Automated Assessment of Bone Age Using Deep Learning and Gaussian Process Regression

Tom Van Steenkiste[1], Joeri Ruyssinck[1], Olivier Janssens[1], Baptist Vandersmissen[1], Florian Vandecasteele[1], Pieter Devolder[2], Eric Achten[2], Sofie Van Hoecke[1], Dirk Deschrijver[1], and Tom Dhaene[1]

***Abstract*— Bone age is an essential measure of skeletal maturity in children with growth disorders. It is typically assessed by a trained physician using radiographs of the hand and a reference model. However, it has been described that the reference models leave room for interpretation leading to a large inter-observer and intra-observer variation. In this work, we explore a novel method for automated bone age assessment to assist physicians with their estimation. It consists of a powerful combination of deep learning and Gaussian process regression. Using this combination, sensitivity of the deep learning model to rotations and flips of the input images can be exploited to increase overall predictive performance compared to only using the deep learning network. We validate our approach retrospectively on a set of 12611 radiographs of patients between 0 and 19 years of age.**

I. Introduction

Bone age assessment is used in medicine to measure skeletal and biological maturity of children [1]. It can be used, among others, to estimate the final adult height [2], to measure therapeutic effect in patients with endocrine disorders [3] or to estimate the age of asylum seekers [4].

In the traditional method, a trained physician compares hand and wrist bones with normal age level images by radiography of the left hand and wrist in combination with reference standards. An example of such a reference standard is the hand atlas of Greulich and Pyle (G&P) [2]. However, the use of such a reference is a lengthy process and leaves room for interpretation, leading to large inter-observer and intra-observer differences. Average spread of inter-observer differences has been reported up to 11.5 months for the G&P method [5]. This causes issues when comparing estimations across patients or of the same patient over time. Furthermore, using different methods leads to variations in the estimated bone age [1].

To reduce these variations, the potential of automated methods to assist the physician has been identified and explored by the community [6], [7], [8]. These methods rely on the segmentation and extraction of typical bone age features from the images. However, including a segmentation step in the processing pipeline can be a significant disadvantage as it is challenging to make these methods robust to large variations in image quality.

[1]Tom Van Steenkiste, Joeri Ruyssinck, Olivier Janssens, Baptist Vandersmissen, Florian Vandecasteele, Sofie Van Hoecke, Dirk Deschrijver and Tom Dhaene are with Ghent University - imec, IDLab, Technologiepark-Zwijnaarde 15, B-9052 Ghent, Belgium tomd.vansteenkiste@ugent.be

[2]Pieter Devolder and Eric Achten are with University Hospital (UZ) Ghent, Department of Radiology, De Pintelaan 185, B-9000 Ghent, Belgium

In other medical domains, deep learning [9] has been proven to be a successful method for image analysis. An example is the automated detection of mitosis in breast cancer histology images [10]. In bone age assessment, recent examples with deep learning include [11] where an automated tool is demonstrated to enhance efficiency of reviewers and [12] where a fully automated setup is discussed for which estimates are accurate within 1 year 92.23% of the time and an average spread of 10.52 months is achieved for patients between 5 and 18 years of age. In this work, we explore a novel machine learning approach for bone age estimation to improve upon the standard state-of-the-art deep learning performance. Our method is based on a powerful combination of deep learning with Gaussian Process Regression (GPR) [13] to exploit sensitivity of the deep learning predictions to rotations and flips of the radiographs.

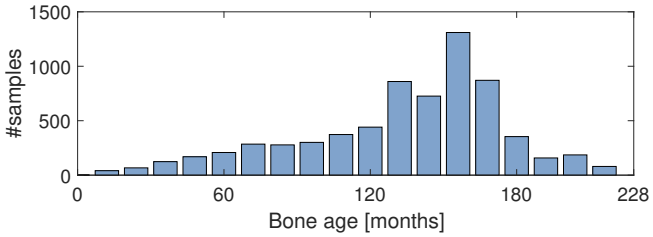
In Section II, the dataset is described. In Section III, the methodology is explained and in Section IV the results of our tests are provided and discussed. Finally, future work is detailed and conclusions are made in Section V.

II. Dataset

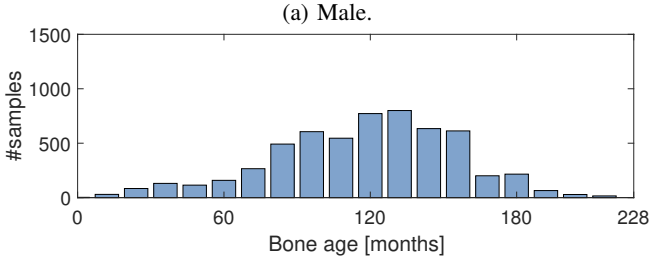
The dataset used in this work consists of 12611 radiographs of the hand and wrist collected by the Radiological Society of North America in the context of the Pediatric Bone Age Prediction Challenge [14]. The institutional review boards of the organizing committee approved the study. The dataset contains 6833 radiographs of male patients and 5778 radiographs of female patients. The annotated estimated bone ages, assessed by trained physicians using the G&P hand atlas, range between 0 and 228 months. The age distribution of the dataset is not uniform as shown in Fig. 1.

Fig. 2 shows several examples of radiographs in the dataset. The size, orientation, brightness and contrast differ across the samples. In some cases, additional artifacts are visible on the radiographs such as watches, plaster casts, surgical screws and assisting nurses. Sometimes parts of the hand, such as fingers, are missing. These artifacts heavily complicate the traditional segmentation methods discussed in Section I.

Given the size of the dataset, we chose to split the data into a train/validation/test set, as opposed to performing a k-fold cross-validation, to reduce computational demands. First, the data is split by gender. Next, an age-stratified split is generated for each gender based on the bone age estimated by trained physicians. Table I provides an overview of the distribution of the patients in the various sub-datasets.



(a) Male.



(b) Female.

Fig. 1: Distribution of estimated bone age assessed by trained physicians. The bone ages are not uniformly distributed.

TABLE I: Distribution of male and female patients in train, validation and test set.

	Train	**Validation**	**Test**	***Total***
Male	5532	691	610	6833
Female	4681	571	526	5778
Total	10213	1262	1136	12611

III. Methodology

In this section, we explore a novel method for automated bone age prediction using a radiograph of the hand and wrist. Our approach consists of three steps. First the images pass through a preprocessing stage to prepare for the next steps and to augment the dataset. Next, the first modeling stage uses a deep learning approach for estimating the bone age. Finally, in a third step, Gaussian process regression is used to refine the estimated bone age by exploiting variations in the predictions of the deep learning stage when predicting rotations and flips of the same input radiograph.

A. Preprocessing and Augmentation

First, the radiographs are rescaled to a size of 224 x 224 pixels, as required by the deep learning VVG16 [15] network that was used. During scaling, the aspect ratio of the images is not altered as they are padded with black pixels.

Subsequently, edge enhancement is applied to highlight the bones in the radiographs. The enhancement consists of a convolution of the pixel data with the enhancement matrix E in Equation (1) where e represents the enhancement strength. The parameter e has to be tuned, which is discussed in the next subsection. A larger value results in sharper edges but also enhances noise present in the images

Finally, the enhanced image passes through a data augmentation stage which is essential as training deep learning networks requires large amounts of data [12]. Here, a single radiograph is rotated 18 times in a range of $[-90, 90]$ degrees. A mirror image along the y-axis of the resulting 19 radiographs is also generated, resulting in a total of 38 images from 1 radiograph. These two steps simulate the cases in the dataset where the hand was not lined up correctly with the y-axis or where a mirror image of the radiograph has been digitized.

$$E = \begin{pmatrix} -1 & -1 & -1 \\ -1 & e & -1 \\ -1 & -1 & -1 \end{pmatrix} \tag{1}$$

B. Modeling with Deep Learning

Deep learning has proven to drastically outperform traditional image recognition and detection methods based on manual feature engineering. In this paper, deep learning is applied for the initial estimation of the bone age using a single radiograph.

We use VGG16, a popular deep learning architecture often used for image processing and recognition. VGG16 is a convolutional neural network consisting of 16 layers. For a full discussion we refer to [15]. As the number of images in the dataset is too limited to completely train such a deep architecture from scratch, we apply transfer learning [16] by first training the VGG16 network with the Imagenet [17] dataset. As the VGG16 architecture is designed for classification, we need to modify the architecture towards regression. For this, we append two dense layers of size n_1 and n_2. Finally a dense layer with 1 output is appended for the final regression output of the model. To prevent overfitting, we append a dropout layer with dropout probability of p_1 and p_2 after the first two appended dense layers respectively. The first two appended dense layers have rectified linear unit (ReLU) activation functions as commonly used in deep learning networks and the output layer has a linear activation function, as typically used in regression settings.

The parameters n_1, n_2, p_1 and p_2 represent hyperparameters of the network. To reduce overfitting and fully harness the power of the deep learning model, these parameters have to be optimized. For this, we use Bayesian optimization, an efficient optimization method for tuning the hyperparameters of machine learning models [18]. For a detailed description of Bayesian optimization, we refer to [18], [19]. The ranges of the hyperparameters and final optimal hyperparameters are shown in Table II. Other hyperparamters such as the number of extra appended layers were not tuned to reduce the computational demand.

TABLE II: Range of hyperparameters and optimal hyperparameters for the image preprocessing and deep learning modeling stages, tuned using Bayesian optimization.

parameter	**type**	**min**	**max**	**optimal**
k	integer	9	10	9
n1	integer	64	1024	512
n2	integer	64	512	512
p1	float	0.1	0.5	0.12
p2	float	0.1	0.5	0.14

The complete network, including the pre-trained VGG16 part, is trained using the Adam optimization algorithm [20].

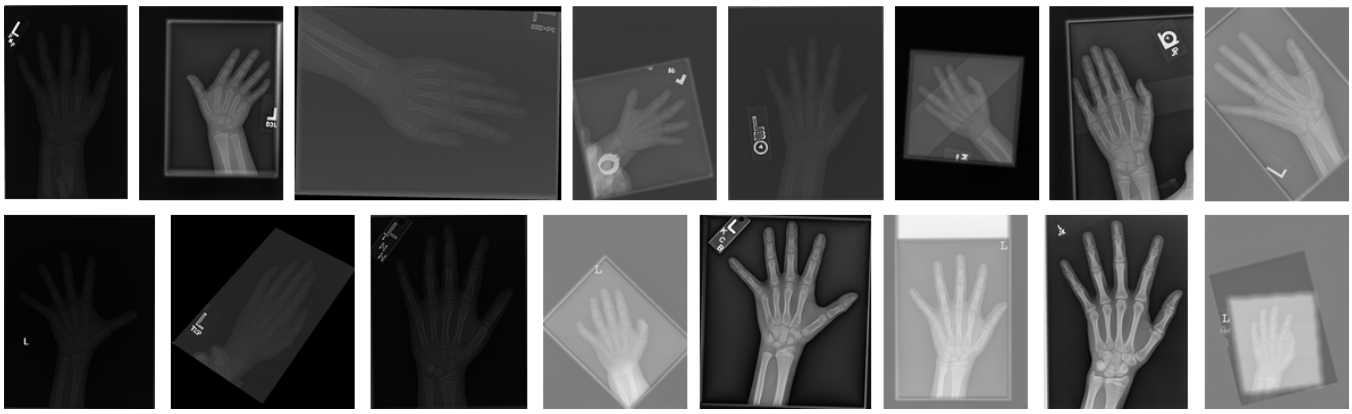


Fig. 2: Example of radiographs in the dataset showing variations in size, contrast, orientation and brightness. Sometimes additional artifacts are visible.

The used loss function is the Mean Absolute Difference (MAD), also known as Mean Absolute Error, defined as:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y_i}|, \tag{2}$$

where n represents the number of samples, y_i represents the physician-estimated bone age of sample i in months and $\tilde{y_i}$ represents the model-estimated bone age for sample i. A lower MAD score represents a closer match with the annotation of a trained physician.

All training data is augmented and randomly permuted after which training is performed in minibatches of 64 radiographs.

C. Modeling with Gaussian Process Regression

The preprocessing step, data augmentation step and deep learning step provide a powerful method for bone age prediction of a single image. However, the predictions of the method are still sensitive to random rotations or flips of the input image. Fortunately, the sensitivity to this type of variations can be exploited to improve overall predictions. Each radiograph is preprocessed and augmented as discussed in Section III-A. Then, for each augmented version of a single radiograph, a prediction is made with the previously described deep learning model. Hence, the data augmentation is not only used during training but also during prediction. This results in a set of 38 predictions for a single radiograph.

Typically, these results are then aggregated by averaging [10]. However, to further improve results we explore the novel method of using a Gaussian Process Regression (GRP) [13] model for aggregation. A Gaussian process is a powerful kernel-based method for modeling and interpolation [21]. The form of the kernel can be chosen and tuned to fit the application needs. In this work, we use the Matèrn$\frac{5}{2}$ kernel as it is commonly used in GPR. A detailed description of GPR and the Matèrn$\frac{5}{2}$ kernel is given in [13].

The parameters for the GPR model are trained using maximum likelihood estimation [13]. The Gaussian process learns to estimate the physician-based bone age using a vector of prediction scores for rotated and mirror images of a single radiograph. This allows the GPR model to fully exploit the original sensitivity of the deep learning model to random rotations and flips of the input image. The final complete model was evaluated using the previously introduced MAD score in Equation (2).

IV. Results and Discussion

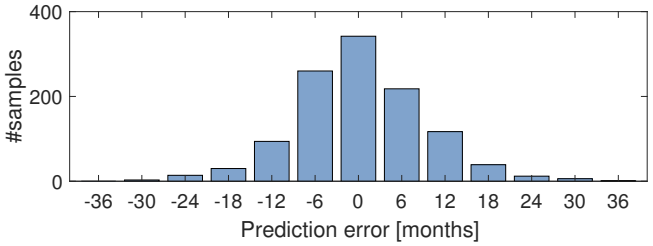
An overview of the final results for the train, validation and test set is shown in Table III where the difference with the original method of only using deep learning is noted between brackets. Note that the test set has not been used to train any of the models or tune any of the hyperparameters. It represents a set of new, unseen patients. The results show that the performance of the test set is in line with the performance of the train and validation sets.

The first row of the table indicates the results when only the original radiograph is used for prediction with deep learning as in the current state-of-the-art. These results are comparable to the results of other deep learning solutions for bone age prediction. The second row of the table indicates the results when the radiograph is augmented during prediction and the average of several predictions is used. Note that this augmentation substantially improves the accuracy of the model. Finally, the third row of the table indicates the results when Gaussian process regression is used to exploit variations in the prediction output of the deep learning model to rotations and flips of the same radiograph. It is clear that this leads to a further substantial improvement, increasing the average accuracy with almost a full month on the unseen test set. Our final model has an accuracy within one year of 94.45% and an average spread of 6.80 months.

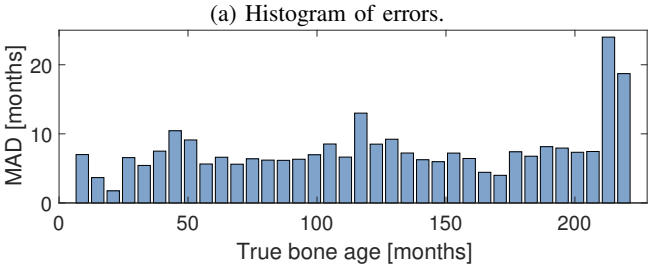
To further analyze the model, a histogram of the errors and a graph of the distribution of the MAD score versus age is shown in Fig. 3. The histogram of the errors indicates that for some samples, large errors are present. The radiographs of these samples are too dark or overexposed, leading to few distinguishable features. Further image preprocessing could improve these results.

TABLE III: Resulting MAD in months. The difference with the original method is noted between brackets. The proposed methodology of using Gaussian process regression for aggregating augmented deep learning results clearly improves the standard state-of-the-art deep learning performance.

	Train	Validation	Test
Original	7.60	8.42	7.74
Augmented + mean	6.94 (-0.66)	7.47 (-0.95)	6.96 (-0.79)
Augmented + GPR	**6.56 (-1.04)**	**7.26 (-1.16)**	**6.80 (-0.94)**



(a) Histogram of errors.



(b) MAD distribution across age.

Fig. 3: Further analysis of the errors on the unseen test set for the complete setup with deep learning and Gaussian process regression for aggregation.

The distribution of the MAD in function of age shows two peaks near the edge of the domain. This indicates a lack of data for the model to make accurate predictions as reflected by Fig. 1. Extending the dataset with samples near the edges could improve these results. For the other ages, scores are typically good and there is little deviation from the trend.

V. Future Work and Conclusions

Bone age assessment is essential in many medical use cases, and even in legal contexts as the rights of asylum seekers depend on their age. However, the current bone age estimation procedures are time consuming and allow for interpretation leading to large inter- and intra-observer variations. Deep learning has proven to be a powerful method for estimating bone age predictions. In this work, we improve upon the state-of-the-art by introducing a novel method for bone age prediction using deep learning by adding a Gaussian process regression stage in which predictions for rotated and flipped versions of the same radiograph are aggregated. Our preliminary results show good performance with a MAD score of 6.80 on the unseen test set and show an increase in accuracy over only using a deep learning network. Future work will focus on using multiple assessments of a single radiograph in the model to make our approach more robust against inter- and intra-observer variations in the training set. Our research contribution shows that variations in the output of a deep learning model, due to rotations and flips of an input image, can be exploited to improve accuracy. This is not only useful for bone age estimation but for other medical or general image analysis use-cases as well.

References

[1] A. M. Mughal, N. Hassan, and A. Ahmed, "Bone age assessment methods: A critical review," *Pakistan journal of medical sciences*, vol. 30, no. 1, p. 211, 2014.

[2] W. W. Greulich and S. I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist," *The American Journal of the Medical Sciences*, vol. 238, no. 3, p. 393, 1959.

[3] D. B. Darling, *Radiography of Infants and Children.*. Ch. C. Thomas, 1962.

[4] P. J. Sauer, A. Nicholson, D. Neubauer, *et al.*, "Age determination in asylum seekers: physicians should not be implicated," 2016.

[5] D. King, D. Steventon, M. O'sullivan, A. Cook, V. Hornsby, I. Jefferson, and P. King, "Reproducibility of bone ages when performed by radiology registrars: an audit of tanner and whitehouse ii versus greulich and pyle methods," *The British journal of radiology*, vol. 67, no. 801, pp. 848–851, 1994.

[6] F. Cao, H. K. Huang, E. Pietka, and V. Gilsanz, "Digital hand atlas and web-based bone age assessment: system design and implementation," *Computerized medical imaging and graphics*, vol. 24, no. 5, pp. 297–307, 2000.

[7] A. Zhang, A. Gertych, and B. J. Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 299–310, 2007.

[8] H. H. Thodberg, "An automated method for determination of bone age," *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 7, pp. 2239–2244, 2009.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[10] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Int. Conf. on Medical Image Computing and Computer-assisted Intervention*. Springer, 2013, pp. 411–418.

[11] J. R. Kim, W. H. Shim, H. M. Yoon, S. H. Hong, J. S. Lee, Y. A. Cho, and S. Kim, "Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency," *American Journal of Roentgenology*, pp. 1–7, 2017.

[12] H. Lee, S. Tajmir, J. Lee, M. Zissen, B. A. Yeshiwas, T. K. Alkasab, G. Choy, and S. Do, "Fully Automated Deep Learning System for Bone Age Assessment," *Journal of Digital Imaging*, pp. 1–15, 2017.

[13] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006.

[14] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, pp. 170–236, 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[18] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[19] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] I. Couckuyt, T. Dhaene, and P. Demeester, "oodace toolbox: a flexible object-oriented kriging implementation," *Journal of Machine Learning Research*, vol. 15, pp. 3183–3186, 2014.