Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks

Vladimir I. Iglovikov[1], Alexander Rakhlin[2], Alexandr A. Kalinin[3(✉)], and Alexey A. Shvets[4]

[1] ODS.ai, San Francisco, CA 94107, USA
iglovikov@gmail.com
[2] Neuromation OU, 10111 Tallinn, Estonia
rakhlin@neuromation.io
[3] University of Michigan, Ann Arbor, MI 48109, USA
akalinin@umich.edu
[4] Massachusetts Institute of Technology, Cambridge, MA 02142, USA
shvets@mit.edu

Abstract. Skeletal bone age assessment is a common clinical practice to diagnose endocrine and metabolic disorders in child development. In this paper, we describe a deep learning approach to the problem of bone age assessment using data from the 2017 Pediatric Bone Age Challenge organized by the Radiological Society of North America. This dataset consists of 12,600 radiological images. Each radiograph in the dataset is an image of a left hand labeled with bone age and sex of a patient. Our approach introduces a comprehensive preprocessing protocol based on the positive mining technique. We use images of whole hands as well as specific hand parts for both training and prediction. This allows us to measure the importance of specific hand bones for automated bone age analysis. We further evaluate the performance of the suggested methods in the context of skeletal development stages. Our approach outperforms other common methods for bone age assessment.

Keywords: Medical imaging · Computer-aided diagnosis (CAD)
Computer vision · Image recognition · Deep learning

1 Introduction

Clinicians use bone age assessment (BAA) in order to estimate maturity of a child's skeletal system since the difference between assigned bone and chronological ages may indicate a growth problem. BAA methods usually include taking a single X-ray image of the left hand from the wrist to fingertips and comparing it with a standardized reference. Over the past decades, BAA has been performed manually by either comparing the patient's radiograph with an atlas of

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 300–308, 2018.
https://doi.org/10.1007/978-3-030-00889-5_34

representative ages [4] or using a scoring system that examines specific bones [16]. Only recently software solutions, such as BoneXpert [17], have been developed and approved for the clinical use in Europe. BoneXpert uses a computer vision algorithm to reconstruct the contours of 13 bones of a hand. However, it is sensitive to the image quality and does not utilize carpal bones, despite their suggested importance for BAA in infants and toddlers [3]. Methods based on classical computer vision reduce time needed for evaluating a single radiograph, but they still require substantial feature engineering, doctoral supervision and expertise.

Recently, deep learning-based approaches demonstrated performance improvements over conventional machine learning methods for many tasks in biomedicine [1,6]. In medical image analysis, convolutional neural networks (CNN) have been successfully used, for example, for diabetic retinopathy screening [9], breast cancer detection [10], and other problems [1]. Deep neural network based solutions for BAA were suggested before [7,8,14]. However, most of these studies did not evaluate model performance using different hand bones or different skeletal development stages. Moreover, the performance of deep learning models depends on the quality of training data. Radiographs are obtained from various medical centers, different hardware, and under variable conditions. They also vary in scale, orientation, exposure, and often feature specific markings (Fig. 4).

In this study, we present a deep learning-based method for BAA. One of the key contributions of this work is rigorous preprocessing pipeline. To prevent the model from learning false associations from artifacts in the image, we first remove background by segmenting the hand. Then, we normalize contrast and detect key points. Then, we apply affine transforms to register segmented images in a common coordinate space. Besides improving the quality of data, this step allows us to accurately identify different regions of the hand. We train several deep networks using different parts of hand images to assess how different hand bones contribute to the models' performance across four major skeletal development stages. Finally, we compare regression and classification, sex-specific and sex-agnostic models, and evaluate overall performance of our approach. We validate our method using data from the 2017 Pediatric Bone Age Challenge organized by the Radiological Society of North America (RSNA) [12]. The suggested method is robust and shows superior performance compared to other proposed solutions.

2 Methods

2.1 Preprocessing

First, we extract a hand mask from every image to remove all extraneous objects. Simple background removal methods did not produce satisfactory results, while machine learning-based segmentation typically requires large manually labeled training set. To alleviate labeling costs, we use positive mining, an iterative procedure that combines manual labeling with automatic processing, see Fig. 1. It

allows us to quickly obtain accurate masks for the whole training set. For segmentation, we employ slightly modified version of the original U-Net architecture [11] that previously proved itself useful for segmentation problems with limited amounts of data [5], making it a good choice for positive mining.

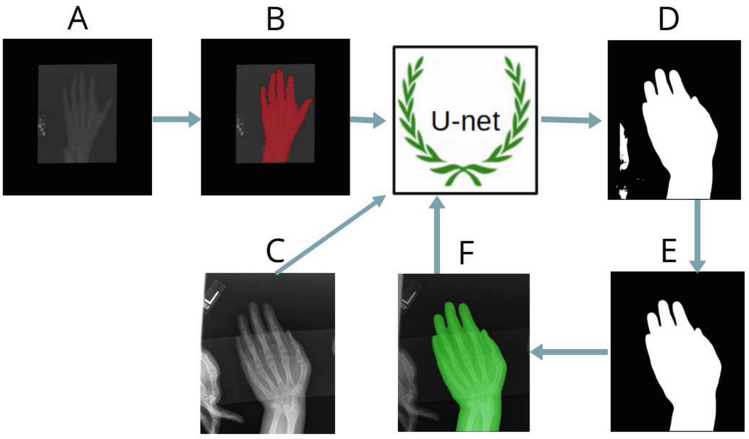


Fig. 1. Iterative procedure of positive mining utilizing U-Net architecture for image segmentation: (A) raw input data; (B) mask manually labeled with the online annotation tool Supervisely [15]; (C) new data; (D) raw prediction; (E) post processed prediction; (F) raw image with mask plotted together for visual inspection.

We train U-Net using a generalized segmentation loss function:

$$L = H - \log J\,, \tag{1}$$

where H is a binary cross entropy that defined as

$$H = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i))\,, \tag{2}$$

where y_i and $\hat{y}_i$ are a binary value (label) and a predicted probability for the pixel i, correspondingly. In the second term of Eq. (1), J is a differentiable generalization of the Jaccard Index

$$J = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i\hat{y}_i}{y_i + \hat{y}_i - y_i\hat{y}_i}\right)\,. \tag{3}$$

By minimizing this loss function, we simultaneously maximize probabilities for correct pixels to be predicted and maximize the intersection between masks and corresponding predictions, which improves overall segmentation performance [5].

First, we manually label 100 hand masks using Supervisely [15]. Then, we train the U-Net model and use it to segment the rest of the training set. For each

prediction we only keep the largest connected component. We manually curate all segmented masks to discard those of poor quality and train the model using the expanded training set with good quality masks. We repeat this procedure 6 times to achieve acceptable quality on the whole training set, see Fig. 1. Finally, we manually label approximately 100 images that U-Net fails to segment correctly.

2.2 Key Point Detection Model

Since original atlas-based methods evaluate specific hand bones, we use several hand regions to assess their importance. In order to correctly locate these regions, radiographs need to be registered in a common coordinate space. For registration, we detect coordinates of several key points of a hand and use them to calculate affine transformation parameters (zoom, rotation, translation, and mirror) (Fig. 2). Three specific points on the image are chosen: the tip of the distal phalanx of the third finger, tip of the distal phalanx of the thumb, and the center of the capitate. All images are re-scaled to the same resolution: 2080×1600 and padded with zeros, when necessary. To create training set for key points model, we manually label 800 radiographs. Pixel coordinates of key points serve as training targets for our regression model. Key point detection model is based on a VGG-like architecture [13] with 3 VGG blocks and 3 fully connected layers with dropout Fig. 3. The VGG module consists of 2 convolutional layers with the Exponential Linear Unit (ELU) activation function [2] and max-pooling. The model is trained with Mean Squared Error loss function (MSE). We downscale input images to 130×100 pixels and apply rotation, translation and zoom as augmentations. The model outputs 6 coordinates (2 for every key point) that are used to calculate affine transformations for all radiographs. We register them such that: (1) the tip of the middle finger is aligned horizontally and positioned approximately 100 pixels below the top edge of the image; (2) the capitate is aligned horizontally and positioned approximately 480 pixels above the bottom edge of the image. The key point for the thumb is used to detect mirrored images and adjust them. The results of the segmentation, normalization, and registration are shown in Fig. 4.

2.3 Bone Age Assessment Models

We compare bone age regression and classification using two VGG-style CNNs [13] with 6 convolutional blocks followed by 2 fully connected layers (see Fig. 3). The input size varies depending on the considered region of an image, Fig. 2. Both networks are trained by minimizing Mean Absolute Error (MAE) with augmentations (zoom, rotation shift). The regression network has a single output predicting bone age in month, which is scaled in the range $[-1, 1]$. The classification model (Fig. 3) is similar to the regression one, except for two final layers. First, we assign each bone age a class. As bone ages expressed in months, we assume 240 classes total. The second to the last layer is a softmax layer that outputs vector of probabilities for 240 classes. In the final layer, probabilities are multiplied by a vector of bone ages uniformly distributed over integer values

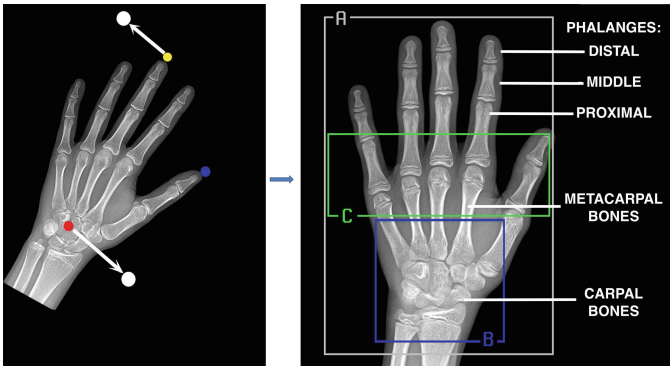


Fig. 2. Image registration. (Left) Key points: the tip of the middle finger (the yellow dot), the center of the capitate (the red dot), the tip of the thumb (the blue dot). Registration positions: for the tip of the middle finger and for the center of the capitate (white dots). (Right) A registered radiograph with three specific regions: (A) a whole hand; (B) carpal bones; (C) metacarpals and proximal phalanges.

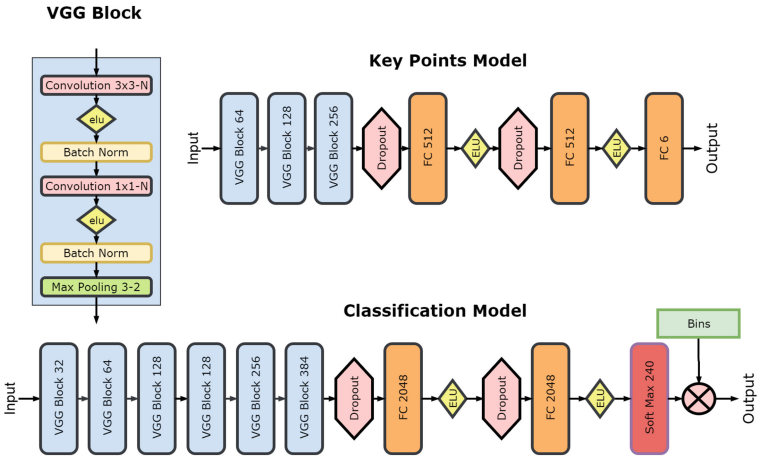


Fig. 3. VGG-style neural network architectures for regression (top) and classification (bottom) tasks.

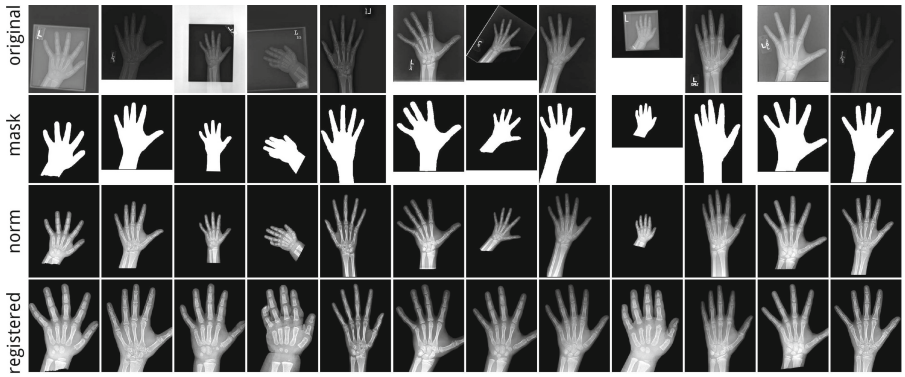


Fig. 4. Preprocessing pipeline: (first row) original images; (second row) binary hand masks that are applied to the original images to remove background; (third row) masked and normalized images; (bottom row) registered images.

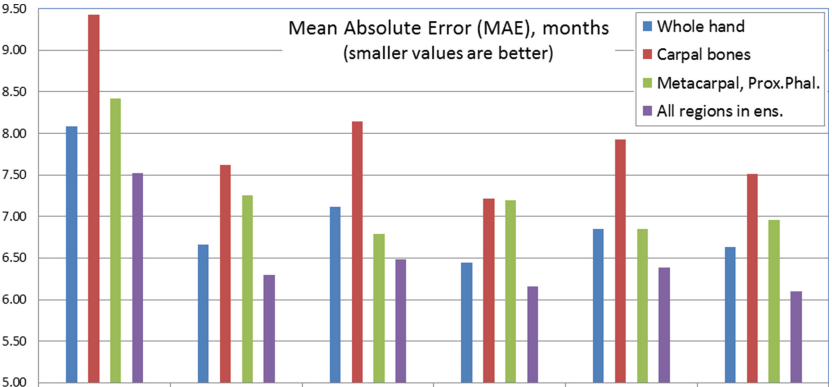
[0..239]. The model outputs single value that corresponds to the expectation of the bone age. Training protocol is the same as for the regression model.

According to the features of skeletal development stages described in [3,4,16], we crop three specific regions from registered radiographs, as shown in Fig. 2: (1) whole hand; (2) carpal bones; and (3) metacarpals and proximal phalanges. We split labeled radiographs into training (11,600 images) and validation (1,000 images) sets, preserving sex ratio. We create several models with a breakdown by: (1) prediction type; (2) sex (males, females, both); and (3) a region (A, B, C). Given these conditions, we produce 18 basic models ($2 \times 3 \times 3$). Furthermore, we construct several meta-models by averaging different regional models.

3 Results

As shown in Fig. 4, original images varied in quality and often had artifacts. In order to assess the effect of preprocessing on prediction performance, we evaluate the regression network on original images, segmented and normalized images, and segmented, normalized and registered images. Corresponding MAEs of 31.56, 8.76, and 8.08 months accordingly demonstrate performance improvement due to the preprocessing. All further results were obtained on the preprocessed data.

The performance of all models evaluated on validation data set is shown in Fig. 5. The region of metacarpals and proximal phalanges (region C in Fig. 2) shows higher accuracy using both regression and classification models. Classification performs better than regression, while the linear ensemble of three regional models outperforms each separate model. The regional pattern MAE(B) > MAE(C) > MAE(A) > MAE (ensemble) is observed for different model types and patient sexes with few exceptions.



	Regression mixed	Regression males	Regression females	Classific. males	Classific. females	Regression Classific. ensemble
Whole hand	8.08	6.67	7.12	6.44	6.85	6.63
Carpal bones	9.43	7.62	8.15	7.22	7.92	7.52
Metacarpal, Prox.Phal.	8.42	7.25	6.79	7.19	6.85	6.96
All regions in ens.	7.52	6.30	6.49	6.16	6.39	6.10

Fig. 5. Mean absolute errors on the validation data set for regression and classification models for different bones and sexes. Colors correspond to different regions. Table: regions are shown in rows, models in columns. There is a total of 15 individual models and 9 ensembles.

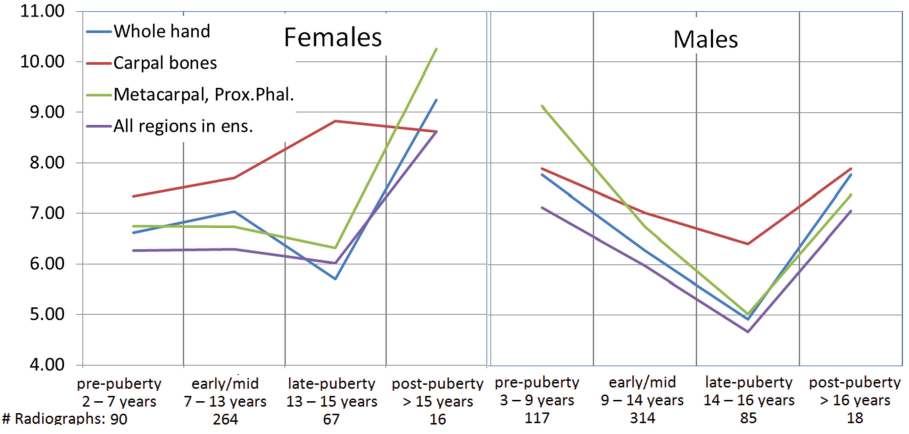


Fig. 6. Mean absolute error in months as a function of skeletal development stages for different sexes. Different colors on the plot correspond to different regions of a radiograph. For males and females the development stages are labelled at the bottom of each plot.

Following [3,8], we also consider four major skeletal development stages: pre-puberty, early-and-mid puberty, late puberty, and post-puberty, see Fig. 6. Infant and toddler categories were excluded due to scarcity of data. Unlike Lee *et al.* [8], we do not observe better results when training on carpal bones compared to other areas. With two exceptions (pre-puberty for males and post-puberty for females), metacarpals and proximal phalanges provide better accuracy than carpals do. Gilsanz and Ratib [3] suggest carpal bones as the best predictor of skeletal maturity only in infants and toddlers. Thus, we find no sound evidence to support the suggestion that carpal bones can be considered the best predictor in pre-puberty. For both sexes the accuracy peaks at late-puberty, the most frequent age in the dataset, showing the influence of the dataset size on the performance.

In the RSNA2017 Pediatric Bone Age Assessement challenge, our solution has been evaluated using the test set consisting of 200 radiographs. Based on organizers' report our method achieves MAE of 4.97 months, higher than local validation, possibly due to the better image or label quality in the test set.

4 Conclusion

In this study, we suggest a deep learning-based approach to the problem of the automatic BAA. Despite the challenging quality of the radiographs, our approach demonstrates robust results and surpasses existing automated models in performance. By using different hand zones, we find that BAA can be done just for carpal bones or for metacarpals and proximal phalanges with around 10–15% increase in error compared to the whole hand. Our approach can be improved by either using more powerful deep networks or increasing the training set size.

References

1. Ching, T., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface **15**(141) (2018)
2. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (ELUs). arXiv preprint arXiv:1511.07289 (2015)
3. Gilsanz, V., Ratib, O.: Hand Bone Age: A Digital Atlas of Skeletal Maturity. Springer Science & Business Media, Heidelberg (2005). https://doi.org/10.1007/b138568
4. Greulich, W.W., Pyle, S.I.: Radiographic atlas of skeletal development of the hand and wrist. Am. J. Med. Sci. **238**(3), 393 (1959)
5. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
6. Kalinin, A.A., et al.: Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics **19**(7), 629–650 (2018)
7. 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, 170236 (2017)

8. Lee, H., et al.: Fully automated deep learning system for bone age assessment. J. Digit. Imaging, 1–15 (2017)
9. Rakhlin, A.: Diabetic retinopathy detection through integration of deep learning classification framework. In: bioRxiv, p. 225508 (2017)
10. Rakhlin, A., Shvets, A., Iglovikov, V., Kalinin, A.A.: Deep convolutional neural networks for breast cancer histology image analysis. In: Campilho, A., Karray, F., ter Haar Romeny, B. (eds.) ICIAR 2018. LNCS, vol. 10882, pp. 737–744. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93000-8_83
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. RSNA Pediatric Bone Age Challenge. http://rsnachallenges.cloudapp.net/competitions/4 (2017). Accessed 16 July 2017
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in X-ray images. Med. Image Anal. **36**, 41–51 (2017)
15. Supervisely. https://supervise.ly/. Accessed 16 July 2017
16. Tanner, J., Whitehouse, R., Cameron, N., Marshall, W., Healy, M., Goldstein, H.: Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method). Academic Press, London (1983)
17. Thodberg, H.H., Kreiborg, S., Juul, A., Pedersen, K.D.: The BoneXpert method for automated determination of skeletal maturity. IEEE Trans. Med. Imaging **28**(1), 52–66 (2009)