Bone Age Detection via Carpogram Analysis using Convolutional Neural Networks

Felipe Torres*, [1] María Alejandra Bravo*, [1] Emmanuel Salinas, [2] Gustavo Triana, [2] Pablo Arbeláez[1]

[1] Universidad de los Andes, Cra 1 N° 18A - 12, Bogotá, Colombia;
[2] Fundación Santa Fe de Bogotá, Cl. 116, Bogotá, Colombia;

ABSTRACT

Bone age assessment is a critical factor for determining delayed development in children, which can be a sign of pathologies such as endocrine diseases, growth abnormalities, chromosomal, neurological and congenital disorders among others. In this paper we present *BoneNet*, a methodology to assess automatically the skeletal maturity state in pediatric patients based on Convolutional Neural Networks. We train and evaluate our algorithm on a database of X-Ray images provided by the hospital Fundación Santa Fe de Bogotá with around 1500 images of patients between the ages 1 to 18. We compare two different architectures to classify the given data in order to explore the generality of our method. To accomplish this, we define multiple binary age assessment problems, dividing the data by bone age and differentiating the patients by their gender. Thus, exploring several parameters, we develop BoneNet. Our approach is holistic, efficient, and modular, since it is possible for the specialists to use all the networks combined to determine how is the skeletal maturity of a patient. BoneNet achieves over 90% accuracy for most of the critical age thresholds, when differentiating the images between over or under a given age.

Keywords: Bone Age, Carpogram, Convolutional Neural Network, Pediatric radiology

1. INTRODUCTION

To determine the maturation and growth status in pediatric patients, specialists analyze hand radiographs known as *carpograms*, as shown in Figure 1. The maximum height that can be achieved by an individual is defined by several factors including physiology, nurturing, and ancestry. Moreover, physicians can predict the maximum height of a patient according to ossification behavior that the cartilages present, mostly in the epiphysis of several bones. Therefore, ossification defines one important parameter to diagnose growth abnormalities.

With the implementation of diagnostic imaging, doctors can assess and identify the current situation of an individual using non-invasive diagnostic methods, which reduces the need of surgeries and their complications. Additionally, diagnostic imaging allows specialists to have a good insight on the status of certain organs and structures. X-Ray imaging, for example, is a suitable and low-cost tool that physicians use to have a good overview of bones and other tissues. However, even with the advances in image acquisition the medical analysis of diagnostic images is not an easy task as it requires expertise and time.

Figure 1. Greulich and Pyle's Image corresponding to a "normal" X-ray for an 8 year old male child

Throughout history, medicine has recognized the importance of tracking skeletal maturity of pediatric patients for observing and monitoring growth and development. Bone age is a measure of the maturation of the bones and this value varies according to gender, age, and race. Several methods have been proposed to obtain this medical parameter. Currently, Greulich and Pyle's atlas (GP)[1] and Tanner and Whitehouse's algorithm (TW)[2] are commonly used in clinical practice; however, both methods rely on the specialist's ability to interpret carpograms. Unfortunately, this task is so hard that different physicians may give inconsistent diagnostics. Discordance between specialists can be generated by differences in experience and to the atlas used to compare the images. Besides, most references

* Equal Contribution

are from middle-class or upper-class children and are not representative of the complete population.[3] Once the bone age is obtained, doctors can predict the maximum height achievable by a person by relying on a set of tables produced by TW in 1975.

Bone age can be evaluated manually or by automated methods that rely on machine learning. Automated methods are more accurate than manual measurements,[4] as they improve diagnostic performance in terms of bone age alterations and save time in interpreting results. Additionally, they are appropriately integrated into the work-flow and decrease the costs of patient care. We hereby present BoneNet as a methodology to assess the bone age estimation with the application of machine learning to perform carpogram analysis. To accomplish this goal, we levarage recent and high-performance Convolutional Neural Network (CNN) architectures and explore different strategies to obtain an accurate prediction for bone age. We implemented our method using two CNN, VGG-VD-16[5] and Inception-v4,[6] that were originally designed for image classification on datasets such as Imagenet.[7] These CNNs are retrained for the task of bone age assesment using a dataset provided by the hospital Fundación Santa Fe de Bogotá, which consists on over 1500 hand X-Rays from pediatric patients. Each of our CNNs divides the data into two classes: younger or older than a specific age. We consider multiple age references during the critical growth period of 7 to 14 years. BoneNet directly analyzes the information on the whole hand X-Ray in order to predict the correct maturity age. We trained a set of nine CNNs that combined are used to approximate the bone age of a patient. BoneNet provides a method for accurate prediction of bone age as it eliminates the bias produced by the replicability and expertise of specialists and establishes a model adapted to the Colombian population and its ethnics.

2. RELATED WORK

Before machine learning influenced medical practice, Greulich and Pyle[1] built the first systematic attempt to create a methodology to assess bone age in 1959. They developed an atlas of hand X-ray models for each age. This atlas (Figure 2) is composed by a collection of carpograms used as reference for certain ages. In addition, each model has an approximation of the bone age with standard deviation and possible ranges. However, this method has several shortcomings such as intraobserver variability, time consuming diagnostics, and low reproducibility. It is important to notice that Greulich and Pyle's atlas reports information of non-regional, nor specific standards for the Latin American population.

As a complement to GP's atlas, Tanner and Whitehouse's method[2] considers more specific factors of the images to compare. This method requires the specialist to consider the special configuration, size, and shape of several bones from the wrist and the metacarpal. However, this method still suffers from GP's weakness in reproducibility and objectiveness of the diagnostic. Figure 2 shows the differences between TW and GP methods. In contrast with GP and TW's methods, BoneNet does not generate intraobserver variability and is fully replicable as we eliminate the human factor that can produce bias in the experiments.

Few studies that have been performed in the Colombian population for the evaluation of bone age, one of them[8] presents results showing a Gaussian distribution of bone ages. This study has certain limitations: a low sample size (126 radiographs), an age range limited to 5 to 20 years and the absence of national referents. BoneNet on the contrary, takes into account every single age in a range from 0 to 18 which is the measurement for bone age. Also, with the results obtained by our method we establish a reference model adapted to Colombia.

Several attempts to automate bone age assessment have been developed over the years. Nonetheless one that stands out above the rest was created in 2009 by Hans Henrik Thodberg and Sven Kreiborg. As a result, the authors produced the software called BoneXpert.[9] BoneXpert was generated using mostly patients from a Danish cohort, therefore the reliability is not guaranteed when assessing data from different countries. BoneXpert uses an algorithm that follows four steps as described below:

1. Border Recognition.
2. Border Validation.
3. Hand Segmentation.
4. Probability function assignment to each image according to the normalized histogram.

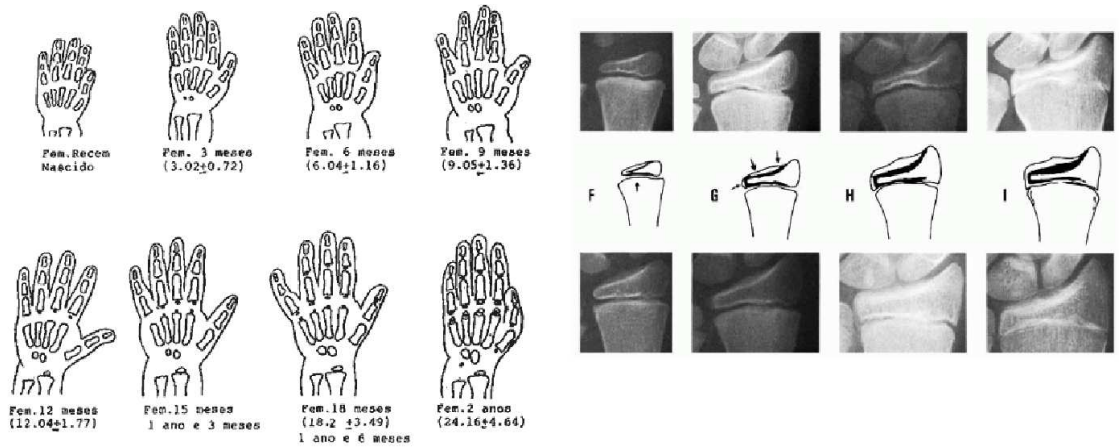

Figure 2. Comparison between two different atlas of bone age examples: Left images correspond to GP's reference and right images correspond to TW's reference.

Once these four steps are completed, the program generates a dynamic model of appearance for each bone; then, the output data is compared with the model references in the database. This modern alternative is already deployed in the European market. In contrast with BoneXpert, BoneNet establishes a model based on artificial intelligence that works on a Colombian cohort. Our method can be trained with patients from different countries as the only parameter that would need to be adjusted would be the data to train the CNN.[9]

Other imaging modalities, such as MRI and CT, have been considered as input for automated bone age assessment. For example, the Aachen University group developed a software to determine bone age using the methodology designed by *IRMA* (Image Retrieval in Medical Applications). This software finds each bone joint in the image, compares the pattern with a reference dataset, and gives scores to each reference. Finally, the program chooses the category with the highest probability and assigns the image to the corresponding bone age.[10] In contrast to this software, BoneNet relies purely on hand X-Rays making the assessment more affordable; additionally, our CNN-based approach is significantly more robust than active appearance models. Finally, the prediction time is reduced as the process only needs a forward passing of the image through the network instead of selecting regions of interest and performing feature engineering.

A recent method,[11] approached the bone age regression problem based on MRI analysis. This study shows very promising results for the estimation of biological age and emphasizes the relevance of an in-depth automatic study for the analysis of diagnostic images. However, the high cost of MRI limits its practical application. In contrast, our method explores the use of a single channel and inexpensive input (X-ray) to train multiple binary models and benefits naturally from global information by identifying the big changes on osseous structures, as well as from fine information present in consecutive ages. Additionally, BoneNet is trained on a Colombian database producing an optimal tool for diagnosis in Latin America.

3. DATASET

The images collected for this project were retrieved from the diagnostic imaging department of the hospital Fundación Santa Fe de Bogotá. These images are saved in their original format (DICOM) which stores them preserving their quality and metadata. Nonetheless, as this format contains sensible information of the patient, anonymization is required and to do so, during the extraction phase, only the fields related to the image, gender of the patient, physiological age and access number are retrieved.

Overall 1402 images were retrieved, most of them belong to physiological and bone ages centered around 9 to 13 years, which are the ages around most of the growth tracking in children takes place. In Figures 3 and 4, we present the distributions of images for each gender. As we aim to differentiate the data that are the most critical for the bone ages

(being the ones closest to the threshold), our experiments focus on data near the boundaries for those reference ages. To split the database for training and validation sets, we randomly selected half of the patients for training and the other half for validation, making sure we had same number of images for each class in both sets.

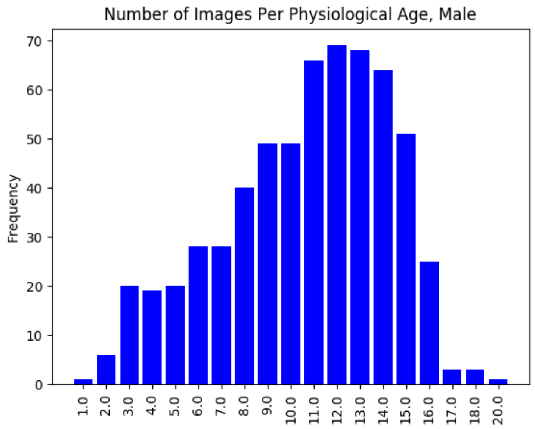

Figure 3. Data distribution for male gender

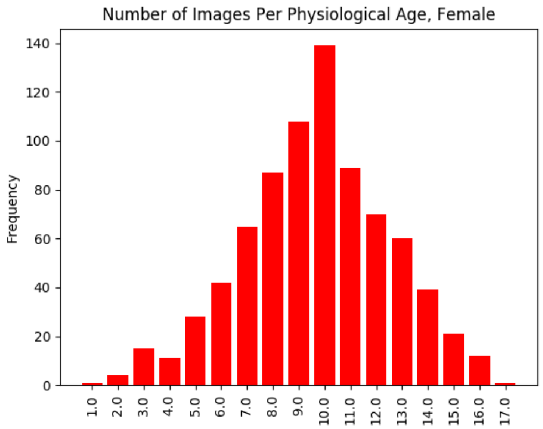

Figure 4. Data distribution for female gender

4. APPROACH

4.1 Class Distribution

According to the bone ages that constitute the database, we formulate skeletal maturity determination as holistic binary classification to build on current high-capacity CNN architectures.[5,6] Nevertheless, compared to the ImageNet classification challenge for which these CNNs were designed, bone age determination requires more detailed and fine grain specialization. Therefore, we start by learning a high-level representation and fine grain information from carpograms by removing the fully connected layers of an ImageNet architecture such as VGG-16[5] and InceptionV4[6] and changing them into fully connected layers with a two-sized vector. The resulting vector gives the probability of the patient to have a higher or lower bone age compared to a specific age. This approach is the commonly used by GP and TW's methods.

Since BoneNet is composed of nine binary CNNs, we split the training dataset into two classes for nine different threshold ages, from 7 to 15 years. These ages correspond to the critical period in which parents mostly track their children's growth, something that can be observed on the distribution of carpograms in Figures 3 and 4. This increase in interest can be explained by considering that during this age interval most parents grow weary about the maximum height that their children can achieve and about the overall growth. Thus, a bone age value below the corresponding age threshold can mean an early diagnosis of a growth abnormality.

4.2 Baseline

As baseline for automated bone age assesment, we train a Support Vector Machine (SVM) using the Bag Of Words descriptor (BoW). BoW can be understood as a set of features that describe an object, these features correspond to specific characteristics that differentiate each class of objects from the rest. For example, to describe a face we would consider features that would contain elements such as nose, ears, eyes and lips. This visual Bag of Words works on the basis of a vocabulary generated by extracting SIFT descriptors at multiple scales on a dense grid in the image and clustering them with k-means. Image-level descriptors are then computed on a spatial pyramid and classified with an SVM.[12]

4.3 BoneNet

We start our method from learned CNNs trained in ImageNet and extend the CNNs with a binary output softmax layer predicting if the maturity age is higher or lower than a given age. To achieve the correct classification of ages relying on the bone age, we use different reference ages as thresholds to train our models for bone age estimation. This binary approach

consider nine reference ages from 7 to 15 years. We experiment training two different CNNs architectures to explore the generalization of our model, VGG-16[5] and inception-4.[6] Thus, we build a model able to classify images between two categories, lower or higher than a specific age.

To train BoneNet we do data augmentation of the images by first cutting right and left hands separately from a single carpogram; then, we flip the right hand in order to have the same position of the hand to classify (only left hands). We rotate the hand varying from -24 to 24 degrees and we use the complement of the image to have the higher numbers corresponding to the bones. For each split, we balance the two classes in the training set using this augmentation.

Using the nine CNNs ensemble we build a multiclass BoneNet. It takes the advantage of multiple trained CNNs for binary classification and given the individual predictions is capable of approximating the maturity age of a patient. During the last stage we consider the predictions of both hands present in the carpogram by averaging their score. We take the best results obtained in the binary phase of the best performing networks and by considering their scores we select the predicted age. This age corresponds to the least reference age in which the consecutive binary models predicted a higher age. For this algorithm we only consider patients of ages 7 to 15.

5. EXPERIMENTAL RESULTS

In this section, we present the experimental results varying the architecture and threshold used for the binary classification. We present classification results, evaluated with average classification accuracy (ACA), in Table 1 and detection results, evaluated with the maximal F-measure (Fmax) and average precision (AP), in Table 2.

Table 1. Binary classification results for the different approaches, varying architecture and threshold

Sex	Male			Female		
Reference Age	Baseline	VGG16	Inception4	Baseline	VGG16	Inception4
7 Years	73.90%	69.00%	**91.12%**	59.46%	67.84%	**87.20%**
8 Years	61.06%	70.61%	**91.08%**	60.58%	67.54%	**93.50%**
9 Years	65.09%	73.38%	**93.07%**	53.21%	70.03%	**90.32%**
10 Years	58.65%	72.00%	**92.01%**	61.34%	68.78%	**86.77%**
11 Years	87.03%	76.10%	**93.11%**	70.06%	65.77%	**91.16%**
12 Years	82.94%	69.86%	**90.19%**	76.00%	57.16%	**85.33%**
13 Years	72.41%	72.75%	**89.23%**	57.90%	69.28%	**85.69%**
14 Years	65.82%	70.38%	**95.80%**	63.37%	68.07%	**91.72%**
15 Years	70.04%	71.03%	**91.25%**	62.21%	65.96%	**91.47%**
Average	70.77%	71.67%	**91.87%**	62.68%	66.71%	**89.24%**

5.1 Baseline

For our baseline, we extracted SIFT descriptors for each image and we trained nine binary SVMs using an intersection Kernel with the implementation of the Bag of Words model (BoW), one SVM for each reference age. Our best results, in Table 1, were 87.03% ACA on male using the 11 years threshold and 76.00% ACA on female with the 12 years threshold. The average score obtained for our baseline are 70.77% ACA on the male set and 62.68% ACA on the female set. Although these scores are higher than the random score of 50%, they are still far from being useful for medical diagnosis. These results indicate that this problem surpasses the capacity of models such as BoW, where handcrafted descriptors are crucial for a correct classification and the algorithm is not learning these descriptors. From the experimentation and the data used in this problem, we found out that most of the difficulties were generated due to the variability of the data itself, especially as the different bone ages start to variate as the skeletal maturity becomes apparent.

5.2 BoneNet

5.2.1 VGG16

VGG-16 is a CNN composed by 16 layers that use 3x3 filters with stride and pad of 1, nonlinear layers ReLU, along with 2x2 maxpooling layers with stride 2. These small-sized filters decrease in the number of parameters compared to other CNNs. The input size of this CNN is of 227x227 pixels and we trained the CNN initializing the weights randomly. We

found batch normalization to be beneficial, as it allows the CNN to learn faster and in a more robust way, as it updates statistics when each batch is completed. We used a batch size of 20 images and a learning rate initialized with a value of 10^{-3} that decreased every 10 epochs with a rate of 10^{-2}.

The results, on Table 1, indicate that VGG surpasses our baseline in most of the cases. It obtained the highest accuracy values at the threshold 11 on the male set with a value of 76.10% ACA and on the female set at threshold 9 with a value of 70.03% ACA. On average, the classification scores were 71.67% ACA on male set and 66.71% ACA on female set. Clearly these scores are higher than the baseline but they are still far for medical use.

5.2.2 Inception-v4

Inception is a CNN that uses the power of parallel computations of different sizes. It is composed of inception models which combine different size of filters per layer in a parallel form and concatenates their results. This CNN is one of the deepest and wider well-known networks.[6] For training Inception-v4 we used a batch size of 25 images and a starting learning rate of 0.001.

The classification results of Inception-v4 obtained the highest accuracy (Table 1). For male, reference age of 14 years achieved 95.80% ACA and for female, reference age of 8 years achieved 93.50% ACA. The average score for male was 91.87% ACA and for female it was 89.24% ACA, these scores are very high when compared to the other methods and are promising for medical diagnosis.

In order to obtain further insights, we evaluated our results as a detection problem, as shown in Table 2. Inception-v4 significantly outperforms VGG-16 in both F-max and AP metrics. The scores obtained with Inception-v4 were: for male 0.864 (Fmax) and 0.962 (AP) in average, and for female 0.808 (Fmax) and 0.925 (AP) in average. The high results show that the algorithm not only classifies correctly but also with a high confidence.

It is important to notice that female scores were lower than male scores, which indicates a relevant difference in bone growth. Also, for some reference ages, the accuracy was lower, indicating that those are the critical ages in which it is more difficult to determine the correct bone age.

Table 2. Binary detection results for the different approaches, varying architecture and threshold

Sex	Male				Female			
Architecture	VGG16		Inception4		VGG16		Inception4	
Reference Age	Fmax	AP	Fmax	AP	Fmax	AP	Fmax	AP
7 Years	0.872	0.937	**0.953**	**0.996**	0.888	0.945	**0.943**	**0.998**
8 Years	0.820	0.864	**0.938**	**0.991**	0.860	0.940	**0.955**	**0.997**
9 Years	0.779	0.857	**0.918**	**0.992**	0.756	0.806	**0.885**	**0.988**
10 Years	0.703	0.847	**0.927**	**0.991**	0.706	0.867	**0.863**	**0.978**
11 Years	0.538	0.559	**0.906**	**0.989**	0.579	0.669	**0.850**	**0.971**
12 Years	0.400	0.375	**0.826**	**0.964**	0.397	0.431	**0.713**	**0.893**
13 Years	0.277	0.247	**0.852**	**0.966**	0.277	0.221	**0.607**	**0.828**
14 Years	0.155	0.101	**0.842**	**0.901**	0.156	0.146	**0.822**	**0.953**
15 Years	0.086	0.065	**0.615**	**0.869**	0.085	0.083	**0.635**	**0.718**
Average	0.514	0.539	**0.864**	**0.962**	0.523	0.567	**0.808**	**0.925**

Figure 5 presents our multi-way classification results for nine classes, 7 to 15 years, taking the average scores of both hands. For male patients the average classification accuracy (ACA) reached 43.03% and for female patients it reached 41.04% ACA. The results indicate the potential that BoneNet has for approximating correctly the bone age. Discrepancy on performance between genders can be explained by the size difference of the sets, as females provide more data to train the models.

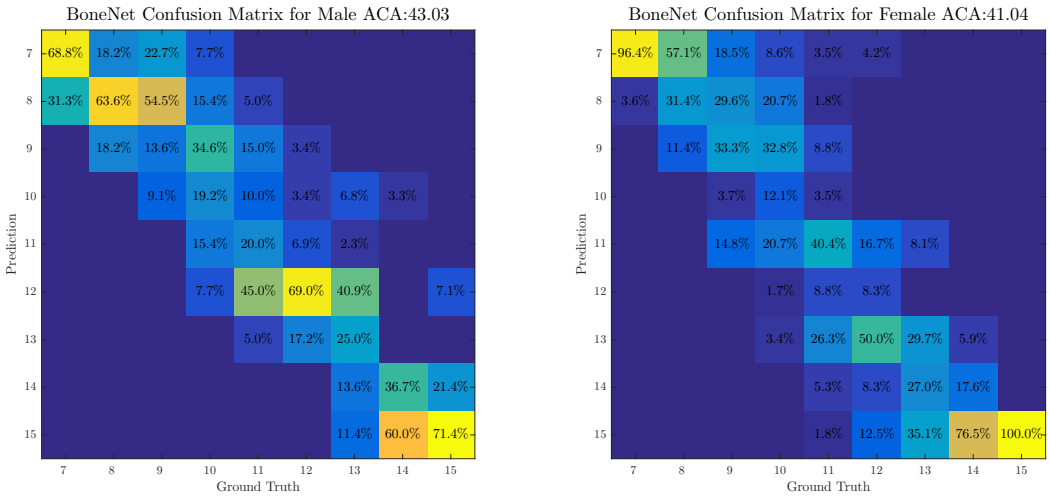

Figure 5. Confusion Matrix of BoneNet for ages 7 to 15. On the left male results and on the right female

Looking deeper at the classification scores of Figure 5, we recognized that the ages at which BoneNet got the highest values correspond mainly to the edges, 7 and 15, and most of the confusion was between consecutive ages.

6. CONCLUSION

We have presented a methodology that is capable of fulfilling the task of assessing the bone age on a binary task. This methodology is robust, reliable and can work with a high precision on both male and female genders. Our method also demonstrates the big performance difference that can be achieved by using a CNN against using methods such as SVMs and can be benefit from larger and more efficient CNNs.

7. ACKNOWLEDGMENTS

The authors gratefully acknowledge NVIDIA Corporation for donating the GPUs used in this project.

REFERENCES

1. W. W. Greulich and S. I. Pyle, *Radiographic atlas of skeletal development of the hand and wrist, based on the brush foundation study of human growth and development*, 1 ed., 1959.
2. J. M. Tanner, R. H. Whitehouse, W. A. Marshall, and B. S. Carter, "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), pp. 14–26, 1975.
3. T. W. Todd, "Age changes in the pubic bone. i. the male white pubis," *American Journal of Physical Anthropology* **3**(3), p. 285–334, 1920.
4. C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in x-ray images," *Medical Image Analysis* **36**, pp. 41–51, 2017.
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., pp. 1097–1105, Curran Associates, Inc., 2012.
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.
7. 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*, pp. 248–255, IEEE, 2009.

8. J. Carrillo, L. Caro, E. Villamor, J. C. Morales, M. Ireton, and J. Monroy-A, "Bone age in a group of colombian schoolchildren," *Acta Medica Auxologica* **33**(2), pp. 113–120, 2001.
9. H. Thodberg, S. Kreiborg, A. Juul, and K. Pedersen, "The bonexpert method for automated determination of skeletal maturity," *IEEE Transactions on Medical Imaging* **28**(1), pp. 52–66, 2009.
10. IRMA, "Bone age assesment," 2008. Berlin: http://ganymed.imib.rwth-aachen.de/irma3_production/bone_age_assessment_demo.current/index.php.
11. D. Štern, C. Payer, V. Lepetit, and M. Urschler, "Automated age estimation from hand mri volumes using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 194–202, Springer, 2016.
12. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR06)* .