An Empirical Study on Global Bone Age Assessment

Felipe Torres[1], Cristina González[1], María Camila Escobar[1] Laura Daza[1], Gustavo Triana[2], Pablo Arbeláez[1]

[1] Universidad de los Andes, Cra 1 N° 18A - 12, Bogotá, Colombia
[2] Fundación Santa Fe de Bogotá, Cra. 7 ##117 -15, Bogotá, Colombia

ABSTRACT

Bone Age Assessment (BAA) is a task performed by physicians to estimate the skeletal development of a pediatric patient. Tipically physicians perform this exam by doing a manual analysis of the X-ray image of the non-dominant hand of a child, either by taking the image as a whole or paying attention to certain anatomical Regions Of Interest (ROIs). Over the years, several datasets have been proposed in order to generate automated methods to perform this task. Most notably, in 2017 the Radiological Society of North America (RSNA)[1] created the Pediatric Bone Age Challenge, which encouraged the development of machine learning approaches for this task. In this paper, we present *GPNet* a convolutional neural network capable of performing BAA precisely and effectively by analyzing the whole hand in a single forward pass. We train GPNet using the training data available from the dataset created in the RSNA challenge and evaluate our method using the validation set. We use the testing set to compare our performance with the current state-of-the-art and find that GPNet significantly outperforms previous methods. During our architecture search we perform several experiments to demonstrate the effect of different layers, proving that some blocks do not contribute to the performance of the network, but instead they affect it. As a result, we are able to develop a method that reduces the number of trainable parameters by nearly 82.15 M in comparison to the state-of-the-art, while improving the performance.

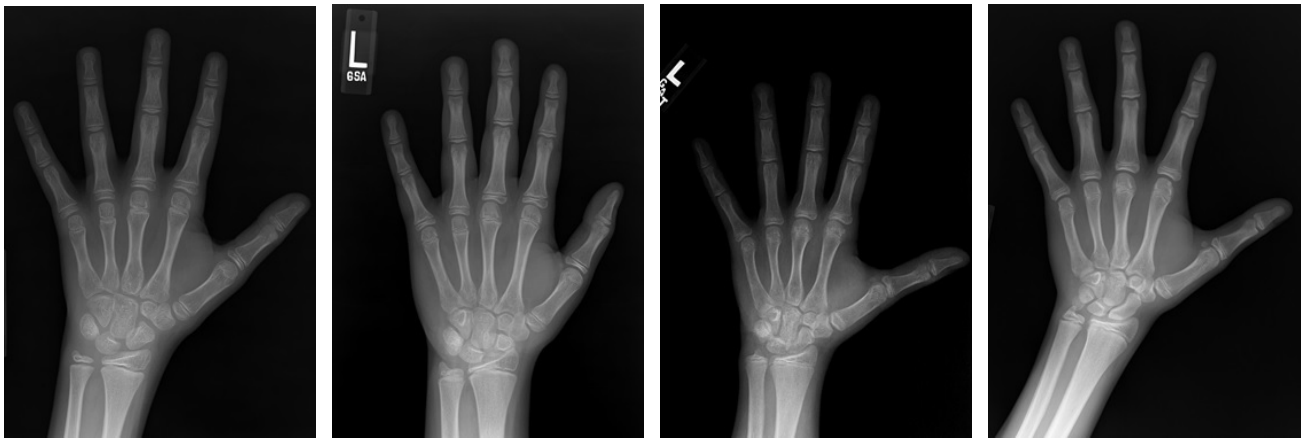
Keywords: Bone Age Assessment, Computer Aided Diagnosis, Convolutional Neural Networks, Machine Learning, Pediatrics

1. INTRODUCTION

The bone age of a child is a measurement of his or her skeletal development. Unlike chronological age, this indicator considers factors such as age, race and gender. In pediatric patients, bone age is useful to diagnose growth and endocrine disorders, and to predict the final height of the patient. Bone Age Assessment (BAA) is performed by comparing a radiograph of the non-dominant hand and wrist against standardized images corresponding to different bone ages. The most commonly used atlases for BAA are Greulich and Pyle (G&P)[2] and Tanner and Whitehouse (TW2).[3] The former is composed of images at different ages for each gender, while the latter involves the scoring of 20 Regions Of Interest (ROIs) to determine the final bone age. However, since the images are compared manually, these approaches are observer dependent and may vary depending on the physician's experience. Furthermore, these manual methods are prone to intra and inter observer errors given the radiologist's expertise or possible variations on the image. An example of the difficulty of this task is shown in Figure 1, where we display four different hands from similar bone ages.

With the success of machine learning techniques in computer vision tasks, such as image classification,[4] detection[5] and semantic segmentation,[6] the use of automatic methods has been proposed to improve the accuracy of BAA. Early methods were manually designed to extract specific ROIs or features and used them to classify the images.[7–9] More recently, the use of deep learning has been proposed to avoid the necessity of feature engineering and allow the fully automation of BAA. The introduction of the Radiological Society of North America (RSNA) in the Bone Age Pediatric Challenge in 2017[10] encouraged the development of deep learning and machine learning approaches for BAA. The winners of this challenge[11] used a deep natural image classification system with an additional branch to include gender information, and achieved a Mean Absolute Difference (MAD) of 4.26 months on the test set.

Regarding the image classification task, recent deep learning approaches have greatly improved the accuracy of machine learning models on benchmarks such as ImageNet.[12] More specifically, methods such as ResNet,[13] DenseNet,[14] Inception-v3[15] and ResNext,[16] have introduced novel ideas that improve the effectiveness of convolutional neural networks in producing accurate classification. Inspired by the capability of these methods to perform a global analysis of the



(a) (b) (c) (d)

Figure 1: Radiographs of male patients with bone age values of (a) 12 years, (b) 13 years,(c) 14 years and (d) 15 years. G&P's method requires visual comparison of these images with the normalized examples contained in the atlas.

image, we develop a network able to perform BAA following G&P's approach, exploiting the global information found in hand radiographs and generating a precise bone age estimation.

In this paper we present **GPNet**, a deep learning approach for BAA that uses Inception-V3 as backbone to extract features from the hand radiograph, and then fuses that information with the gender data. We evaluate our method on the 2017 RSNA Boneage Pediatric Challenge test set and obtain state-of-the-art results using an architecture significantly smaller than the previous leading method.

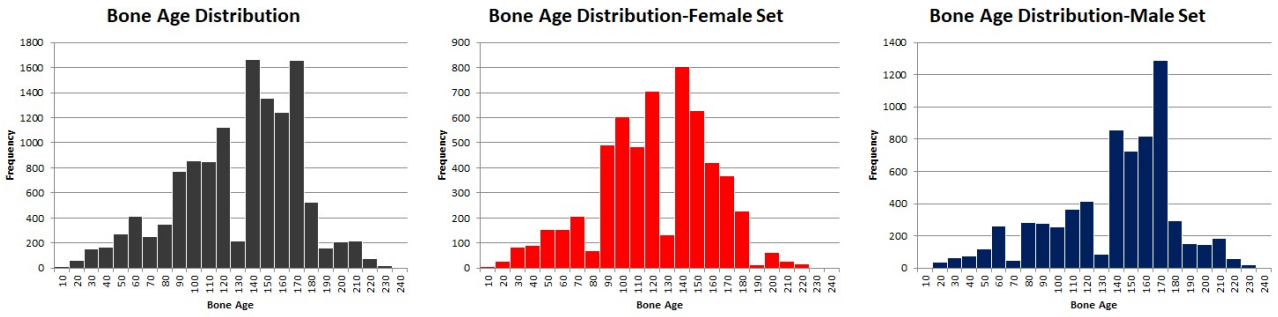
2. RELATED WORK

2.1 Bone age assessment datasets

The first methods used to determine skeletal maturity were based on BAA atlases. These atlases contain standard plates which represent the central tendency of a specific chronological age group. G&P established a normalized atlas from radiographs of upper middle-class boys and girls acquired by the Brush Foundation Study of Human Growth and Development from 1931 to 1942 in the United States.[17] The construction of this guideline was done under the assumption that in healthy children skeletal maturation is uniform, *i.e.* all bones have the same appearance at a certain chronological age and the development of the bone centers follow fixed patterns.[18] Therefore, the use of this rule has limitations in applicability.

In 2005, Vicente Gilsanz and Osman[18] proposed a set of digital data as an alternative to reference atlases. For its development, they considered the central skeletal maturity indicators found in six ossification centers for each sex and age group. In this way, they generated one single idealized image that represents the perfect average for each age group, spaced every six months from 2 to 6 years, and annually from 6 to 17 years. To establish an algorithm that avoids bias towards a particular population, it is necessary to use a dataset that includes high race variability. Thus, in 2007, the Children's Hospital of Los Angeles (CHLA)[19] collected a total of 1,390 left-hand radiographs of children between 1 and 18 years of Asian, African-American, Caucasian and Hispanic descent. Unfortunately, the data is not publicly available. The mentioned datasets do not contain a significant number of images for their use in the current experimental framework.

In 2017, the RSNA created the dataset for the Pediatric Bone Age Challenge. The data was provided by Stanford University, the University of Colorado and the University of California - Los Angeles. The dataset comprises radiographs in png format taken from the non-dominant hand of both male and female patients between 0 and 240 months of age (0-20 years), with bone age annotations made by trained radiologists.[10] The dataset is composed of 14,236 images divided into 3 sets: 12,611 for training, 1,425 for validation and 200 for testing. 46% of the dataset corresponds to female patients and 54% to male patients. The average size of the images is 1,694 $\times$ 1,284 pixels. For the training subset, the age for both splits of the dataset (male and female) resembles a Gaussian distribution, centered around 117 months for females and 135 months for males. Figure 2 presents the data distribution for the whole training set and for each gender. The dataset contains images with a difference in bone age of less than 3 months, complicating the classification process. Therefore, a



(a) (b) (c)

Figure 2: Data distribution for (2a) the training set, (2b) the female training subset and (2c) the male training subset of the RSNA's Pediatric Bone Age Estimation Challenge

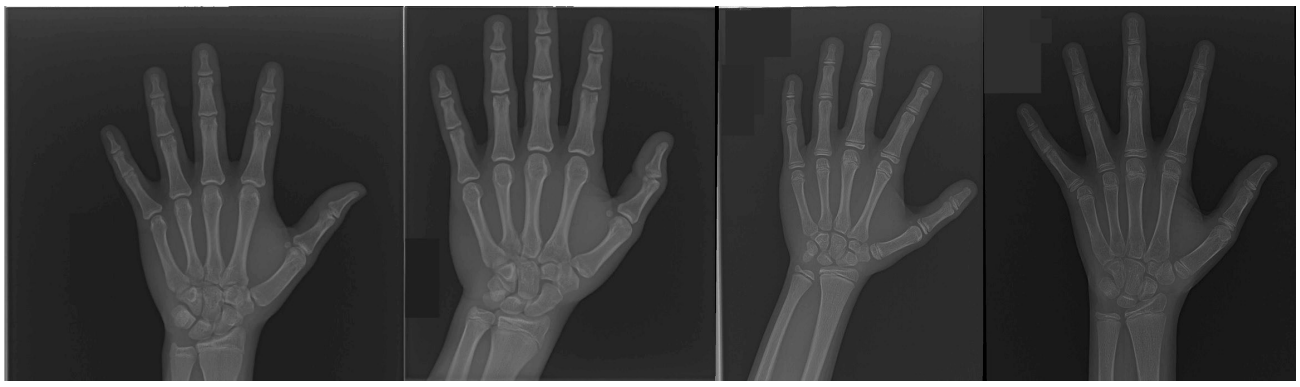
high performing model must take into consideration fine details to correctly classify similar bone ages. Figure 3 shows an example of this observation of the RSNA dataset.

2.2 Bone age assessment methods

Manual assessment. The G&P method consists in comparing the radiograph with a reference atlas to determine the most suitable bone age. This approach is the most popular among radiologists because it is the least time consuming manual method currently available. Nevertheless, BAA through the G&P method can vary in precision depending on the expertise of the radiologist and variations between radiographs. In contrast, in the TW2 method the hand is divided into 20 ROIs, the development of each ROI is considered individually and is classified in a stage. Finally, to determine bone age, the different stages for each ROI are taken into account.[20] The TW2 strategy is more precise than G&P because it makes a local analysis on the radiograph. However, this process tends to be more tedious and time consuming for radiologists.

Automated assessment. Several attempts to automate BAA have been developed over the years. One that is currently used in clinical settings is the private software BoneXpert[9] created in 2009 by Hans Henrik Thodberg and Sven Kreiborg. BoneXpert was generated using patients from a Danish cohort, therefore the reliability is not guaranteed when assessing data from different countries. The BAA is based on edge detection and active appearance models[21] to generate candidates and make comparisons according to G&P and TW2 approaches by taking into account the bone structures on each hand.

Other automated methods for BAA have been developed in the framework of the RSNA Boneage Pediatric Challenge. 16Bit, the winners of the challenge in 2017, used the Inception-V3 network combined with a network for gender information, and added two 1000-neuron densely connected layers before the final month prediction.[11] Another approach,



(a) (b)

Figure 3: Examples of complex classification in the RSNA dataset.
(a) Left side, Hand of a patient with 204 months of bone age. Right side, hand of a patient with 203 months of bone age.
(b) Left side, Hand of a patient with 156 months of bone age. Right side, hand of a patient with 153 months of bone age.

presented by Iglovikov et al,[22] included manually labeling three keypoints on a fraction of the images, and implemented a deep convolutional neural network to create a regression model.

2.3 Image Classification Networks

Inspired by the study of the cells present in the visual primary cortex,[23] Fukushima designed the first convolutional neural network, the Neocognitron.[24] This job introduced the convolutional and spatial averaging layers. Moving on, Yann Lecun made use of these two layers in addition to his backpropagation algorithm[25] to produce a convolutional neural network that successfully generalized its weights on the hand drawn digit recognition task of the U.S. Postal Code.

Since the introduction of AlexNet[4] in 2012, neural networks have grown deeper to extract richer features and significantly improve the accuracy of the methods. To tackle the vanishing-gradient issue caused by adding many layers, He et al.[13] introduced residual connections. As a result, subsequent networks have explored different ways to use skip connections to create more effective blocks.[16,26] In addition to residual connections to tackle the optimization problem, Huang et al. introduced a network called DenseNet that connects each layer to every other layer in the blocks.[14] Consequently, this approach strengthens the feature propagation and reduces the amount of trainable parameters of the model. Moreover, recent models have included densely connected convolutional layers to better use the feature information and reduce the amount of trainable parameters.[27,28]

Furthermore, Szegedy et al. proposed a method called Inception[29] to better utilize the computing resources inside the model, thus increasing the depth and width of the network while keeping the parameter count low. This model introduces Inception layers, which are namely a combination of the outputs of the 1×1, 3×3 and 5×5 convolutional layers concatenated into a single one, allowing these internal layers to assess the best filter size to learn the required information. Recently Szegedy et al. have updated the inception architecture in order to increase the size of the network maintaining a low amount of parameters.[30]

3. OUR METHOD

We propose a global approach to address this task inspired on the methodology of G&P and 16Bit, the winning method of the Pediatric Bone Age Challenge by the RSNA in 2017. 16Bit is a model that uses the Inception-v3 architecture to extract image features from 500×500 hand X-rays. The method combines this information with the gender mapped to a 32 bits learnable embedding by passing it through a 1000-neuron fully connected layer. Finally, they use a 1000-neuron regressor to predict bone age.

We build upon this method by reducing the number of trainable parameters in the model without sacrificing performance. To do this, we decrease the number of feature maps and the spatial resolution of the output by adding a convolutional layer with $stride = 2$ to the final output of the backbone. To consider the gender information in our model, we replicate the binary value 32 times and concatenate it to the image features. Finally, we pass this information through a 500-neuron regressor to predict the final bone age in months. With this modifications we reduce the number of trainable parameters to nearly 82.15 M, in turn, we decrease the training time and overfitting of the model. Additionally, this change allows us to use input images with greater resolution to decrease the loss of information caused by the original resizing. Figure 4 shows an overview of our method. We train *GPNet* using an Adam optimizer with an initial learning rate of 1×10^{-3} until convergence and reducing the learning rate by a factor of 0.1 once a loss plateau is reached.

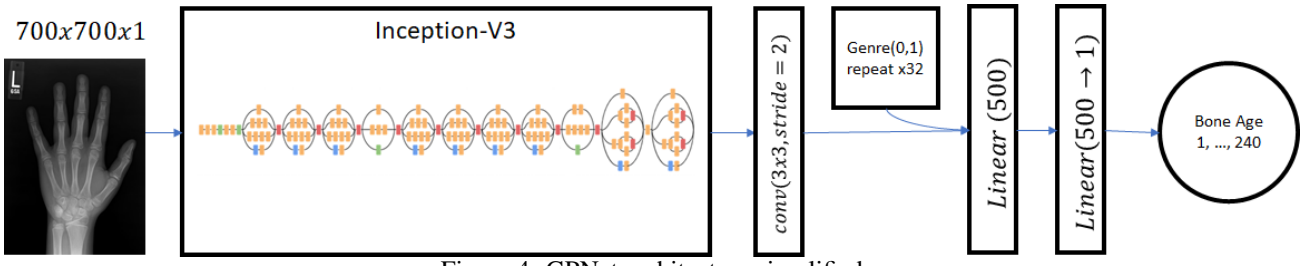


Figure 4: GPNet architecture simplified

4. EXPERIMENTS

To determine the effect of each of the proposed modifications on the baseline, we conducted a comprehensive ablation study in the validation set of the RSNA dataset. For each step of this study we add one of the suggested modifications to the architecture, and train from scratch until convergence. When obtaining a boost in performance with a certain configuration we use it as the starting point for the following experiments. Table 1 shows the results obtained for the ablation study. The first row corresponds to the baseline.

Table 1: Experimental results. The first row shows the performance of the baseline along with the characteristics of the network, the image input size and the number of parameters in the model. The last row shows the performance of the final model with all the modifications included.

Number of fully connected layers	**Fully connected layers size**	**Backbone output size**	**Gender embedding**	**Input image size**	**Backbone**	**Number of parameters**	**MAD**
2	1000	100352	fully connected	500x500	Inception-V3	123.17M	6.506
2	1000	100352	fully connected	500x500	ResNet-101	143.88M	6.858
2	1000	100352	fully connected	500x500	ResNeXt-50	33.00 M	6.610
2	500	100352	fully connected	500x500	Inception-V3	72.48M	6.476
1	500	100352	fully connected	500x500	Inception-V3	71.98M	6.421
2	1000	204800	fully connected	700x700	Inception-V3	227.62M	6.440
2	1000	100352	replicated	500x500	Inception-V3	123.17M	6.384
1	500	33712	fully connected	500x500	Inception-V3	59.11M	6.332
1	500	82976	replicated	700x700	Inception-V3	82.15M	**6.321**

4.1 Backbone Selection

We study the effect of two high performance classification convolutional neural networks: ResNeXt-50 and ResNet-101. The second and third row of Table 1 present the results obtained for these two backbones. The results demonstrate that the Inception-v3 architecture extracts more useful image features for the BAA task. Furthermore, this backbone is less likely to present overfitting to the given data. Therefore, we maintain this architecture as the backbone of our method.

4.2 Size and number of fully connected layers

The 16Bit model has a 1000-neuron fully connected layer to process the image features and the gender information associated to the image and a last one that functions as a regressor to predict the final bone age. In order to determine the effect of the size of these layers, we reduce their size by half. The results show that this modification reduces the number of trainable parameters by more than 50 M, while yielding a slightly better performance. Additionally, to determine if this number of layers are required, we reduce the number of layers in our model to one. Thus, this layer takes the gender information along with the image features extracted from the backbone and predicts the bone age in months. The results show that this change reduces the number of parameters without affecting the performance. Accordingly, with this reduction in the number of trainable parameters of the model, we improve its generalization capacity.

4.3 Image Resolution

The baseline model takes as input images of size 500×500. However, the original images of the dataset have an average resolution of 1,694 $\times$ 1,284 pixels, meaning that during the resizing more than 70% of the spatial information is lost. Figure 5 shows a comparison of the quality of the texture information on some ROIs in the hand when resizing the image to 500×500 and 700×700. When reducing the size of the image the visual information is smoothed and the ossification patterns are lost. As a result, we study the effect of using bigger images as input for our model (700×700). Table 1 shows that w.r.t the baseline the performance is indeed improved by 0.6 moths with this modification, but the number of parameters is nearly doubled.

Figure 5: Texture preservation over the two image resolutions. Best viewed in electronic format.

4.4 Gender information

In the 16Bit model the gender information is passed by a 32-neuron fully connected layer before concatenating it to the image features. This part of the model implies that they are learning an intermediate representation of 32 bits for a binary value. In contrast, we replicate the binary value associated with the gender information 32 times and concatenate it with the image features. This ratio allows gender information to really affect the final prediction. We report the results of this experiment in Table 1. The results show that if we replicate the gender value instead of learning an intermediate representation for it, the model can take advantage of this information in a more appropriate way, which is reflected in a 0.122 month increase in performance.

4.5 Network Output

The baseline model eliminates the final concatenation layer of the Inception-v3 architecture, flattens it and concatenates it with gender information. Processing this data with a 500-neuron fully connected layer generates a considerable amount of trainable parameters for our network. For this reason, we introduce a 3×3 convolution with $stride = 2$ to reduce the spatial resolution as well as the number of channels and produce a 33712 features vector. The eighth row of Table 1 shows that this modification leads to an additional 12M parameters reduction with a performance increase of 0.1.

4.6 Full model

We integrate all the previously mentioned modifications to our final model, GPNet. We show the results obtained for our model in the last row of Table 1. These changes reduce the the number of trainable parameters from 123 M to 82.15 M and increase the performance in 0.66 months. Finally, we compare the performance of GPNet w.r.t 16Bit performance on the test set of the RSNA dataset. These results are shown in the Table 2.

Table 2: Performance comparison of GPNet and 16Bit methods on the RSNA's Boneage Pediatric Challenge test set. The first row demonstrates the performance of our baseline, while the second row demonstrates the performance of GPNet.

Method Name	**Input image size**	**Number of parameters**	**MAD**
16Bit[11]	500×500	123.17M	5.58
GPNet	700×700	**82.15M**	**4.92**

We obtain different testing performances between 16bit's model and our implementation, we note that this difference of performance is a result of our re implementation of the architecture, as the authors of the aforementioned model did not release any code nor model; thus for our experimental setup, we re implemented it according to the description given, to generate a fair comparison between it and our approaches.

5. CONCLUSION

In this work we present GPNet, an automatic approach for BAA that leverages global information of hand X-ray images inspired on the methodology of G&P. Our method has a greater generalization capacity, with a considerable reduction in number of trainable parameters, from 123 M to 82.15 M. Additionally, using higher resolution images maintains local information that is useful for BAA task. Our model takes advantage of gender information more appropriately by using it in a straightforward manner instead of learning an embedding for a binary number. Finally, GPNet outperforms the previous state-of-the-art method on this task. GPNet source code, models, pre-computed results will be available* to boost relevant research in the topic of BAA.

ACKNOWLEDGEMENTS

This project was partially funded with the Administrative Department of Science, Technology and Innovation (COLCIENCIAS) grant 841-2017.

REFERENCES

[1] "Rsna pediatric boneage challenge," (2017).

[2] Greulich, W. W., Pyle, S. I., and Todd, T. W., [*Radiographic atlas of skeletal development of the hand and wrist*], vol. 2, Stanford University Press (1959).

[3] Tanner, J., Whitehouse, R., Marshall, W., and Carter, B., "Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height.," *Archives of disease in childhood* **50**(1), 14–26 (1975).

[4] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in Neural Information Processing Systems 25*], Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds., 1097–1105, Curran Associates, Inc. (2012).

[5] Ren, S., He, K., Girshick, R. B., and Sun, J., "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017).

[6] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in [*Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*], 833–851 (2018).

[7] Tsao, S., Gertych, A., Zhang, A., Liu, B. J., and Huang, H. K., "Automated bone age assessment of older children using the radius," in [*Medical Imaging 2008: PACS and Imaging Informatics*], Andriole, K. P. and Siddiqui, K. M., eds., SPIE (Mar. 2008).

[8] Liu, J., Qi, J., Liu, Z., Ning, Q., and Luo, X., "Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method," *Computerized Medical Imaging and Graphics* **32**, 678–684 (Dec. 2008).

[9] Thodberg, H., Kreiborg, S., Juul, A., and Pedersen, K., "The BoneXpert Method for Automated Determination of Skeletal Maturity," *IEEE Transactions on Medical Imaging* **28**(1), 52–66 (2009).

[10] RSNA, "Pediatric bone age challenge," (2017).

[11] Cicero, M. and Bilbily, A., "Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge," (2017).

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

[13] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *CoRR* **abs/1512.03385** (2015).

[14] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q., "Densely connected convolutional networks," in [*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2017).

[15] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*], 2818–2826 (2016).

*https://biomedicalcomputervision.uniandes.edu.co/index.php/research?id=18

[16] Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K., "Aggregated residual transformations for deep neural networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017).

[17] Gaskin, C. M., Kahn, M. M. S. L., Bertozzi, J. C., and Bunch, P. M., [*Skeletal development of the hand and wrist: a radiographic atlas and digital bone age companion*], Oxford University Press (2011).

[18] Gilsanz, V. and Ratib, O., [*Hand bone age: a digital atlas of skeletal maturity*], Springer Science & Business Media (2005).

[19] Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., and Huang, H., "Bone age assessment of children using a digital hand atlas," *Computerized Medical Imaging and Graphics* **31**(4-5), 322–331 (2007).

[20] Niemeijer, M., van Ginneken, B., Maas, C. A., Beek, F. J., and Viergever, M. A., "Assessing the skeletal age from a hand radiograph: automating the tanner-whitehouse method," in [*Medical Imaging 2003: Image Processing*], **5032**, 1197–1206, International Society for Optics and Photonics (2003).

[21] Cootes, T. F., Edwards, G. J., and Taylor, C. J., "Active appearance models," in [*IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*], 484–498, Springer (1998).

[22] Iglovikov, V. I., Rakhlin, A., Kalinin, A. A., and Shvets, A. A., "Paediatric bone age assessment using deep convolutional neural networks," in [*Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*], 300–308, Springer (2018).

[23] Hubel, D. H. and Wiesel, T. N., "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology* **148**(3), 574–591 (1959).

[24] Fukushima, K., "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics* **36**(4), 193–202 (1980).

[25] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D., "Backpropagation applied to handwritten zip code recognition," *Neural computation* **1**(4), 541–551 (1989).

[26] Zagoruyko, S. and Komodakis, N., "Wide residual networks," *Procedings of the British Machine Vision Conference 2016* (2016).

[27] Larsson, G., Maire, M., and Shakhnarovich, G., "Fractalnet: Ultra-deep neural networks without residuals," (2016).

[28] Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y., "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in [*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (July 2017).

[29] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1–9 (2015).

[30] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A., "Inception-v4, inception-resnet and the impact of residual connections on learning.," in [*AAAI*], 4278–4284 (2017).