for diagnosing diastrophic dysplasia [10], Sotos syndrome [11], dyschondrosteosis, Turner syndrome, and hypochondroplasia [12, 13]. A notable effort was done by A. K. Poznanski who analyzed and published M&P length measurements for healthy individuals and for over 100 disorders [8, 14].

However, manually measuring the 19 M&P bones in a hand radiograph is a laborious and time-consuming process that is also susceptible to human error. An automated method for quantifying M&P bone lengths in pediatric hand radiographs could therefore provide substantial clinical benefit by improving efficiency, consistency, and

diagnostic accuracy. In this study, we introduce a novel automated approach designed specifically for this purpose.

From an image analysis perspective, accurate bone segmentation is a prerequisite for reliable length measurement. Several studies have explored segmentation of individual hand bones to enable quantitative skeletal analysis. Early model-based methods demonstrated feasibility using anatomical priors [15–17]. Later studies addressed pediatric and clinical contexts using increasingly advanced segmentation techniques. Phalangeal regions and epiphyseal/metaphyseal areas in children were segmented for growth assessment using image enhancement and boundary detection algorithms [18], while active shape models were also applied for individual bones [19, 20]. More recently, deep learning–based methods have enabled automated bone-level segmentation: fully automatic networks have been developed for phalanges segmentation in bone age analysis, and convolutional neural network (CNN) frameworks have been used to segment metacarpals and phalanges in arthritis and bone age studies [21–25]. On the length measurement topic, current studies have focused on the metacarpal bones [26, 27], while the phalangeal bones have received little to no consideration.

However, existing studies on hand-bone segmentation and measurement have primarily focused on cohorts with normal skeletal morphology, without evaluating model performance on dysmorphic bones—a limitation likely driven by the scarcity of such data.

In this work, we introduce a new pipeline for automated segmentation and length measurement of the 19 M&P bones and assess its performance on normal and also dysmorphic hand radiographs that are frequently observed in genetic and endocrine bone disorders. Our proposed approach provides a step toward quantitative and objective evaluation of hand radiographs, with potential to enhance clinical assessment in these complex patient populations. Figure 1 shows a schematic of our automatic system which we call *Auto-Bone-Caliper*.

2 Materials and Methods

2.1 Data Collection

We used three public image data sources in this study.

RSNA Bone Age Dataset: The primary dataset used in this work is from the Radiological Society of North America (RSNA) Pediatric Bone Age Challenge 2017 [28]. This public dataset contains over 14,000 high-resolution pediatric hand X-rays with bone age and sex information [29]. The majority of these X-rays belong to individuals with normal hand bone morphology. From the RSNA dataset, we selected a subset of 550 X-rays with an equal number of male and female subjects. By randomly selecting images from sex-specific bins, the subset covers the entire range of bone ages. The radiographs were visually inspected to ensure quality and normal bone morphology. We used hand X-rays from the RSNA for the development of our segmentation model.

GestaltMatcher Database: We obtained hand X-rays of patients with skeletal conditions from the GestaltMatcher Database (GMDB) [30, 31]. GMDB is a collection

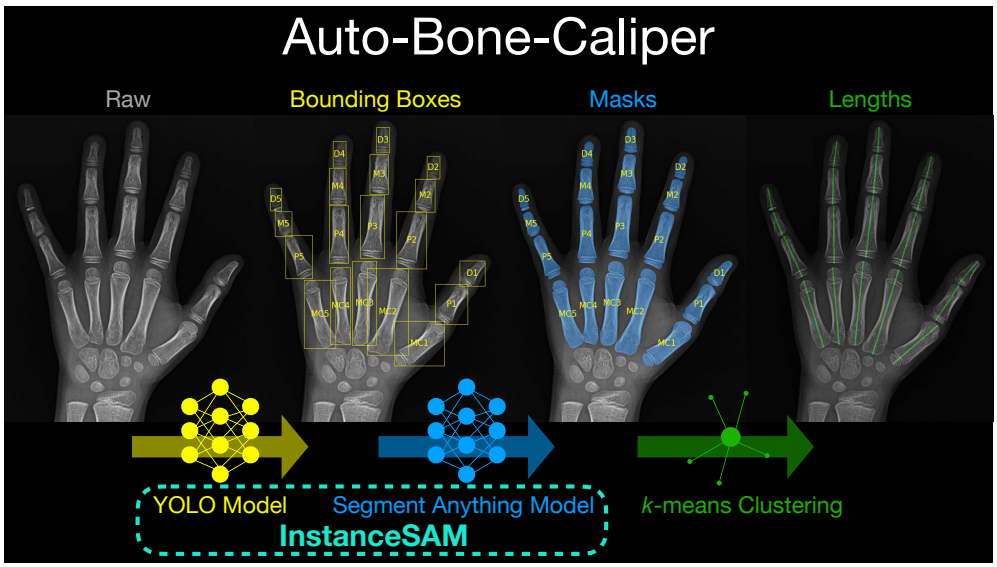

Fig. 1 A schematic of our system, *Auto-Bone-Caliper*, for automatic instance segmentation and length measurements of the M&P bones. The raw radiograph is from the RSNA dataset. MC: Metacarpal bone, P: Proximal phalanx, M: Middle phalanx, D: Distal phalanx.

of curated medical imaging of genetic syndromes as a resource for clinicians and researchers. From the GMDB, we used 50 hand X-rays from 12 different conditions, namely Turner syndrome (n=13), achondroplasia (n=10), pseudohypoparathyroidism (n=10), Proteus (mild phenotype, n=4), Ollier (n=3), McCune-Albright syndrome (n=3), Down syndrome (n=2), Marfan syndrome (n=1), Noonan syndrome (n=1), Hypochondroplasia (n=1), Leri-Weil syndrome (n=1), and Pseudopseudohypoparathyroidism (n=1). We used hand X-rays from the GMDB for testing the performance of our segmentation and length measurement methods on dysmorphic bones.

Digital Hand Atlas Dataset: As an additional test set of normal morphology bones for testing our length measurement approach, we used the publicly available Los Angeles Digital Hand Atlas (DHA) dataset [32], which contains 1,390 radiographs acquired between 1997 and 2008 at the Children's Hospital Los Angeles, USA. The DHA dataset contains radiographs from individuals of four different ancestry groups: Caucasian, Asian, African American, and Hispanic. Since both the RSNA dataset and the data of Poznanski (1984) [8] (see below) are primarily based on children of European ancestry, we restricted our analysis to the corresponding ancestry group in the DHA dataset, comprising 330 hand X-rays.

In addition to these image data sets, we also used the reference M&P bone length measurements from Poznanski (1984) [8]. These references contain mean and standard deviation values of bone lengths for different age ranges and for both males and females

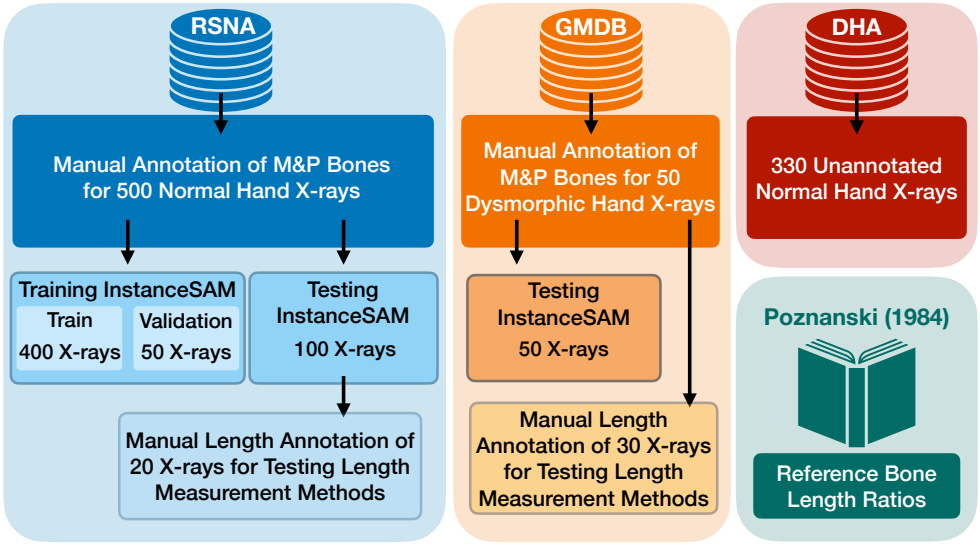

Fig. 2 The datasets used in this study. They are all publicly avilable. RSNA: Radiological Society of North America, GMDB: GestaltMatcher Database, DHA: Digital Hand Atlas.

from a healthy cohort. It also contains ratios of the lengths of the 19 M&P bones which enables a scale-independent analysis. We used these reference ratios to test the performance of *Auto-Bone-Caliper* when applied on an independent healthy cohort (from the DHA dataset). Figure 2 summarizes the datasets used in this study.

2.2 Data Preparation

2.2.1 Bone Mask Annotations

The image datasets mentioned above provide neither bone annotations nor bone masks. Consequently, we used the Labelbox Annotation Tool[1] to manually annotate segmentation masks and labels for the 19 M&P bones in all 550 hand X-rays from the RSNA and 50 hand X-rays from the GMDB, resulting in 11,400 bone masks with labels.

2.2.2 Manual Length Measurements

We manually measured the lengths of the M&P bones, using Labelbox, in 20 normal and 30 dysmorphic hand radiographs (10 cases each for Turner syndrome, achondroplasia, and pseudohypoparathyroidism). The manual length measurements were done by a clinician-scientist (P.S.) following clinical guidelines [7] guided by a pediatric endocrinologist with more than 40 years of clinical experience (K.M.). These annotations are the ground truth for testing our automated length measurement methods (see

[1] https://www.labelbox.com

below). Throughout this paper, we follow the standard anatomic naming approach. Finger numbers start with the thumb (1) and go to the little finger (5).

2.3 Deep Learning Model for Bone Segmentation

We develop a two-stage approach for automatic instance segmentation of the 19 M&P bones consisting of the YOLO v9 model [33] followed by the Segment Anything Model v1 (hereafter, SAM) [34]. First, we fine-tuned a YOLO v9 model with pre-trained weights (COCO2017) for 50 epochs with data augmentations (horizontal flipping, random scaling, random brightness). For this purpose, we used 400 RSNA hand X-rays for training, 50 for validation, and 100 for test. This model was trained to predict the bounding boxes and class labels for each of the 19 M&P bones. The model's performance was measured on the average precision with an intersection over union (IoU) threshold set to 0.5. Subsequently, the outputs from the fine-tuned YOLO model, i.e. the predicted bounding boxes, are used as input for SAM, to give it a cue to the location of the bone to segment. We used the ViT-H SAM model pre-trained on the SA-1B dataset. We refer to the combination of our fine-tuned YOLO model and SAM as InstanceSAM (Figure 1) and use the Dice score as the metric for quantifying its performance.

2.4 Algorithms for Bone Length Measurements

Once InstanceSAM predicts the segmentation masks for each bone, the next challenge is accurately calculating bone lengths. While this may seem straightforward, adhering to clinical measurement standards is essential. There are two main challenges in measuring bone length: identifying the outline and defining the bone axis along which the length is measured. Simple methods, such as identifying the longest line within the mask, are insufficient, as this would yield a diagonal line. Clinical guidelines require length to be measured along the bone's main axis, terminating at its outermost points. We developed three algorithms to determine the main axis and length of the bone. These methods are described below and depicted in Figure 3.

Statistical Shape Model:
In this method, we build and use 19 Statistical Shape Models (SSM), one per bone type, to place 100 points around each bone mask. These 100 points are placed evenly along the contour of each bone mask. Those contours are then aligned through iterative affine transformations and mean shape matching. We then perform Procrustes analysis with Principal Component Analysis (PCA) retaining 95% of the variance to get a model for the bone shapes. The same PCA transformations are then applied to contours of new segmentation masks, to ensure consistent point placement around the segmentation mask. The main axis on the test set is defined by the pair of points that minimizes the axis error in the training set.

Bounding Box:
This method applies a function to compute a bounding box enclosing the bone best.

The minimum-area bounding rectangle of a given set of points is determined by identifying the smallest enclosing rectangle that fully contains the shape while allowing arbitrary rotation. This process involves analyzing the convex hull of the points and applying the Rotating Calipers algorithm to find the rectangle with the least area. The resulting rectangle is characterized by its center coordinates, dimensions, and orientation angle. The centers of the bounding box's top, bottom, left, and right edges are then determined based on its orientation. Finally, the height and width of the bone are measured by calculating the Euclidean distances between these centers.

k-means:
In this method, we use k-means to find the main bone axis by splitting the pixels in the segmentation mask of each bone into two clusters. The idea is that one cluster will form at the proximal end, the other at the distal end of the bone mask. We use k-means clustering with $k = 2$, giving us the main axis by connecting the cluster centers. To find the length, we then rotate the segmentation mask to align the main axis with the y-axis of the image. The length can then be extracted from the extent of the bone in y-Direction.

We evaluate each bone length estimation method using the manual annotations (see Section 2.2.2) by calculating the average relative length and axis angle errors.

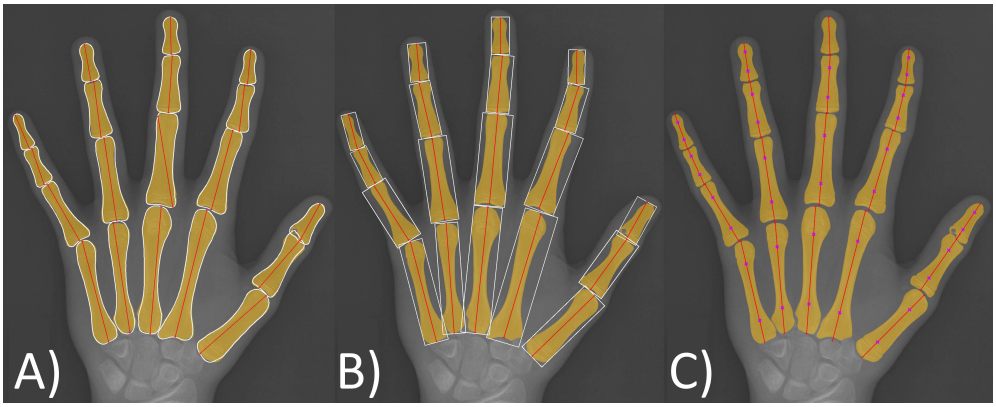

Fig. 3 Length measurements for a hand with normal bone morphology. The red lines show the lengths, the bounding boxes and contours are shown in white, and the cluster centers are pink crosses. The methods used are: A) Statistical Shape Model, B) Bounding Boxes, and C) k-means.

Given that the image scales across different imaging devices can be different, or the physical size of each pixel might be absent in some images, we need to work with dimensionless (scale-independent) values, i.e., length ratios, for the following analyses.

2.5 Comparison with Poznanski Reference Values

As mentioned in Section 2.1, Poznanski (1984) [8] provides sex- and age-specific M&P bone length ratios and their standard deviations for a healthy cohort. As a conservative test of *Auto-Bone-Caliper*, we consider Poznanski (1984) values as healthy references and compare our M&P length ratio measurement on a subset of the DHA dataset which was not part of our training process.

For this purpose, we compute pairwise bone length ratios derived from segmented bone measurements. The 19 M&P bones in each radiograph yield $N = 171$ unique pairwise ratios $R_{x,y}$ defined as

$$R_{x,y} = \frac{l_x}{l_y}, \tag{1}$$

where l_x and l_y denote the lengths of bones x and y in pixels, respectively.

We then quantify the relative difference of our automated measurement for a healthy cohort from the reference values of Poznanski $R^{\text{ref}}_{x,y}$ [8] via

$$\Delta R_{x,y} = \frac{(R_{x,y} - R^{\text{ref}}_{x,y})}{R^{\text{ref}}_{x,y}}. \tag{2}$$

We then analyze the distribution of these relative differences by measuring its mean and standard deviation and comparing it with the distribution from Poznanski (1984) using Wasserstein-1 distance.

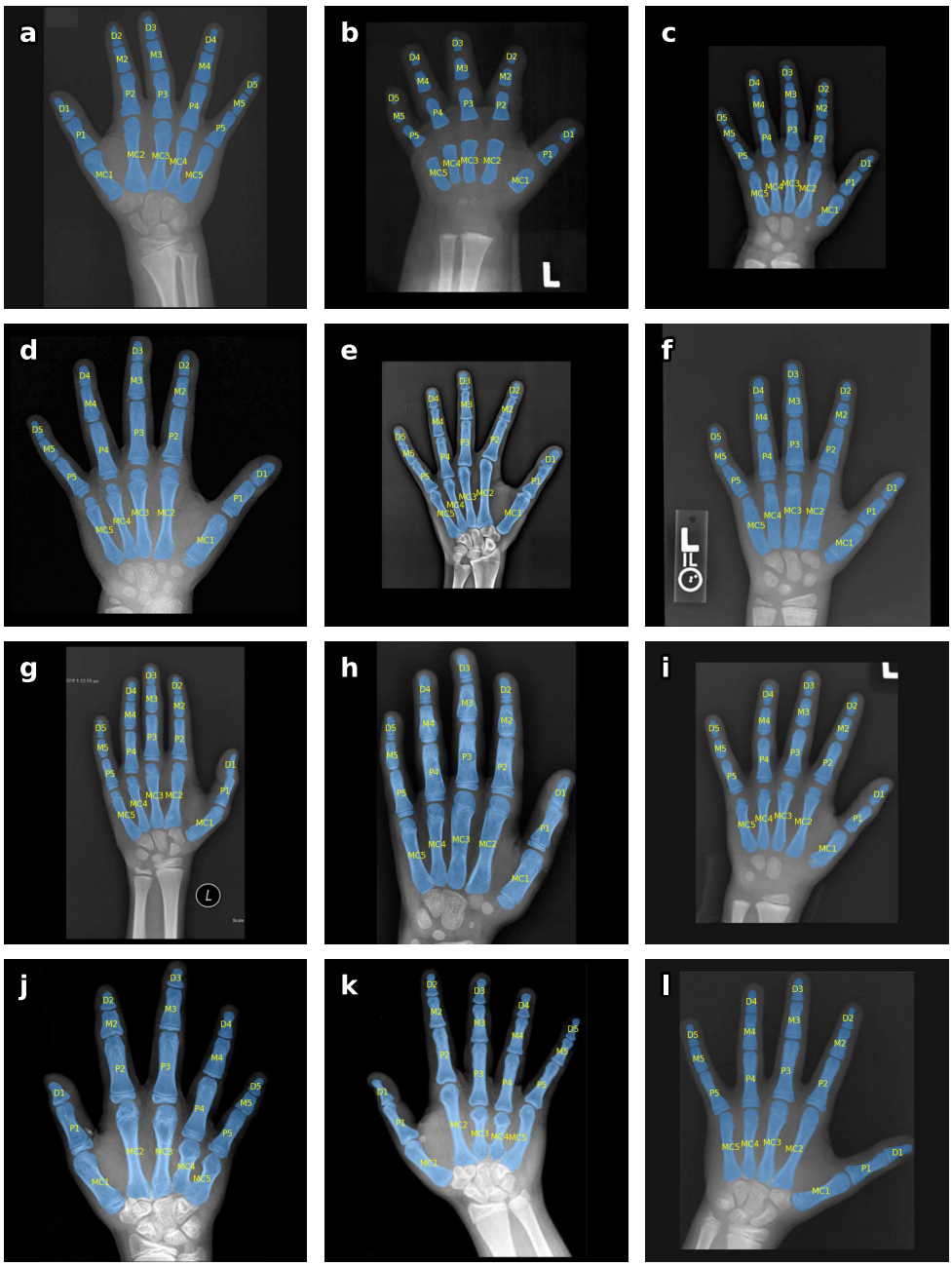

Fig. 4 Examples of performance of InstanceSAM on dysmorphic hand X-rays. a: achondroplasia, b: Down syndrome, c: hypochondroplasia, d: Leri-Weill syndrome, e: Marfan syndrome, f: McCune–Albright syndrome, g: Noonan Syndrome, h: Ollier disease, i: Proteus syndrome, j: pseudohypoparathyroidism, k: pseudopseudohypoparathyroidism, l: Turner syndrome. The raw images are obtained via the GMDB [31], note the variations in scales and quality. MC: Metacarpal bone, P: Proximal phalanx, M: Middle phalanx, D: Distal phalanx.

2.6 M&P Profiles

To obtain the M&P profiles, we first define the relative bone length $r_{m,x}$ of bone x in radiograph m as:

$$r_{m,x} = \frac{l_{m,x}}{\sum_{i=1}^{19} l_{m,i}} \tag{3}$$

where $l_{m,x}$ is the bone length in pixels. This essentially normalizes the length of each M&P bone in a hand with the sum of the lengths of all the 19 M&P bones in that hand, creating scale-independent values. Then we use the mean bone length ratio $\overline{r}_x^s$ across all radiographs of a sample s. This is then used to infer M&P profiles that can be visualized by plotting the z-score:

$$z_x^s = \frac{\overline{r}_x^s - \overline{r}_x^{\text{ref}}}{\sigma_x^{\text{ref}}} \tag{4}$$

using the mean $\overline{r}_x^{\text{ref}}$ and standard deviation σ_x from a reference healthy cohort.

3 Results

3.1 Bone Segmentation

The mean Dice score of InstanceSAM for normal hands was 98.7% and for dysmorphic bones in the GMDB test set was 95.0%. In Table 1 we also list the average DICE score for each of the 19 bones in the normal and dysmorphic cohorts. For the normal cohort, the mean DICE score for all bones is above 98.0% and the general trend is a decreasing DICE score for smaller bones. This trend shows itself more drastically in the dysmorphic cohort with the lowest DICE score, 84.3%, belonging to the smallest bone in the hand i.e. the distal phalanx of the 5th finger.

Examples of the performance of InstanceSAM segmentation on dysmorphic bones of 12 different disorders in the GMDB test set are shown in Figure 4.

3.2 Bone Length Prediction

Performances of the three automatic bone length measurement methods are provided in Table 2. The result reveals notable differences between normal and dysmorphic groups. Across all methods, the relative error is consistently lower for normal hand radiographs compared to dysmorphic ones. The k-means–based method achieved the best performance in both axis detection and length prediction, yielding errors of only 2.2% and 1.6°, respectively, for normal hands, and 4.5% and 2.9°, respectively, for dysmorphic hands. Therefore, we choose the k-means approach as the length measurement method of our *Auto-Bone-Caliper* and use it for all the subsequent analyses.

3.3 Comparison with Poznanski's reference values

For the length-ratio comparisons, the age and sex of the cases must be taken into account. Poznanski (1984) provides bone length ratios for age groups of 1, 4, 9, and

Table 1 Performance of InstanceSAM. The Mean Dice score for all 19 bones and for each bone type is listed for both normal and dysmorphic hand radiographs.

Bone	Normal [%]	Dysmorphic [%]
All 19 Bones	98.7	95.0
Metacarpal 1	99.0	97.1
Metacarpal 2	98.6	97.4
Metacarpal 3	98.3	97.5
Metacarpal 4	98.3	97.1
Metacarpal 5	98.8	97.1
Proximal Phalanx 1	99.0	95.6
Proximal Phalanx 2	99.3	97.4
Proximal Phalanx 3	99.4	97.8
Proximal Phalanx 4	99.3	97.5
Proximal Phalanx 5	99.2	95.7
Middle Phalanx 2	98.0	96.1
Middle Phalanx 3	98.4	94.5
Middle Phalanx 4	98.7	94.1
Middle Phalanx 5	98.3	94.9
Distal Phalanx 1	98.8	94.6
Distal Phalanx 2	98.3	90.7
Distal Phalanx 3	98.5	94.3
Distal Phalanx 4	98.5	91.1
Distal Phalanx 5	98.2	84.3

Table 2 Average relative length and axis angle errors in percentage [%] and degrees [°], respectively, for the three automatic length measurement methods.

Method	Normal (n=20)		Dysmorphic (n=30)		Total (n=50)	
	Length	Angle	Length	Angle	Length	Angle
Statistical Shape Model	3.1%	5.6°	6.4%	10.3°	5.1%	8.4°
Bounding Box	3.4%	3.5°	5.6%	5.1°	5.7%	4.6°
k-means	**2.2%**	**1.6°**	**4.5%**	**2.9°**	**3.6%**	**2.4°**

18 years. By setting N=10 as the minimum number of images required for this comparison, we restrict our analysis to cases aged 4, 9, and 18 years in the DHA dataset, comprising 10, 15, and 20 radiographs, respectively.

Figure 5 presents the distribution of relative differences (Eq. 2) for the 45 radiographs. For comparison, we also show a reconstructed distribution corresponding to Poznanski (1984), derived from the reported standard deviations under the assumption of Gaussianity. Treating Poznanski (1984) as the reference, its relative difference with respect to itself has zero mean and a standard deviation of ± 0.0593.

The distribution of relative differences obtained via our *Auto-Bone-Caliper* closely matches that of Poznanski (1984), with a mean of $-0.012^{+0.049}_{-0.072}$, corresponding to a standard deviation of ± 0.060. The Wasserstein-1 distance between the two distributions is 0.012, consistent with their nearly identical dispersions.

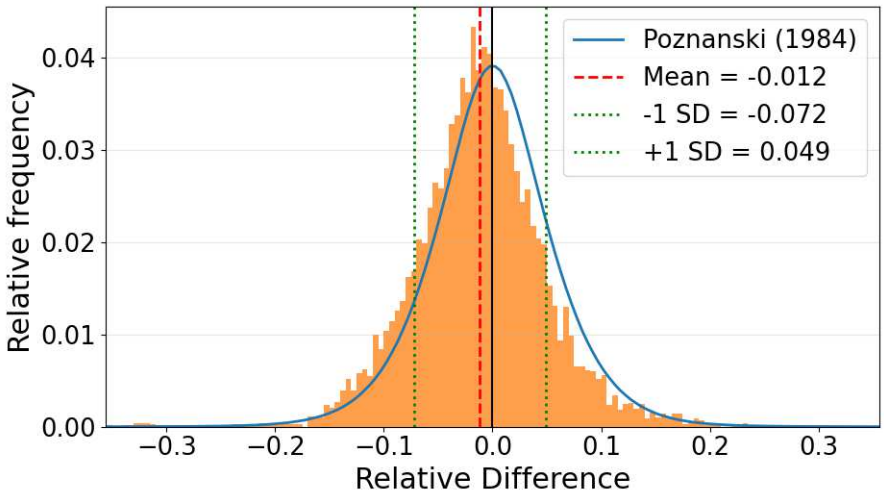

Fig. 5 Distribution of relative differences (Eq. 2) between predicted bone-length ratios on 45 radiographs (ages 4, 9, and 18) from the DHA dataset and Poznanski (1984) [8] reference values.

3.4 M&P Profiles

As an example of the use-case of our *Auto-Bone-Caliper* we show in Figure 6 M&P profiles of three different dysmorphic groups i.e. Turner syndrome, achondroplasia, and pseudohypoparathyroidism. The results from the previous section indicate a high general concordance between the reference values of Poznanski (1984)[8] and healthy cases in DHA. Therefore, we now take our automatic measurements from all the 330 DHA radiographs as reference (Z-score=0, see Eq. 4) for building M&P profiles.

The profiles show characteristic differences. Fore example, the most common abnormalities for pseudohypoparathyroidism, a (relative) shortening of D1, MC4, and MC5 [35], are visible through negative z-scores for these bones. For achondroplasia, absolute brachydactyly cannot be demonstrated, as we could only use relative bone lengths (Eq. (3)). The same holds for Turner syndrome; however, the characteristic (relative) shortening of MC4 characteristic of this syndrome [35] is clearly visible.

4 Discussion

Our segmentation model, InstanceSAM, achieves a substantially higher Dice score than those reported in previous studies [24, 25], both of which target bone age assessment. Owing to differences in experimental settings and datasets, a direct quantitative comparison is challenging. In the two-stage approach presented in [25], which was applied to data from a local hospital in China, hand bones are first extracted using the OSA-YOLOv5 network and subsequently segmented with GRU-UNet. Although that study also includes segmentation of the radius and ulna, only the M&P bones of the 1st, 3rd, and 5th digits are segmented. Based on their Table 6, their average Dice score for the M&P bones is around 95%. In [24], the authors propose a lightweight encoder–decoder semantic segmentation model with a multi-scale attention block, achieving a Dice score

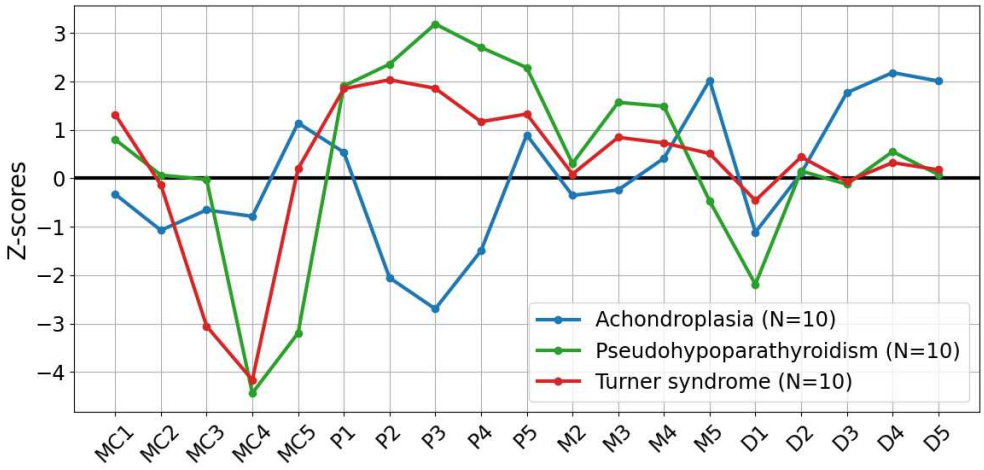

Fig. 6 The M&P profiles from the mean relative bone lengths for three dysmorphic groups: Achondroplasia, Pseudohypoparathyroidism, and Turner syndrome, with the DHA dataset (male and female) as reference (Z-score=0).

of 93.9% on the RSNA and MURA [36] datasets. It is important to note, however, that [24] additionally segments the radius, ulna, and carpal bones, whereas our work focuses exclusively on the M&P bones for which length-based measurements can be reliably defined and evaluated.

At the time of writing, we are not aware of any segmentation studies specifically addressing dysmorphic hand bones, precluding direct comparison. The lower performance of our segmentation model on dysmorphic bones compared to normal ones can be attributed to two main factors: (1) the model was trained exclusively on normal bone morphology, and (2) the hand X-rays in GMDB are sourced from figures in publications and therefore exhibit substantially reduced image quality both in resolution and intensity. Regarding the first point, the model's ability to generalize to dysmorphic cases despite never having been exposed to them is noteworthy, especially given the difficulty of assembling sufficiently large collections of radiographs representing all rare bone disorder groups for dedicated training. Nevertheless, future work should aim to evaluate InstanceSAM on larger dysmorphic cohorts and investigate fine-tuning strategies tailored to specific dysmorphic subtypes.

Regarding the length-measurement methods, our results show a clear advantage of the k-means–based approach—for both normal and dysmorphic bones—over the SSM and bounding-box techniques, which exhibit comparable performance. Accurate detection of the bone axis is essential for reliable length estimation, and the axis-prediction error of the k-means method is approximately half that of the other two approaches. For this reason, we selected k-means as the length-measurement component of *Auto-Bone-Caliper*.

A key result of this study is the evaluation of our *Auto-Bone-Caliper* on normal hand radiographs from the DHA dataset—which was not used during training—using

the reference values provided in the Poznanski (1984) [8] catalog. Two considerations are important here. First, the Poznanski reference values were derived from manual measurements performed with caliper on physical hand radiographs in the 1970s and 1980s, rather than digital images. Second, there is more than a decade between the data collection of the DHA dataset (1997–2008) and the acquisition and publication of the RSNA dataset, during which imaging protocols may have changed substantially. For these reasons, we regard this evaluation to be a conservative one.

Although based on a limited number of hand radiographs, the M&P profiles for achondroplasia, pseudohypoparathyroidism, and Turner syndrome presented in Figure 6 illustrate the potential of our automated system for supporting the clinical assessment of bone disorders. Larger cohorts will be needed to establish robust sex-, age, and ancestry-specific reference M&P profiles for these conditions. Nevertheless, *Auto-Bone-Caliper* can be used to detect significant outliers and anomalies.

5 Conclusion

This study introduces *Auto-Bone-Caliper*, an automated system for segmentation and length measurement of the M&P bones. We first present InstanceSAM, a two-stage segmentation pipeline that detects 19 M&P bones in pediatric hand X-rays using a fine-tuned YOLO model, followed by segmentation with SAM. Although InstanceSAM was trained exclusively on normal hand radiographs, it generalizes successfully to dysmorphic bones from patients with various bone disorders.

We then designed and evaluated three bone-length measurement methods: statistical shape modeling, bounding box estimation, and a k-means–based approach. The k-means method demonstrated superior accuracy and was therefore selected as the length-measurement component of our *Auto-Bone-Caliper*.

To assess the reliability of *Auto-Bone-Caliper*, we compared bone-length ratios derived from normal hand radiographs in the independent DHA dataset with the reference values for healthy individuals reported in the Poznanski (1984) [8] catalog. The close agreement between the two supports the validity of *Auto-Bone-Caliper*.

Bone-length analysis in hand radiographs, such as in M&P Profile analysis, is known to aid in the diagnosis and evaluation of patients with various skeletal disorders; however, the burden of manual measurements has limited its routine clinical use. Automated solutions such as the *Auto-Bone-Caliper* have the potential to streamline and standardize the process, enabling objective and scalable assessment for both diagnostic and monitoring purposes and also to function as an accessible screening tool for detecting patients with bone anomalies.

Data Availability

The data used in this study were obtained from four publicly available datasets: RSNA, GMDB, DHA, and Poznanski (1984) catalog.

Code Availability

Implementations of the methods are available on our GitHub repository github.com/bone2gene/Auto-Bone-Caliper.

Competing Interests

P.G., A.K., E.B., P.S., and B.J. are co-inventors on a pending patent application containing parts of the methods described in this manuscript.

Funding

J.G. received funding from the Federal Ministry of Education and Research within the project "BNTrAinee" (funding code 16DHBK1022). This study is part of the Bone2Gene project (PI: B. Javanmardi) funded by the German Federal Ministry of Research, Technology, and Space within the GO-Bio initial program.

Authors Contributions

P.G. developed the segmentation model. P.G. and P.S. performed the manual annotations. A.K. developed the k-means and the SSM methods and F.H. developed the bounding box method. E.B., P.G., A.K., P.S., F.H., J.G., A.H., and B.J. analyzed the data and the results. The first draft of the manuscript was written by P.G., A.K., E.B., P.S., F.H., A.H., and B.J.. E.A., T.S., K.M., and P.K. provided intellectual input on computer vision, clinical genetics, pediatric endocrinology, as well as translational and ethical aspects. B.J. conceived the study. All co-authors contributed to the study design and have critically reviewed the manuscript.

References

[1] Offiah, A.C., Hall, C.M.: Radiological diagnosis of the constitutional disorders of bone. as easy as a, b, c? Pediatric Radiology **33**(3), 153–161 (2002) https://doi.org/10.1007/s00247-002-0855-8

[2] Creo, A.L., Schwenk, W.F.I.: Bone age: A handy tool for pediatric providers. Pediatrics **140**(6), 20171486 (2017) https://doi.org/10.1542/peds.2017-1486

[3] Rassmann, S., Keller, A., Skaf, K., Hustinx, A., Gausche, R., Ibarra-Arrelano, M.A., Hsieh, T.-C., Madajieu, Y.E.D., Nöthen, M.M., Pfäffle, R., Attenberger, U.I., Born, M., Mohnike, K., Krawitz, P.M., Javanmardi, B.: Deeplasia: deep learning for bone age assessment validated on skeletal dysplasias. Pediatric Radiology **54**(1), 82–95 (2024) https://doi.org/10.1007/s00247-023-05789-1

[4] Rassmann, S., Abashishvili, L., Melikidze, E., Sukhiashvili, A., Lartsuliani, M., Chkhaidze, I., Tskhvediani, N., Gordeziani, T., Kvaratskhelia, E., Kheladze, N., Rekhviashvili, M., Rodonaia, S., Sukhitashvili, N., Urushadze, N., Krawitz, P.,

Tkemaladze, T., Javanmardi, B.: Population-specific calibration and validation of an open-source bone age ai. Scientific Reports **15**, 32673 (2025) https://doi.org/10.1038/s41598-025-20148-w

[5] Spranger, J.W., Brill, P.W., Hall, C., Nishimura, G., Superti-Furga, A., Unger, S.: Bone Dysplasias: An Atlas of Genetic Disorders of Skeletal Development. Oxford University Press, Oxford (2018). https://doi.org/10.1093/med/9780190626655.001.0001

[6] Cavallo, F., Mohn, A., Chiarelli, F., Giannini, C.: Evaluation of bone age in children: A mini-review. Frontiers in Pediatrics **9**, 580314 (2021) https://doi.org/10.3389/fped.2021.580314

[7] Garn, S.M., Hertzog, K.P.., Poznanski, A.K., Nagy, J.M.: Metacarpophalangeal length in the evaluation of skeletal malformation. Radiology **105**(2), 375–381 (1972) https://doi.org/10.1148/105.2.375

[8] Poznanski, A.K.: The Hand in Radiologic Diagnosis: With Gamuts and Pattern Profiles, 2nd edn., p. 925. W. B. Saunders, Philadelphia (1984)

[9] Butler, M.G., Meaney, F.J., Kaler, S.G.: Metacarpophalangeal pattern profile analysis in clinical genetics: An applied anthropometric method. American Journal of Physical Anthropology **70**(2), 195–201 (1986) https://doi.org/10.1002/ajpa.1330700206

[10] Butler, M.G., Gale, D.D., Meaney, F.J., Opitz, J.M., Reynolds, J.F.: Metacarpophalangeal pattern profile analysis in diastrophic dysplasia. American Journal of Medical Genetics **28**(3), 685–689 (1987) https://doi.org/10.1002/ajmg.1320280316

[11] Takahashi, Y., Imaizumi, K., Takada, F., Kuroki, Y.: Metacarpophalangeal pattern profile analysis in 14 japanese children with sotos syndrome. Japanese Journal of Human Genetics **39**(1), 187–191 (1994) https://doi.org/10.1007/bf01915954

[12] Laurencikas, E., Söderman, E., Davenport, M., Jorulf, H., Sävendahl, L.: Metacarpophalangeal pattern profile analysis as a tool for early diagnosis of turner syndrome. Acta Radiologica **46**(4), 424–430 (2005) https://doi.org/10.1080/02841850510020905

[13] Laurencikas, E., Sävendahl, L., Jorulf, H.: Metacarpophalangeal pattern profile analysis: useful diagnostic tool for differentiating between dyschondrosteosis, turner syndrome, and hypochondroplasia. Acta Radiologica **47**(5), 518–524 (2006) https://doi.org/10.1080/02841850600635939

[14] Poznanski, A.K., Gartman, S.: A bibliography covering the use of metacarpophalangeal pattern profile analysis in bone dysplasias, congenital malformation

syndromes, and other disorders. Pediatric Radiology **27**(5), 358–365 (1997) https://doi.org/10.1007/s002470050152

[15] Levitt, T.S., Hedgcock, M.W.J., Vosky, D.N., Shadle, V.M.: Model-based prediction of phalanx radiograph boundaries. In: Medical Imaging · Image Processing (Proc. SPIE Vol. 1898), pp. 670–678 (1993). https://doi.org/10.1117/12.154555

[16] Chassignet, P., Nitescu, T.-M., Hassan, M., Stanescu, R.: Hand radiograph analysis for fully automatic bone age assessment. In: Image Processing (Proc. 1999) [MIIP?] (1999). https://dblp.org/rec/conf/miip/ChassignetNHS99

[17] Kauffman, J.A., Slump, C.H., Bernelot Moens, H.J.: Segmentation of hand radiographs by using multi-level connected active appearance models. In: Medical Imaging Conference 2005 (MIC), Proceedings of SPIE Vol. 5747, pp. 1571–1581 (2005). https://doi.org/10.1117/12.595640

[18] Hsieh, C.W., Chen, C.Y., Jong, T.L., Liu, T.C., Chiu, C.H.: Automatic segmentation of phalanx and epiphyseal/metaphyseal region by gamma parameter enhancement algorithm. Measurement Science Review **12**(1), 21–27 (2012) https://doi.org/10.2478/v10048-012-0003-z

[19] Dendere, R., Kabelitz, G., Douglas, T.S.: Model-based segmentation of the middle phalanx in digital radiographic images of the hand. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 3702–3705 (2013). https://doi.org/10.1109/EMBC.2013.6610347

[20] Lindner, C., Thiagarajah, S., Wilkinson, J.M., Wallis, G.A., Cootes, T.F., Consortium: Accurate bone segmentation in 2d radiographs using fully automatic shape model matching based on regression-voting. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, vol. 16, Part 2, pp. 181–189. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_23

[21] Simu, S.A., Lal, S., Fadte, K., Harlapur, A.: Fully automatic segmentation of phalanges from hand radiographs for bone age assessment. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization **7**(1), 59–87 (2019) https://doi.org/10.1080/21681163.2017.1416491

[22] Rambojun, A., Tillett, W., Campbell, N., Shardlow, T.: Development of an automated segmentation algorithm to identify bones of the hand. Annals of the Rheumatic Diseases **78**(Suppl.2), 589 (2019) https://doi.org/10.1136/annrheumdis-2019-eular.2336

[23] Liu, Z., Hu, Z., Wu, T., Ye, G., Tang, Y., Zeng, Z., Ouyang, Z., Li, Y.: Bone age recognition based on mask r-cnn using xception regression model. Frontiers in Physiology **14**, 1062034 (2023) https://doi.org/10.3389/fphys.2023.1062034

[24] Nagaraju, Y., Venkatesh, Thanu Shree Yadav, P.R., Vaishnavi, A., Tejashree, S.V.: A feature-enhanced multiscale attention approach for automated hand bone segmentation. Multimedia Tools and Applications (2024) https://doi.org/10.1007/s11042-024-19647-1

[25] Du, H., Wang, H., Yang, C., Kabalata, L., Li, H., Qiang, C.: Hand bone extraction and segmentation based on a convolutional neural network. Biomedical Signal Processing and Control **89**, 105788 (2024) https://doi.org/10.1016/j.bspc.2023.105788

[26] Areeckal, A.S., Sumam David, S., Kocher, M., Jayasheelan, N., Kamath, J.: Fully automated radiogrammetric measurement of third metacarpal bone from hand radiograph. In: 2016 International Conference on Signal Processing and Communications (SPCOM 2016). IEEE, Piscataway, NJ, USA (2016). https://doi.org/10.1109/SPCOM.2016.7746608

[27] Haghnegahdar, A., Pakshir, H.R., Zandieh, M., Ghanbari, I.: Computer assisted bone age estimation using dimensions of metacarpal bones and metacarpophalangeal joints based on neural network. Journal of Dentistry **25**(1), 51–58 (2024) https://doi.org/10.30476/dentjods.2023.95629.1882

[28] Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., Kitamura, F.C., Thodberg, H.H., Chen, L., Shih, G., Andriole, K., Kohli, M.D., Erickson, B.J., Flanders, A.E.: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019) https://doi.org/10.1148/radiol.2018180736 . PMID: 30480490

[29] Larson, D.B., Chen, M.C., Lungren, M.P., Halabi, S.S., Stence, N.V., Langlotz, C.P.: Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology **287**(1), 313–322 (2018) https://doi.org/10.1148/radiol.2017170236

[30] Hsieh, T.-C., Bar-Haim, A., Moosa, S., Ehmke, N., Gripp, K.W., Pantel, J.T., Danyel, M., Mensah, M.A., Horn, D., Rosnev, S., Fleischer, N., Bonini, G., Hustinx, A., Schmid, A., Knaus, A., Javanmardi, B., Klinkhammer, H., Lesmann, H., Sivalingam, S., Kamphans, T., Meiswinkel, W., Ebstein, F., Krüger, E., Küry, S., Bézieau, S., Schmidt, A., Peters, S., Engels, H., Mangold, E., Kreiß, M., Cremer, K., Perne, C., Betz, R.C., Bender, T., Grundmann-Hauser, K., Haack, T.B., Wagner, M., Brunet, T., Bentzen, H.B., Averdunk, L., Coetzer, K.C., Lyon, G.J., Spielmann, M., Schaaf, C.P., Mundlos, S., Nöthen, M.M., Krawitz, P.M.: Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors. Nature Genetics **54**(3), 349–357 (2022) https://doi.org/10.1038/s41588-021-01010-x

[31] Hsieh, T.-C., Lesmann, H., Hustinx, A., Moosa, S., Marchi, E., Martin, M.d.P.C., Abdelrazek, I., Pantel, J.T., Klinkhammer, H., Hagen, M.t., Thong, M.-K.,

Mazlan, R.A.B., Tae, S.K., Kamphans, T., Meiswinkel, W., Li, J.-M., Javanmardi, B., Knaus, A., al.: Gestaltmatcher database - a global reference for facial phenotypic variability in rare human diseases (2024) https://doi.org/10.21203/rs.3.rs-4438861/v1

[32] Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Computerized Medical Imaging and Graphics **31**(4/5), 322–331 (2007) https://doi.org/10.1016/j.compmedimag.2007.00027

[33] Wang, C.-Y., Yeh, I.-H., Mark Liao, H.-Y.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, pp. 1–21. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72751-1_1

[34] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Piscataway, NJ, USA (2023). https://doi.org/10.1109/ICCV51070.2023.00371

[35] Chavhan, G.B., Miller, E., Mann, E.H., Miller, S.F.: Twenty classic hand radiographs that lead to diagnosis. Pediatric Radiology **40**(5), 747–761 (2010) https://doi.org/10.1007/s00247-009-1520-2

[36] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs (2017)