Deeplasia: prior-free deep learning for pediatric bone age assessment robust to skeletal dysplasias

Sebastian Rassmann[1], Alexandra Keller[2], Kyra Skaf[3], Alexander Hustinx[1], Ruth Gausche[4], Miguel A. Ibarra-Arrelano[1], Tzung-Chien Hsieh[1], Yolande E. D. Madajieu[3], Markus M. Nöthen[5], Roland Pfäffle[6], Klaus Mohnike[3], Peter M. Krawitz[1], and Behnam Javanmardi*[1]

[1]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Germany
[2]Kinderzentrum am Johannisplatz, Leipzig, Germany
[3]Medical Faculty, Otto-von-Guericke-University Magdeburg, Germany
[4]CrescNet - Wachstumsnetzwerk, Medical Faculty, University Hospital Leipzig, Germany
[5]Institute of Human Genetics, University Hospital Bonn, Germany
[6]Department for Pediatrics, University Hospital Leipzig, Germany

Skeletal dysplasias collectively affect a large number of patients worldwide. The majority of these disorders cause growth anomalies. Hence, assessing skeletal maturity via determining the bone age (BA) is one of the most valuable tools for their diagnoses. Moreover, consecutive BA assessments are crucial for monitoring the pediatric growth of patients with such disorders, especially for timing hormone treatments or orthopedic interventions. However, manual BA assessment is time-consuming and suffers from high intra-and inter-rater variability. This is further exacerbated by genetic disorders causing severe skeletal malformations. While numerous approaches to automatize BA assessment were proposed, few were validated for BA assessment on children with abnormal development. In this work, we present Deeplasia, an open-source prior-free deep-learning ensemble approach. After training on the public RSNA BA dataset, we achieve state-of-the-art performance with a mean absolute difference (MAD) of 3.87 months based on the average of six different reference ratings. Next, we demonstrate that Deeplasia generalizes to an unseen dataset of 568 X-ray images from 189 patients with molecularly confirmed diagnoses of seven different genetic bone disorders (including Achondroplasia and Hypochondroplasia) achieving a MAD of 5.84 months w.r.t. to the average of two references. Further, using longitudinal data from a subset of the cohort (149 images), we estimate the test-retest precision of our model ensemble to be at least at the human expert level (2.74 months). We conclude that Deeplasia suits assessing and monitoring the BA in patients with skeletal dysplasias.

1 Introduction

The estimation of bone age (BA), which evaluates skeletal maturity, is one of the most important and valuable tools in assessing children's health. Usually, it is one of the first steps in the diagnosis of pediatric growth disorders [1]. In particular, for conditions in which hormonal therapy or orthopedic interventions are being considered, the timing of the treatment depends on the assessed BA [2]. The BA can be estimated by observing the ossification centers of a child's skeleton. The main body parts used for BA assessment are the hands, wrists, and knees. BA estimates from the hand and wrist are more closely correlated with the child's overall growth progress and puberty onset than estimates from the knee. Hence, the BA estimated from hand X-rays is more effective in assessing delayed or advanced growth [3] and is therefore used as a routine diagnostic and monitoring method [4, 5]. The Greulich-Pyle [6] (hereafter GP) and Tanner-Whitehouse [7, 8, 9] (hereafter TW) are the two most commonly used hand and wrist BA estimation methods. While the TW method is considered to be more accurate, the GP method is generally regarded to be faster [10]. Nevertheless, both methods are

*Corresponding author. E-mail: behnam.javanmardi@uni-bonn.de

time-consuming and show high degrees of inter- and intra-rater variability [10, 11].

Numerous Machine Learning (ML) approaches were proposed to automate BA assessment, especially based on a publicly available dataset released in 2017 by the Radiological Society of North America (RSNA) for their pediatric BA challenge [12, 13]. While an approach using end-to-end Deep Learning (DL) without any human prior or a particularly task-specific design won the competition [12, 14], ML approaches emphasizing anatomical features used in human BA assessment showed some improvement in more recent studies (e.g. [15, 16, 17, 18]).

A major indication to perform BA assessments are suspected growth or developmental anomalies. This is often connected with the phenotype of a skeletal dysplasia [19] due to the common causation by a genetic disorder. Although these disorders are individually rare [20], collectively they affect a large number of children [21] with an estimated total number of around 25 million worldwide. Especially in such patients, reliable and precise BA estimations are vital, both, to obtain an initial diagnosis and to monitor the maturation progress over time [22]. As skeletal dysplasias obstruct precise BA assessment [4], automated and ML-based tools require specific validation on patients with such disorders. However, this problem is still understudied and many approaches to automatic BA assessment were developed for and tested on datasets composed of predominantly normally-developing children. The public dataset released as part of the 2017 BA challenge contains only 0.21% cases of reported skeletal dysplasias [12, 13] and the recent study by [11] included $< 1.4\%$ patients with congenital diseases. [23] and [24] proposed and tested DL methods on patients with abnormal growth, however, their study was limited to Korean and Chinese populations, respectively, and the test sets included no or only small numbers ($n < 10$) of images from patients with severe skeletal malformations such as Achondroplasia (ACh) and Hypochondroplasia (HyCh). Furthermore, the altered hand morphology impedes the automatic identification of individual bones or regions of interest (ROIs) in dysplastic hands. Hence, ML approaches relying on the identification of particular ROIs within the hand might be unsuitable for precise BA assessment on dysplastic hands. For example, the routinely-used BA assessment tool BoneXpert [25, 18] rejects malformed bones and fails to generalize to patients with skeletal dysplasias such as ACh and HyCh [26].

In this work, we design and introduce Deeplasia: a DL-based tool specifically validated on dysmorphic hands of patients with skeletal dysplasias. Given the intrinsic scarcity of data from patients with rare diseases, we aim at presenting an open-source tool that while trained on data of normal hands can reliably be used for assessing BA of patients with rare bone diseases. To this end, and to overcome the dependence on potentially corrupted individual features commonly used for human BA assessment, we employ an end-to-end DL approach. This allows learning a broader range of features rather than emphasizing or completely depending on predefined shapes or ROIs. We demonstrate that our prior-free learning approach is at least as powerful as other approaches which incorporate priors from human BA assessment and require extensive additional annotation.

We show that Deeplasia

- achieves state-of-the-art (SOTA) performance on the RSNA BA test dataset composed of predominantly healthy patients
- generalizes to patients from unseen cohorts and with a variety of genetically-confirmed skeletal dysplasias
- is demonstrated to be applicable to longitudinal data from patients with skeletal dysplasias for the purpose of progressive growth monitoring.
- can assess the BA in less than 10 seconds on consumer-grade hardware without GPU acceleration.

This paper is organized as follows: In Section 2, we introduced the three different datasets that we use in our study, and in Section 3, we present the details of our methodology for all the subsequent developments and analyses. We present the results and discussion in Section 4, and we finally conclude in Section 5.

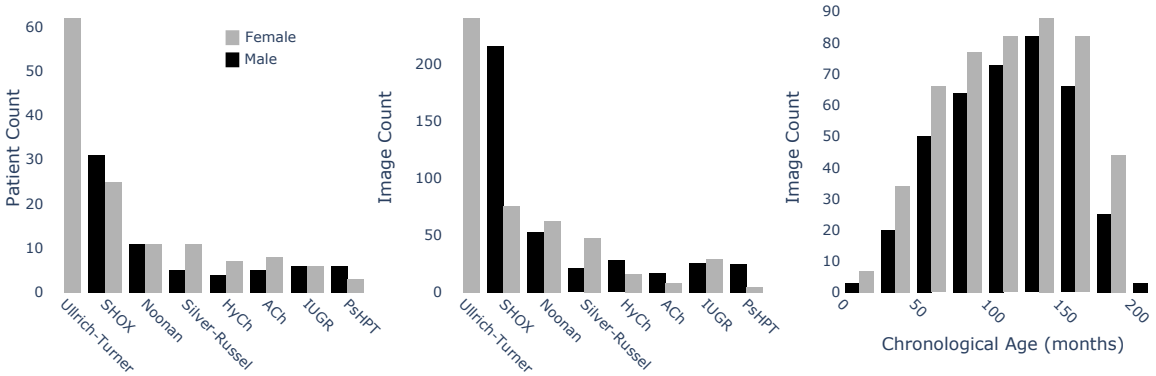

Figure 1: The GDBD distribution of patient and image counts between disorders, and the cross-cohort distribution of chronological age.

2 Data

2.1 RSNA Bone Age Dataset

We use the BA dataset published in 2017 by the RSNA for their Pediatric BA ML Challenge [12, 13]. We use the original training, validation, and test sets containing 12,611, 1,425, and 200 images, respectively. The training and validation sets were obtained from Children's Hospital Colorado (Aurora, CO, USA) and Lucile Packard Children's Hospital at Stanford (Palo Alto, CA, USA), while the test set is only from the latter hospital. The mean chronological age of patients in the training and validation set is 10.8 ± 3.5 years, and that of the test set is 11.3 ± 3.8 years [12]. For each image, the sex and a ground truth GP BA estimate are provided. For determining the ground truth BA, one estimate from the original clinic of the data, a second estimate from the same rater at least one year later, and four independent estimates were obtained. To form the final consensus BA estimate, a weighted mean based on the performance of each reviewer is calculated (for more details see [13]). The mean estimated BA of the training and validation set is 10.6 ± 3.4 years and that of the test set is 11.0 ± 3.6 years [13], and the distribution of ground truth BA for males and females is similar in the training, validation, and test sets.

2.2 Los Angeles Digital Hand Atlas

As an external test set for normally-developing children, the publicly released Los Angeles Digital Hand Atlas (DHA, [27, 28]) was used. It consists of 1,390 images acquired between 1997 and 2008 at the Children's Hospital Los Angeles, USA. The study cohort included four ethnicities and ground truth BA estimates were obtained by two raters using the GP atlas. The ground truth BA was defined as the average of the two ratings. We excluded seven images due to lacking or completely implausible ground truth BA assessment (BA of 99 years, BA of 0 years for children of 9 years of chronological age, and two images with a difference to a third manual assessment by KM of > 2 years).

2.3 German Dysplastic Bone Dataset

To compile a dataset for validating the BA prediction models on dysplastic hands, we retrospectively (2006 - 2022) collected hand X-rays from patients referred to the pediatric endocrinology of two German university hospitals (Magdeburg and Leipzig) due to a suspected growth disorder. The X-ray images were acquired as hard copies and thereafter digitized. The study was approved by the ethics committee of the medical faculties of the universities Magdeburg (vote 27/22) and Leipzig (vote 121/22-ek). We term this dataset the German Dysplastic Bone Dataset (GDBD). In total, it contains 568 hand X-ray images from 189 patients with molecularly confirmed diagnoses of one of the following disorders: ACh, HyCh, Pseudohypoparathyroidism (PsHPT), Noonan, Silver-Russel (SRS), and Ullrich-Turner (UTS) syndromes, and a mutation in the SHOX gene. Further, 55 images from 12 patients with intrauterine growth restriction (IUGR) were included. The number of images and patients as well as the distribution of their chronological age are shown in Fig. 1. An example X-ray per each disorder is shown in Fig. 2. We also included 79 images from 79 children without a diagnosed growth disorder, but who had been referred to pediatric endocrinologists due to a suspected growth disorder. The ethnic

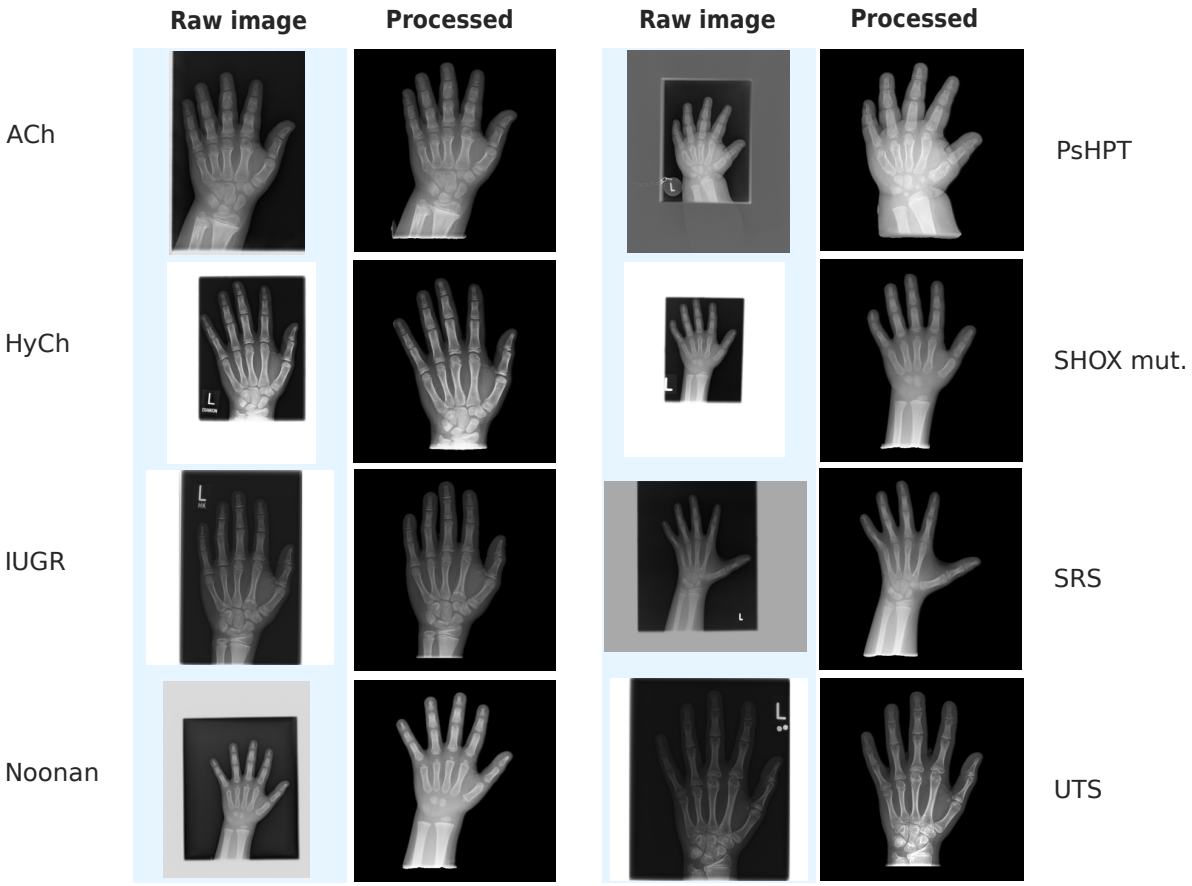

Figure 2: Example hand X-ray for each bone dysplasia in the GDBD. For each image, the original raw image and the preprocessed version (see Section 3.1) are shown. Note, that the hands have a wide range of relative scales within the images, image qualities vary, and show artifacts such as labels and white or gray backgrounds caused by scanning. SHOX mut. stands for mutation of the SHOX gene.

background of the GDBD patients is not available, however, we suspect a large portion of them to be Caucasians.

The BA reference grading for the GDBD was obtained using the GP standard by two pediatric endocrinologists with more than 40 (KM) and more than 20 (AK) years of clinical experience. For 643 out of 702 images, one of the assessments was obtained from the initial clinical report. The BA ratings for the remaining images and the second reference ratings were obtained from a dedicated session, in which the images were presented a) using the same preprocessing procedures as for testing the models (see Section 3.1), b) in a randomized order, and c) blinded for the chronological age, the clinical report, and the diagnosis.

3 Method

In this section, we describe our methodology for preprocessing and preparation of images (3.1), DL model training and implementation (3.2), longitudinal analysis of BA (3.3), and metrics for assessing the performance of the models (3.4).

3.1 Image background removal and preprocessing

Imaging and scanning induce artifacts to the input images, e.g. department-specific markers or white borders surrounding the X-ray carrier in the scan. Such artifacts (which are present in many of the images in our GDBD set) have been shown to bias DL models (e.g. [29]). Further, high-intensity borders can potentially skew the image normalization for inference. To prevent these problems, we trained *TensorMask* [30] and *Efficient-UNet* models [31] to automatically extract the hand from the

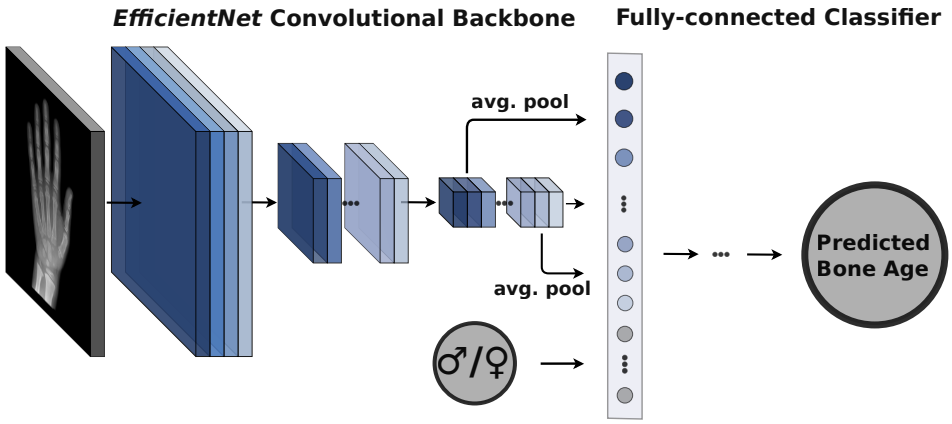

Figure 3: Model architecture for bone age prediction. The gray-scale input image is passed through an *Efficientet* backbone model. The obtained features are combined with an inflated representation of the sex and passed into a fully-connected network to obtain the bone age.

scan by masking the background. The details of our hand segmentation are described in Appendix A. All images were standardized to a mean intensity of zero and a standard deviation of one. We chose this standardization over the commonly used min-max normalization as the latter would be highly susceptible to few high-intensity pixels e.g. due to remaining scanning or imaging artifacts. Fig. 2 shows some examples of the results of our preprocessing on dysplastic hands. It can be seen how the preprocessing removed the backgrounds and cropped the GDBD images to the predicted mask (to assert a constant scale of the hand in the follow-up analysis). The training masks and the code for the hand segmentation are publicly available via [32] and github.com/aimi-bonn/hand-segmentation, respectively.

3.2 Bone age model training and implementation

As a baseline approach for the BA model, we follow the design principle winning the 2017 RSNA Pediatric BA ML Challenge [13, 14]. The model architecture, outlined in Fig. 3, is composed of a fully Convolutional Neural Network (CNN) as a feature extractor, channel-wise average pooling of the extracted features, and concatenation of a representation of the patient sex inflated to 32 neurons[1]. The results are passed through a variable set of fully-connected (FC) layers to achieve a final prediction. We employ *EfficientNets* [33] as backbone feature extractors. In comparison to previously proposed end-to-end learning methods [14, 34], our applied average pooling reduces the dimensionality of the learned features and, thus, decreases the model size. For example, the largest of our BA models has a feature dimensionality of $1,792$ resulting in a total network size of $23 * 10^6$ parameters, while the configuration proposed by [34] uses a feature dimensionality of $33,712$ and $82 * 10^6$ parameters. All the details of our model training are described in Appendix B, and the code for training the BA models is available at github.com/aimi-bonn/Deeplasia.

The models were implemented in *PyTorch* ([35], v1.10) using the *lightning* framework (v1.6, [36]). We used the *Detectron2* (v0.4, [37]) implementation of *Tensormask* [30]. Data augmentation was conducted using the *Albumentations* library (v1.1, [38]). Image pre- and postprocessing was conducted in *OpenCv* (v4.5, [39]).

3.3 Longitudinal analysis

To detect small changes in the developmental process like slow-downs or growth spurts, BA measurements are required to have high test-retest reliability. Directly measuring the test-retest reliability would require a dedicated image session which would be unethical due to the unnecessary radiation exposure. However, assuming linear progress of the BA over time, the test-retest reliability can be estimated retrospectively from regular check-ups within the testing cohort. For estimating the upper bound of the expected error in assessing the BA, the method proposed by [40] was used. No patients were excluded due to therapies or other interventions. Additionally, the potentially variable growth

[1] We further discuss the effect of patients's sex on BA assessment in Appendix F.

patterns due to the disorders of the patients included in the analysis might give a non-linear growth pattern. To account for this, we set the maximum time difference for the derivation triplets to 14 months, the lowest threshold to achieve $n \geq 100$ triplets. For analyzing the rater performance, only triplets derived from either the clinical ratings or from a single rater within the blinded re-rating session were included to avoid rater-rater biases or biases between clinical and blinded reviews.

3.4 Metrics and statistical analysis

For model selection and benchmarking the mean absolute difference (MAD) was used. It is calculated as the L_1-norm of the difference between predicted BA $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$ and the respective ground truth $Y = (y_1, y_2, \ldots, y_n)^T$:

$$MAD\,(\hat{Y}, Y) = ||\hat{Y} - Y||_1 = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{1}$$

Further, the root-mean-square error (RMSE) was used as a metric that is more sensitive to outliers. It is defined as:

$$RMSE\,(\hat{Y}, Y) = \sqrt{\frac{1}{n}||\hat{Y} - Y||_2^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

For the statistical analysis, we assume the signed error to be normally distributed and, thus, derive the confidence intervals of the RMSE from the corresponding χ^2 distribution. As an additional, clinically more interpretable metric, we define a 1-year accuracy. Let 1_{cond} denote the indicator function (a function that evaluates to 1 if and only if *cond* is true) and assume the BAs $\hat{Y}$ and Y to be denoted in years, then

$$Acc_{1y}(\hat{Y}, Y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{|\hat{y}_i - y_i| \leq 1} \tag{3}$$

Note that we do not conduct a symbolic perturbation, so the measure is conservative w.r.t. the model performance as the models are, in contrast to human raters, unlikely to assign integer BAs.

4 Results and Discussion

4.1 Model explorations for building the ensemble

To build the model ensemble for Deeplasia, we experimented with three different training conditions, a) the *baseline*: *EfficientNet-b0* with 512×512 input resolution, b) *large CNN*: *EfficientNet-b4* with 512×512 input resolution, and c) *high-resolution*: *EfficientNet-b0* with 1024×1024 input. For each of these conditions, we trained models with three sets of FC layers: $[256]$, $[512, 512]$, $[1024, 1024, 512, 512]$. Therefore, in total, we trained nine CNN models for BA estimation. The details of our training experimentations are described in Appendix C.

To choose the models for building an ensemble, we analyzed the pairwise correlations between the predicted BA of these nine models, visualized in Fig. 4. This revealed that there is a higher correlation between the predictions of the models within the same training condition (i.e. the *baseline*, *large CNN*, and *high-resolution*) compared to the ones across these conditions. As dissimilar prediction patterns in a model ensemble are advantageous due to partial compensation of predictive errors, it is beneficial to construct an ensemble composed of models across different training conditions. Consequently, we choose to pick the best-performing model from each of the three training conditions for building our model ensemble. The ensembled BA is the average of these three models.

4.2 Performance on the RSNA test set

On the RSNA test set comprising 200 X-rays, Deeplasia achieved a MAD of 3.87 months, RMSE of 5.14 months, and a 1-year accuracy of 98.5%. These results, listed in Table 1, are on-a-par with the current SOTA (3.91 months, [16]) and commercial tools cleared for clinical use (4.1 months, [25, 18]). Interestingly, even the individual three models achieved test accuracies (MAD of 4.2, 4.1, and 4.3 months) comparable to other approaches incorporating human priors.

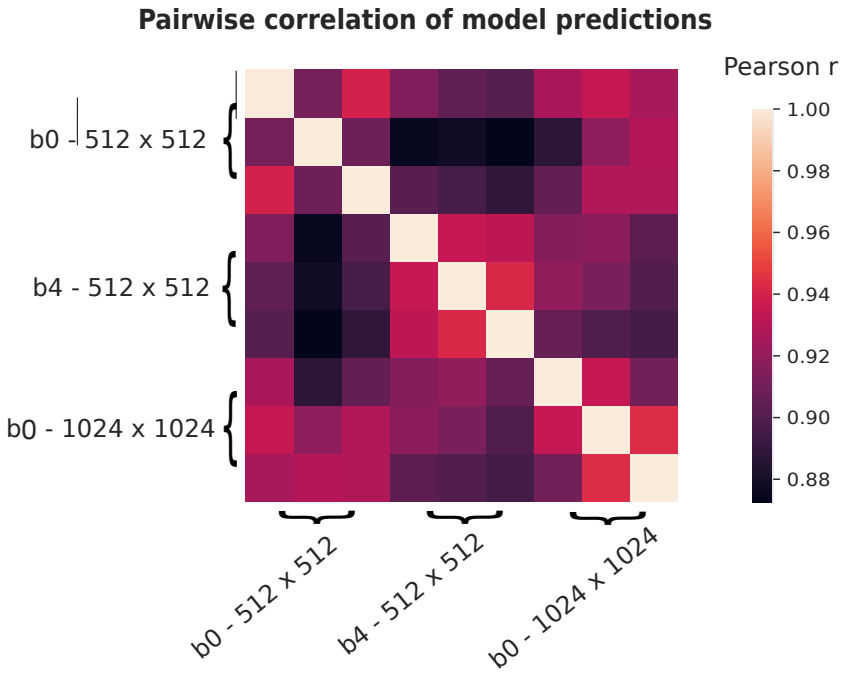

Figure 4: Pairwise correlations of the predicted bone ages (BAs) on the RSNA validation dataset of nine models with different *EfficientNet* backbone models (*EfficientNet-b0* and *-b4*) and at different image resolutions (512×512 and 1024×1024). For each combination of backbone and resolution models with various sets of FC layers ([256], [512, 512], [1024, 1024, 512, 512], top to bottom / left to right) were trained and validated. The correlation of the predicted BAs is stated as Pearson's correlation coefficient.

Dataset	Num. Ref. ratings	n	Model accuracy (months, ↓)		Inter-rater accuracy (months, ↓)	
			MAD	RMSE	MAD	RMSE
RSNA	6	200	3.9	5.1 ([4.7, 5.7])	4.8-7.0*	-
DHA	2	1,383	5.8	7.7 ([7.4, 8.0])	4.4	7.0 ([6.8, 7.3])
GDBD	2	702	6.0	7.7 ([7.3, 8.1])	9.5	12.8 ([12.2, 13.6])
GDBD (gen. dis.)	2	568	5.8	7.5 ([7.1, 7.9])	9.5	12.8 ([12.1, 13.6])

Table 1: Accuracy of Deeplasia and inter-rater accuracy across different test datasets. For the GDBD additionally, the scores for images from patients with molecularly confirmed genetic disorders (gen. dis.) are provided. The RMSE is stated with the 95% confidence interval. n refers to the number of individual X-rays in each set. * Estimated range for the accuracies of the assessed single raters.

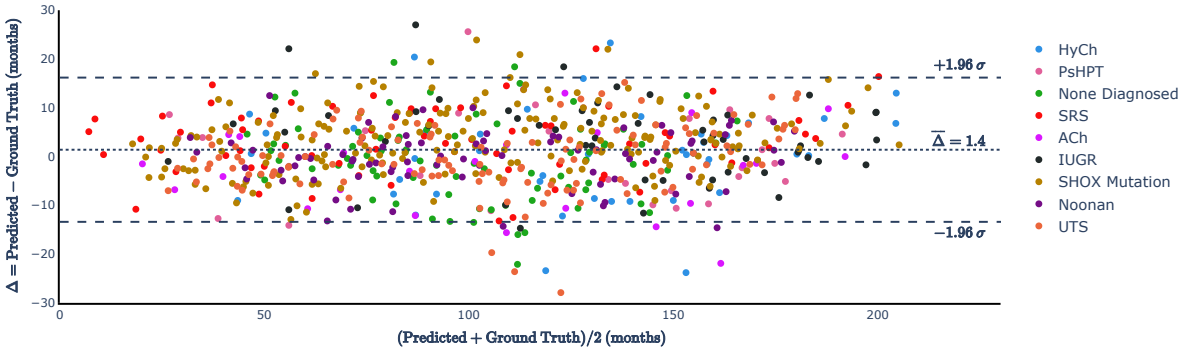

Figure 5: *Bland-Altman* plot comparing the predictions of Deeplasia and the ground truth values (which are the average values of the two manual reference gradings on the GDBD). The dotted line denotes the mean deviation and the dashed lines the 95% confidence interval. Individual disorder categories are color-coded as indicated. An interactive version of the plot can be accessed via `https://aimi-bonn.github.io/website/deeplasia/results.html`.

4.3 Performance on the DHA dataset

To assess the generalizability to external test cohorts and potentially unseen ethnicities, we evaluated Deeplasia on the DHA dataset [27, 28]. We used 1,383 X-ray images from children (age 0-18 years) with different ethnic backgrounds and their corresponding BA ratings. On this dataset, Deeplasia achieves a MAD of 5.81 months, RMSE of 7.67 months, and a 1-year accuracy of 92.9% (see the second row of Table 1). Note that for this dataset the ground truth BA estimates are based on two rather than six raters for the RSNA test set.

4.4 Performance on the GDBD

Finally, we evaluated the performance on the GDBD to assess the generalization of Deeplasia to patients with skeletal malformations. Overall, the GDBD contains 568 images from patients with a molecularly confirmed genetic disorder, 55 images from patients with IUGR, and 79 images from individuals without any known disorders, but who had been referred to pediatric endocrinologists due to a suspected growth disorder. All reference BA ratings were performed by the same two raters (KM and AK).

Comparing the accuracy of the model and the ground truth estimate defined by the average of two raters, the accuracy is 5.96 months (MAD), 7.67 months (RMSE), and 90.2% (1-year accuracy) for the full set and 5.84 months (MAD), 7.48 months (RMSE), and 90.1% (1-year accuracy) for the subset of patients with molecularly confirmed disorders. These values (also listed in the third and fourth rows of Table 1) are similar to those from the performance on the DHA dataset and in the range of the single rater estimated in the annotation of the RSNA BA challenge [13]. Consequently, the error of the model w.r.t. the average of two reference ratings is smaller than the assessed inter-rater error (Table 1). In Fig. 5, we illustrate the *Bland-Altman* plot, which shows the difference between the BA predictions from Deeplasia and the ground truth values (from the two raters) vs. the average of the two methods. The signed mean difference of the two methods is $\overline{\Delta} = +1.4$ months (shown by a dotted line), and the plot reveals no systematic over- or underestimation of the BAs for different skeletal disorders. The difference between the predicted BA and the ground truth ratings is within 1.96 standard deviations (i.e. the 95% confidence interval) for 95.6% of the predicted BAs.

Analyzing the models' predictive error for individual disorders, listed in Table 2, shows no significant drop of performance in comparison to the children with no diagnosed disorder. However, a tendency of increased RMSE and MAD can be observed for ACh, HyCh, and PsHPT, while a significantly decreased error can be observed for Noonan and UTS. This can likely be attributed to the accuracy of the reference grading, given that the inter-rater errors (columns 7 and 8 of Table 2) are also higher for ACh, HyCh, and, PsHPT and lower for Noonan syndrome and UTS. Of note, for each disorder, the average error of the model in comparison to a single manual rating (column 5 of Table 2) is smaller than the average difference between the two manual raters. Hence, our model ensemble is at least as accurate as the assessed human raters for all assessed disorders and, at the same time, retains the accuracy on disorders causing severe malformations (ACh, HyCh, and, PsHPT), while those disorders increase the inter-rater disagreement. This demonstrates the generalizability of Deeplasia to

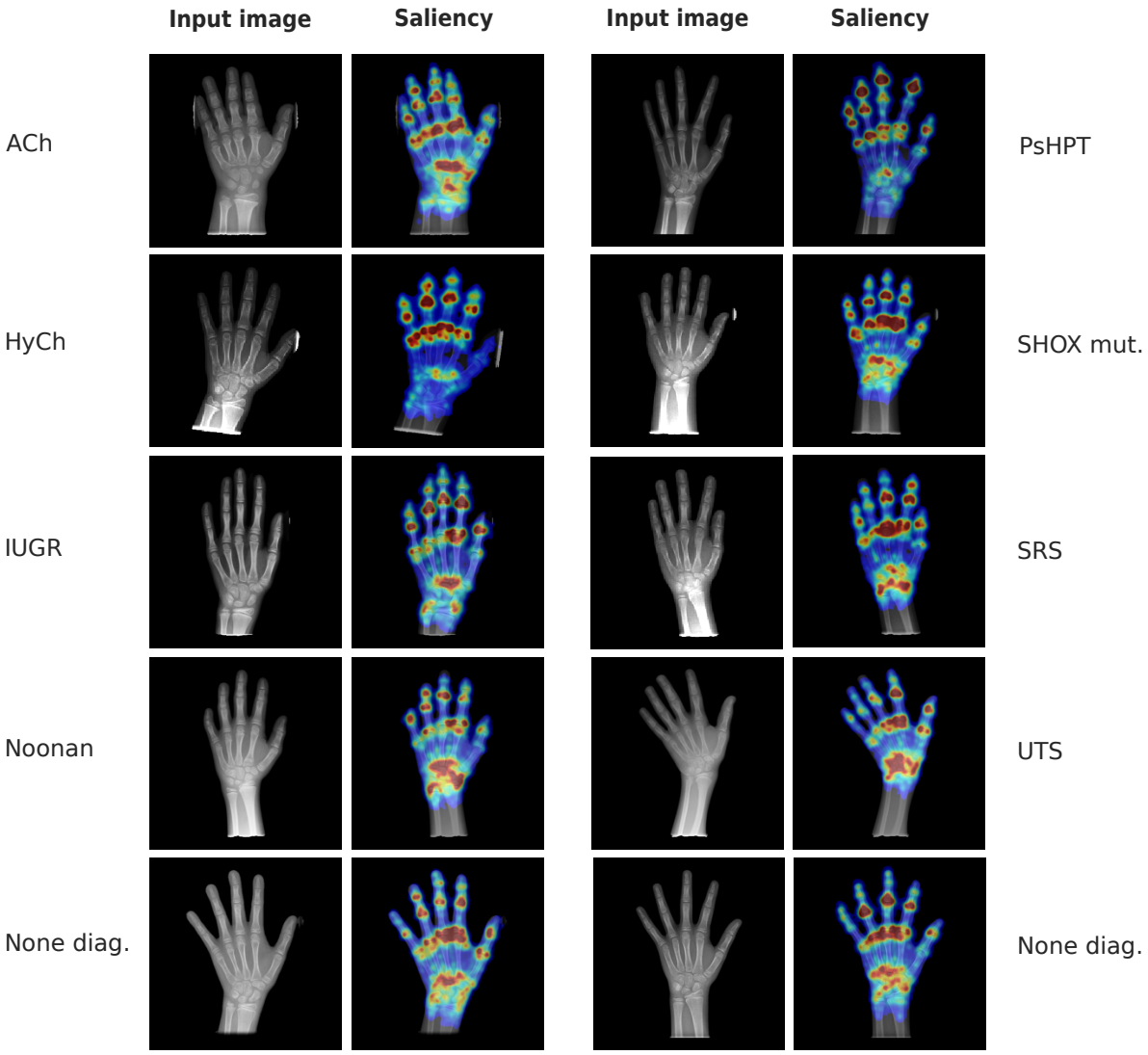

Figure 6: Attribution of images from female patients with the indicated disorders (upper four rows) and without any diagnosed disorder (the lowest row). All patients shown here have a ground truth bone age between 90 and 120 months. None diag. stands for no genetic disorder diagnosed, and SHOX mut. stands for mutation of the SHOX gene.

unseen bone morphologies.

The attribution maps, illustrated in Fig. 6 (see Appendix D for details on generating these maps), show that the attentions of the models are mainly on the phalangeal and metacarpal joints, as well as the carpal bones. Thereby, the observable patterns in the saliency maps of the dysplastic hands remain unaltered in comparison to the hands with no diagnosed disorder. Visualizing the learned feature embeddings (via t-SNE, see Appendix E and Fig. 8) shows that the extracted features of the dysplastic hands follow the patterns of similar-aged hands without any diagnosed disorder, while no clustering based on the disorder label can be observed. Together, this shows that the activation patterns within the model are invariant to the dysmorphologies represented in the GDBD and the extracted features remain unaffected by the anomalies. Combined with the results of the unaltered performance, this underlines the invariance of the models to the presence of skeletal disorders in the input images.

4.5 Longitudinal analysis

In clinical scenarios, determining the BA is not only important for receiving an initial diagnosis but further for monitoring the development and maturation. This requires a high test-retest reliability for

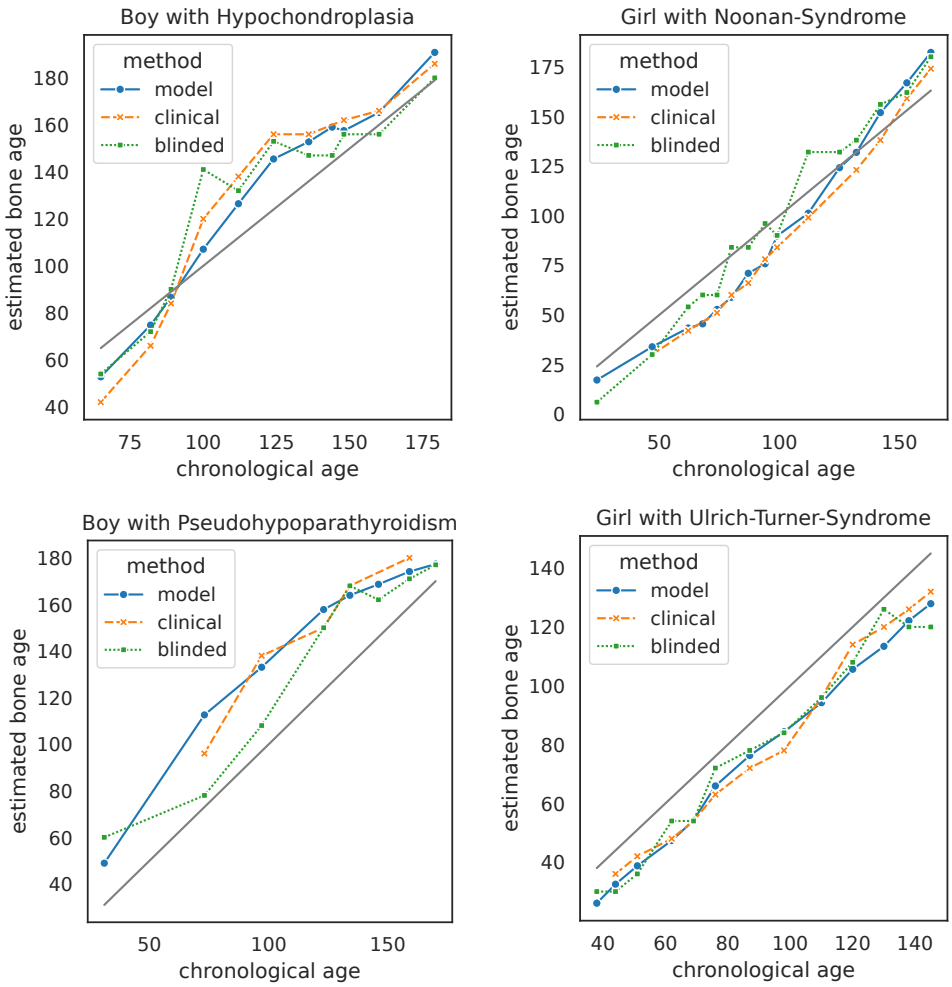

Figure 7: Exemplary plots of bone age maturation progress of individual patients within the GDBD estimated by Deeplasia, the clinical, and a blinded manual assessment. Bone age and chronological age are denoted in months.

the measured BA. We retrospectively estimate the test-retest reliability from regular check-ups within our cohort, employing the method proposed by [40]. In brief, this method assumes a linear progress of BA between two measurements and compares the measured BA to the interpolation between adjacent BA estimates.

The results from this analysis are summarized in Table 3. Based on the GDBD we estimate the test-retest precision on patients with genetic disorders to be at most 2.74 months (95% confidence interval [2.46, 3.09], $n = 149$). Comparing our results to the ground truth rating shows that the precision of Deeplasia is on-a-par with the clinical assessment. Nevertheless, in a clinical scenario, the patient's identity, diagnosis, and BA results from the previous examinations are known and can be used to smooth the next reported BA. If the ratings are conducted blinded and in a randomized order without additional information, the precision of the human BA reading drops significantly (Table 3) and the noise in manual BA assessment is clearly visible (Fig. 7). Thus, automatic BA prediction using Deeplasia is significantly more precise and reliable than a manual rating in a blinded scenario (Table 3).

5 Conclusion

In this work, we presented Deeplasia, a deep learning approach for BA assessment specifically validated on patients with skeletal dysplasias. While designing and optimizing Deeplasia, we experimented with nine different CNN models, and in the end, chose three that maximize the performance (on the validation set) and at the same time minimize the prediction patterns overlap to build a model ensemble.

Disorder	n	Model accuracy w.r.t. two raters (months, ↓)		Model accuracy w.r.t. a single rater (months, ↓)		Inter-rater accuracy (months, ↓)	
		MAD	RMSE	MAD	RMSE	MAD	RMSE
Achondroplasia	25	7.3	9.2 ([7.2, 12.7])	10.0	13.1 ([10.9, 16.2])	13.3	18.5 ([14.5, 25.5])
Hypochondroplasia	44	7.2	9.5 ([7.9, 12.0])	8.4	11.8 ([10.3, 13.9])	11.6	13.9 ([11.5, 17.6])
Noonan syndrome	80	4.3	5.6 ([4.8, 6.6])	7.5	8.0 ([7.2, 9.0])	8.9	11.5 ([9.9, 13.6])
Pseudohypoparathyroidism	30	7.5	8.8 ([7.1, 11.8])	10.3	13.9 ([11.8, 17.0])	13.9	21.5 ([17.2, 28.7])
SHOX mutation	198	5.9	7.5 ([6.8, 8.3])	7.5	9.6 ([9.0, 10.3])	9.5	12.0 ([11.0, 13.4])
Silver-Russell syndrome	69	6.2	7.7 ([6.6, 9.2])	7.7	9.6 ([8.6, 10.8])	7.7	11.5 ([9.8, 13.8])
Ullrich-Turner syndrome	122	5.2	6.9 ([6.1, 7.9])	6.4	8.7 ([8.0, 9.6])	8.4	11.5 ([9.6, 12.3])
Intrauterine Growth Restriction	55	7.2	8.9 ([7.5, 11.0])	8.8	10.8 ([9.6, 12.5])	8.5	12.2 ([10.3, 15.0])
Non diagnosed	79	6.3	8.1 ([7.0, 9.6])	9.3	10.5 ([9.5, 11.8])	10.0	13.5 ([11.6, 15.9])

Table 2: The model accuracy on the GDBD w.r.t. to the average bone age rating of two raters, a single rater, and the inter-rater accuracy. n refers to the number of individual X-rays per disorder. The RMSE is stated with the 95% confidence interval.

Rating method	Precision (months, ↓)	
	full dataset ($n = 149$)	with clinical BA ($n = 106$)
Model ensemble	2.7 ([2.5, 3.1])	2.4 ([2.1, 2.8])
Clinical rating	-	2.6 ([2.3, 3.0])
Blinded rating	**5.6 ([4.9, 6.2])**	**5.8 ([5.1, 6.7])**

Table 3: The test-retest precision of the model ensemble, the clinical, and a blinded manual bone age rating is estimated on patients with genetically-confirmed disorders (i.e. excluding IUGR) within the GDBD. n refers to the number of images within the GDBD for which the interpolation residuals could be estimated. The precision is stated with the 95% confidence interval. Statistically significantly decreased precision w.r.t. to the model ensemble are marked in bold.

Deeplasia achieves a SOTA MAD of 3.87 months on the RSNA test set, demonstrating that our prior-free learning approach is as powerful as other approaches which require extensive annotations, ROI extractions, or human priors. Additionally, we tested the generalizability of Deeplasia on the external DHA dataset that contains hand X-rays of healthy patients from different ethnicities and achieved a MAD of 5.81 months. We then applied Deeplasia on the GDBD - a new dataset comprising hand X-rays with skeletal malformations - achieving a MAD of 5.96 months. These are similar to our results for DHA set which also has two reference ratings, compared to RSNA test set which is based on six reference ratings. The lowest MAD values were for Noonan and UTS (4.3 and 5.2 months, respectively), and the highest MAD values were for PsHPT, ACh, and HyCh (7.5, 7.3, and 7.2 months, respectively). However, the accuracy of human performance on the GDBD drops much more dramatically, so (1) the decrease in the accuracy of the model ensemble can at least be partially attributed to an inaccurate ground truth rating, and (2) the model ensemble is much more accurate than human raters. This underlines the need for automated methods, especially for BA assessment of patients with such disorders. In fact, PsHPT, ACh, and HyCh are amongst the disorders that cannot be processed by the current automated BA tools based on bone segmentation (e.g. BoneXpert, [18]). While segmentation-based approaches offer a higher degree of explainability compared to end-to-end learning methods, the latter is successful in analyzing dysmorphic bones for which the former method does not work.

Finally, we applied Deeplasia to longitudinal data of patients with bone disorders. Although it only considers the X-ray images, we showed that Deeplasia is as accurate as clinical assessments (with prior knowledge of clinical history) for monitoring the growth progress.

As patients with rare skeletal malformations are among the most important groups that require BA assessment, it is vital to insure the applicability and generalizability of automated approaches to these patients in dedicated studies such as this work. While there have been some studies employing DL-based techniques on medical images of patients with rare genetic diseases (e.g. [41, 42, 43]), this field is still understudied perhaps mainly due to the inherently small amount of available data from such

diseases. The current study is limited to only seven different genetic bone diseases. Hence, our future works should expand the current dataset to a broader set of disorders and to patients with varying ethnic backgrounds[2]. We provide the codes we developed for our model ensemble to the community for scrutiny and reuse in their research.

Acknowledgements

This publication has been supported by the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability (ERN-ITHACA). ERN-ITHACA is funded by the EU4Health Program of the European Union, under the Grant Agreement Nr. 101085231. The authors thank Dr. Sven Koitka for the assistance with retrieving the DHA dataset and ground truth annotations, and Dr. Mark Born, Dr. Jörg Schaper, Dr. Alexej Knaus, Prof. Tinatin Tkemaladze, and Prof. Alain Verloes for fruitful discussions.

Author contributions

Sebastian Rassmann: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization. **Alexandra Keller**: Data Curation, Formal analysis, Writing - Review & Editing. **Kyra Skaf**: Data Curation. **Alexander Hustinx**: Software, Writing - Review & Editing. **Ruth Gausche**: Data Curation, Resources. **Miguel A. Ibarra-Arrelano**: Writing - Review & Editing. **Tzung-Chien Hsieh**: Writing - Review & Editing. **Yolande E. D. Madajieu**: Data Curation. **Markus M. Nöthen**: Supervision. **Roland Pfäffle**: Resources. **Klaus Mohnike**: Investigation, Formal analysis, Resources, Writing - Review & Editing, Funding acquisition, Supervision. **Peter M. Krawitz**: Writing - Review & Editing, Resources, Funding acquisition, Supervision, Project administration. **Behnam Javanmardi**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, Supervision, Project administration.

References

[1] Ana L Creo and W Frederick Schwenk 2nd. "Bone Age: A Handy Tool for Pediatric Providers". en. In: *Pediatrics* 140.6 (Dec. 2017).

[2] Paul M Bunch et al. "Skeletal development of the hand and wrist: digital bone age companion-a suitable alternative to the Greulich and Pyle atlas for bone age assessment?" en. In: *Skeletal Radiol.* 46.6 (June 2017), pp. 785–793.

[3] Giorgio Aicardi et al. "Assessment of skeletal maturity of the hand-wrist and knee: A comparison among methods". en. In: *Am. J. Hum. Biol.* 12.5 (Sept. 2000), pp. 610–615.

[4] David D Martin et al. "The use of bone age in clinical practice–part 1". In: *Hormone research in paediatrics* 76.1 (2011), pp. 1–9.

[5] David D Martin et al. "The use of bone age in clinical practice–part 2". In: *Hormone research in paediatrics* 76.1 (2011), pp. 10–16.

[6] W W Greulich and S I Pyle. *Radiographic Atlas of Skeletal Development of the Hand and Wrist.* Stanford University Press, 1959.

[7] J M Tanner. *Growth at adolescence; with a general consideration of the effects of hereditary and environmental factors upon growth and maturation from birth to maturity.* Oxford: Blackwell Scientific Publications, 1962.

[8] J M Tanner et al. *Assessment of skeletal maturity and prediction of adult height (TW2 method).* en. 1975.

[9] J M Tanner et al. *Assessment of skeletal maturity and prediction of adult height (TW3 method).* en. 3rd ed. London: W.B. Saunders, 2001.

[10] Vincenzo De Sanctis et al. "Hand X-ray in pediatric endocrinology: Skeletal age assessment and beyond". en. In: *Indian J. Endocrinol. Metab.* 18.Suppl 1 (Nov. 2014), S63–71.

[2] E.g. via support from FAIR [44] sources such as the GestaltMatcher Database (`db.gestaltmatcher.org`).

[11] David K Eng et al. "Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial". en. In: *Radiology* 301.3 (Dec. 2021), pp. 692–699.

[12] David B Larson et al. "Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs". en. In: *Radiology* 287.1 (Apr. 2018), pp. 313–322.

[13] Safwan S Halabi et al. "The RSNA Pediatric Bone Age Machine Learning Challenge". en. In: *Radiology* 290.2 (Feb. 2019), pp. 498–503.

[14] Mark Cicero and A Bilbily. "Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge". In: *Accessed from https://www.16bit.ai/blog/ml-and-future-of-radiology* (2017).

[15] Marıa Escobar et al. "Hand Pose Estimation for Pediatric Bone Age Assessment". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pp. 531–539.

[16] Dong Wang et al. "Improve bone age assessment by learning from anatomical local regions". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 631–640.

[17] Sven Koitka et al. "Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks". In: *Medical Image Analysis* 64 (Aug. 2020), p. 101743. DOI: 10.1016/j.media.2020.101743. URL: https://doi.org/10.1016/j.media.2020.101743.

[18] D D Martin et al. "Accuracy and self-validation of automated bone age determination". en. In: *Sci. Rep.* 12.1 (Apr. 2022), p. 6388.

[19] J.W. Spranger, A. Superti-Furga, and S. Unger. *Bone Dysplasias: An Atlas of Genetic Disorders of Skeletal Development*. Oxford University Press, 2018. ISBN: 9780190626655. URL: https://books.google.de/books?id=jLhwDwAAQBAJ.

[20] Geert R. Mortier et al. "Nosology and classification of genetic skeletal disorders: 2019 revision". In: *American Journal of Medical Genetics Part A* 179.12 (2019), pp. 2393–2419. DOI: https://doi.org/10.1002/ajmg.a.61366.

[21] Ataf Hussain Sabir and Trevor Cole. "The evolving therapeutic landscape of genetic skeletal disorders". In: *Orphanet Journal of Rare Diseases* 14.1 (Dec. 2019). DOI: 10.1186/s13023-019-1222-2. URL: https://doi.org/10.1186/s13023-019-1222-2.

[22] Mari Satoh and Yukihiro Hasegawa. "Factors affecting prepubertal and pubertal bone age progression". en. In: *Front. Endocrinol.* 13 (Aug. 2022), p. 967711.

[23] Jeong Rye Kim et al. "Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency". en. In: *AJR Am. J. Roentgenol.* 209.6 (Dec. 2017), pp. 1374–1380.

[24] Fengdan Wang et al. "Artificial intelligence system can achieve comparable results to experts for bone age assessment of Chinese children with abnormal growth and development". In: *PeerJ* 8 (2020), e8854.

[25] Hans Henrik Thodberg et al. "The BoneXpert method for automated determination of skeletal maturity". en. In: *IEEE Trans. Med. Imaging* 28.1 (Jan. 2009), pp. 52–66.

[26] Amaka C Offiah and Christine M Hall. "The radiologic diagnosis of skeletal dysplasias: past, present and future". en. In: *Pediatr. Radiol.* 50.12 (Nov. 2020), pp. 1650–1657.

[27] Arkadiusz Gertych et al. "Bone age assessment of children using a digital hand atlas". In: *Computerized medical imaging and graphics* 31.4-5 (2007), pp. 322–331.

[28] Aifeng Zhang et al. "Racial differences in growth patterns of children assessed on the basis of bone age". In: *Radiology* 250.1 (2009), pp. 228–235.

[29] John R Zech et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study". In: *PLoS medicine* 15.11 (2018), e1002683.

[30] Xinlei Chen et al. "Tensormask: A foundation for dense object segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2061–2069.

[31] Bhakti Baheti et al. "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2020, pp. 1473–1481.

[32] Sebastian Rassmann et al. *Hand mask for the RSNA bone age dataset*. Zenodo, Feb. 2023. DOI: 10.5281/zenodo.7611677. URL: https://doi.org/10.5281/zenodo.7611677.

[33] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[34] Felipe Torres et al. "An empirical study on global bone age assessment". In: *15th International Symposium on Medical Information Processing and Analysis*. Vol. 11330. SPIE. 2020, pp. 98–105.

[35] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] William Falcon et al. "Pytorch lightning". In: *GitHub. Note: https://github. com/PyTorchLightning/pytorch-lightning* 3.6 (2019).

[37] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[38] Alexander Buslaev et al. "Albumentations: fast and flexible image augmentations". In: *Information* 11.2 (2020), p. 125.

[39] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[40] Hans Henrik Thodberg and Lars Sävendahl. "Validation and reference values of automated bone age determination for four ethnicities". In: *Academic radiology* 17.11 (2010), pp. 1425–1432.

[41] Yaron Gurovich et al. "Identifying facial phenotypes of genetic disorders using deep learning". In: *Nature Medicine* 25.1 (Jan. 2019), pp. 60–64. DOI: 10.1038/s41591-018-0279-0. URL: https://doi.org/10.1038/s41591-018-0279-0.

[42] Tzung-Chien Hsieh et al. "GestaltMatcher facilitates rare disease matching using facial phenotype descriptors". In: *Nature Genetics* 54.3 (Feb. 2022), pp. 349–357. DOI: 10.1038/s41588-021-01010-x. URL: https://doi.org/10.1038/s41588-021-01010-x.

[43] Nikolas Pontikos et al. "Eye2Gene: prediction of causal inherited retinal disease gene from multimodal imaging using deep-learning". In: (Oct. 2022). DOI: 10.21203/rs.3.rs-2110140/v1. URL: https://doi.org/10.21203/rs.3.rs-2110140/v1.

[44] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016). DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[45] Leonie Henschel et al. "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219 (2020), p. 117012.

[46] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[47] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[48] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[49] Stephen M Pizer et al. "Adaptive histogram equalization and its variations". In: *Computer vision, graphics, and image processing* 39.3 (1987), pp. 355–368.

[50] Alexander Hustinx et al. "Improving Deep Facial Phenotyping for Ultra-rare Disorder Verification Using Model Ensembles". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 5007–5017. DOI: 10.1109/WACV56688.2023.00499.

[51] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[52] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[53] David D Martin et al. "Clinical application of automated Greulich-Pyle bone age determination in children with short stature". In: *Pediatric radiology* 39.6 (2009), pp. 598–607.

[54] Sehyo Yune et al. "Beyond human perception: sexual dimorphism in hand and wrist radiographs is discernible by a deep learning model". In: *Journal of Digital Imaging* 32.4 (2019), pp. 665–671.

Appendices

A Hand segmentation

To extract the hand and mask the background, masks representing the hand were annotated on 528 images randomly drawn from the RSNA training set. The mask annotation used a semi-automatic procedure based on applying intensity thresholds and edge detection. The segmentation was manually controlled and, if needed, corrected. Leveraging this dataset, we trained *TensorMask* [30] and *Efficient-UNet* models [31] for automated mask prediction using 460 masks and leaving the remaining masks for validation. To avoid fitting the BA model to the masks predicted by only one of the masking models and potentially wrongly predicted masks decreasing the effective size of the training set, in each training epoch the masks were randomly selected between either model. To allow for fast mask prediction without hardware acceleration, a light-weight *FastSurferCNN* [45] model was trained on the masks predicted by the *TensorMask* model of the complete RSNA BA training set excluding images with manually edited masks. To reduce the model size, the number of filters was reduced to 32. The models were trained using the ADAM optimizer [46] with a base learning rate (LR) of 10^{-2} scheduled with the *CosineAnnealingLR* ($T_0 = 10$, minimum LR of $5 * 10^{-5}$) for 50 epochs and a batch size of 32. Weight decay was set to 10^{-6}. To reduce biases towards detecting all high-intensity pixels as false positive, the weight of the logistic loss term of the composite loss [45] was adapted to also include all pixels outside of the annotated mask with an intensity higher than the 80th percentile of the annotated mask. Further, we simulated scanning artifacts by drawing artificial boxes and gradient stripes. With this configuration, we achieved a Dice similarity score of 0.993 with respect to the unseen manually edited set of masks. To test the performance of the full BA prediction pipeline, the model evaluation was carried out using *FastSurferCNN* masking.

B Details of bone age model training

All BA models were trained using a mean squared error (MSE) loss and the ADAM optimizer [46]. The initial LR was set to 10^{-3} and decayed using *ReduceLRonPlateau* (factor of 0.2, patience of 10 epochs) to a minimum LR of 10^{-4} tracking the MAD in the validation set. The weights of the final models from each training process were chosen based on the best validation MAD ("checkpointing"). For regularization, dropout ($p = 0.2$) was added to the FC layers and a weight decay of $5 * 10^{-6}$ was applied. We resized the images to a minimum resolution of 512×512 to assert the resoluting potentially relevant fine-grained structures in the bones such as growth gaps. Due to the miss-match of the training resolution with the pre-training of the smaller *EfficientNet-b0* version, the *-b0* models were trained from scratch using the Kaiming initialization [47]. However, the larger *-b4* versions are pre-trained at a similar resolution, so the *-b4* models used *ImageNet* [48] pre-training. The *-b0* versions were trained for 300 epochs, whereas the pre-training of the *-b4* models allowed for faster convergence, so the training was reduced to 100 epochs. The mini-batch size was set to 32 for all models.

As *default* data augmentations we used the approach described by [14] (relative scaling and translation of $\pm 0 - 20\%$, rotation of $\pm 0 - 20°$, shear of $\pm 0 - 1°$ horizontal flipping with $p = 0.5$). We extended this to our *strong* set of augmentations by increasing the maximum scaling and translation to $\pm 30\%$, rotation to $\pm 30°$, and shear to $\pm 10°$. Additionally, non-linear intensity manipulations with either ($p = 0.33$) a Gamma-correction (gamma chosen from [0.7, 1.3]) or ($p = 0.67$) a contrast limited adaptive histogram equalization (CLAHE, [49], clip limit: 3), and image sharpening ($p = 0.2$, alpha choosen from [0.5, 0.75] and lightness chosen from [0.5, 1]) were applied. To compensate for strong regularization inducing a bias towards predicting more extreme BAs on non-augmented samples, the inferred predictions were corrected via a linear regression model fitted on the predictions of the non-augmented training set.

Test time augmentation (TTA) was performed by rotating the input image by $-10, -5, 0, 5, 10°$ and each with and without applying additional horizontal mirroring. Both, model ensembling and TTA, use an unweighted average of all predictions for any given image.

The models included in our final ensemble were chosen based on the best validation MAD score in each training condition.

C Model experimentations

DL model ensembles often show higher performances compared to single models (see e.g. [50, 43]), however, usually multiple experimentations are required to reach a suitable set of models. We took the following steps for investigating the optimum model configurations:

In the first experiment the effect of applying a stronger data augmentation than previously proposed [14] was studied. To this end, we compared the performance of the smallest model configurations (*EfficientNet-b0* with 512×512 input resolution) trained with *default* and *strong* augmentations. These include additional non-linear intensity transformations and edge sharpening (see section 3.2).
Assessing the performance of these models on the internal validation set of the RSNA dataset shows that the *strong* augmentations improve the prediction accuracy across all assessed model configurations (see Table 4 for details). Therefore, we assumed that the *strong* augmentations would improve generalization to unseen data and used these augmentations for all the subsequent experiments as our *baseline* training condition.

Next, we studied the effect of scaling the model size by replacing the *EfficientNet-b0* backbone ($5.3 * 10^6$ parameters) with the larger *-b4* version ($19.3 * 10^6$ parameters) in the *large CNN* condition. Finally, we explored the effect of increasing the input resolution from 512×512 to 1024×1024 as the *high-resolution* condition. Both of the latter modifications show additional improvements compared to the baseline configuration in the RSNA validation set (Table 5).

As a final experiment, we conducted test time augmentation (TTA) similar to the method proposed by [14] for the three chosen models. This improved the ensemble validation accuracy MAD from 6.12 to 6.08 months, see Table 6. Given that the TTA yielded only a marginal improvement but has high computational costs, we decided not to include it in our BA inference approach.

The Tables 4, 5, and 6 list the detailed results of our experimentation for building the model ensemble described in this Appendix and in Section 4.1.

EfficientNet version	Input resolution	FC layers	Val. MAD (months, ↓)	
			default augm.	*strong* augm.
b0	512×512	[256]	6.6	**6.4**
b0	512×512	$[512, 512]$	6.8	**6.7**
b0	512×512	$[1024, 1024, 512, 512]$	6.5	**6.4**

Table 4: Accuracy of single models trained with the *default* and *strong* set of augmentations at different configurations of fully-connected (FC) layers.

Condition name	*EfficientNet* version	Input resolution	FC layers	Val. MAD (months, ↓)
baseline	*b0*	512×512	[256]	**6.4**
			$[512, 512]$	6.7
			$[1024, 1024, 512, 512]$	6.4
large CNN	*b4*	512×512	[256]	6.3
			$[512, 512]$	**6.4**
			$[1024, 1024, 512, 512]$	6.4
high-resolution	*b0*	1024×1024	[256]	**6.3**
			$[512, 512]$	6.4
			$[1024, 1024, 512, 512]$	6.4

Table 5: Comparison of the validation MAD of different training conditions and model configurations. The final model ensemble was selected based on the best score (bold) in each training condition.

Condition name	*EfficientNet* version	Input resolution	FC layers	Val. MAD (months, ↓)	
				no TTA	TTA
baseline	*b0*	512×512	[256]	6.4	6.4
high-resolution	*b0*	1024×1024	[256]	6.3	6.2
large CNN	*b4*	512×512	$[512, 512]$	6.2	6.1
ensemble				6.1	6.1

Table 6: Comparison of the best performing single models in each condition on the RSNA BA validation set with and without applying test time augmentation (TTA). Additionally, an ensemble composed of all models is tested.

D Saliency maps

The attribution maps M were generated by calculating the absolute value of the gradient of the predicted BA $\hat{Y}$ w.r.t. to the input image I:

$$M(I) = \left| \frac{\partial \hat{Y}}{\partial I'}|_{I'=I} \right| = \left| \frac{\partial f^{(w)}(I, S)}{\partial I'}|_{I'=I} \right| \tag{4}$$

The resulting image was subsequently smoothed using a Gaussian kernel with a size 5% of the input image resolution. Subsequently, the maps were normalized by subtracting the minimum intensity, dividing by value of the resulting 99th percentile, and clipping all resulting values to a maximum of 1. For better visualization of the results in the scenario of a masked input images, values less than 0.075 were excluded before applying the color map. Finally, the color maps were blended on the original input images.

E Visualization of learned feature embedding

The internal feature representations were jointly analyzed for the union of the GDBD with the RSNA validation and test sets. To this end, the activation of the first fully-connected layer (i.e. the extracted embedding of the image and the inflated representation of the sex) was obtained and visualized (Fig. 8) using the *scikit-learn* (v.1.0.2, [51]) implementation of the *t-distributed stochastic neighbor embedding* (*t-SNE*, [52]). Keeping the remaining hyper-parameter at the default setting, the perplexity values of $\{20, 50, 100, 200\}$ were tested for each model of the ensemble without a noticeable difference in the clustering patterns.

F Predicting sex from hand X-ray and its effect on BA estimation

Biologically, bone development is highly sex-specific as girls develop and mature earlier and faster than their male peers. Consequently, the same scan read as male rather than female underestimates the BA and vice-versa [9, 6]. Hence, conducting the BA assessment with the wrong sex can cause wrong results both in manual and automatic assessment [53]. While such user errors are usually ignored in the model evaluation, assigning the wrong sex in a clinical setting will result in a highly inaccurate BA estimation. [54] demonstrated that the sex of a patient can be rather precisely predicted from a hand X-ray. Replicating their results, we integrate sex prediction into our BA estimation pipeline rendering the prediction more robust to user errors.

The sex prediction task was formulated as logistic regression. To this end, the baseline model (*EfficientNet-b0* backbone, a single FC layer of 256 neurons) was extended with an additional output neuron for the sex, while the sex was removed as an input. The model was then trained using an additional binary cross-entropy loss on the sex prediction task and the MAD was replaced by the area under the receiver operating characteristic curve (AUROC) as a validation metric.

In line with previous findings, on the RSNA test set our sex prediction model achieves an accuracy of 93.0%, 89.3%, and 81.8% for the RSNA test set, the DHA, and the GDBD, respectively. Using the sex predicted by the model as input to our bone age models, the accuracy in each test set drops considerably (Table 7). Hence, completely omitting sex annotated by the user would result in a

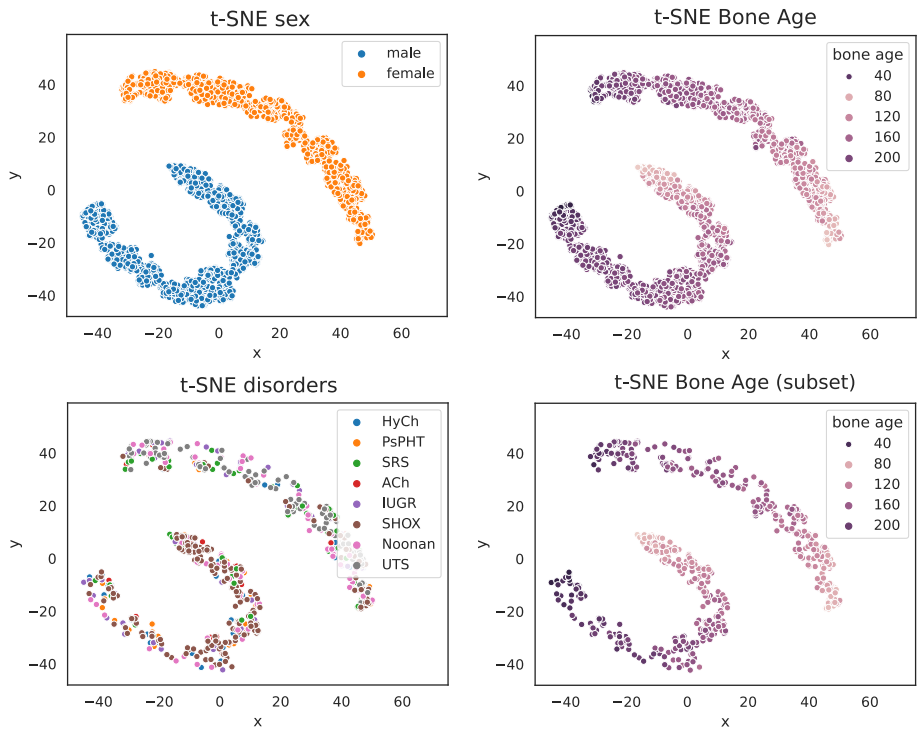

Figure 8: *t-SNE* of the features extracted from the RSNA test and validation sets combined with the GDBD. For drawing the plot in the lower row, healthy cases were omitted. All shown *t-SNE* plots were obtained from the baseline *EfficientNet-b0* model and with the perplexity set to 50. Similar results were obtained for other perplexity values and models.

dramatic loss of accuracy. Therefore, we propose to use the sex prediction as mainly a verification step to mark contradictions between user input and model prediction avoiding the user double-checking and, potentially correcting erroneous inputs.

Dataset	Ground truth sex		Predicted sex	
	MAD (months, ↓)	RMSE (months, ↓)	MAD (months, ↓)	RMSE (months, ↓)
RSNA	3.9	5.1	5.4	8.4
DHA	5.8	7.7	6.7	9.1
GDBD	6.0	7.7	8.4	11.3

Table 7: Performance of the model ensemble on different datasets using either the real biological sex (ground truth) or the sex predicted by the dedicated model.