Transdisciplinary Engineering for Complex Socio-technical Systems – Real-life Applications
J. Pokojski et al. (Eds.)

doi:10.3233/ATDE200080

Automatic Radiographic Bone Age Assessment Using Deep Joint Learning with Attention Modules

Wei TANG[a], Gang WU [b] and Gang SHEN[a,1]
[a] *School of Software Engineering,* Huazhong University of Science and Technology, *Wuhan, China*
[b] *Tongji Hospital, Wuhan, China*

Abstract. Hand and wrist skeletal radiographs serve as an important medium for diversified medical and forensic tasks involving bone age assessment. As an alternative to traditional atlas-based bone age identification techniques, deep learning algorithms automatically classify the radiographs into predefined bone age classes, provided that the deep neural networks (DNN) have been well trained with large scale annotated datasets. Most of the current bone age classification DNNs directly explore the existing network models developed for other computer vision representations and understanding applications, such as VGG, Inception, and ResNet. In this work, we present a multi-scale attention-enhanced classifier with a convolutional neural network backbone, specifically designed for bone age prediction and trained to learn a subject's bone age and gender jointly. The proposed classifier is trained with the dataset provided by the RSNA machine learning challenge, and the low-level semantic features are then transferred to a smaller Tongji dataset collected from a hospital in China. As demonstrated by the experiments, the proposed classifier achieves the MADs of 0.41 years over RSNA data and 0.36 years on Tongji data, outperforming other single model state-of-the-art and baseline algorithms for the same test. It illustrates that joint learning of gender information plays a critical role in refining the bone age assessment, while the convolution-based attention mechanism helps retrieve the key features.

Keywords. Attention mechanism, bone age, convolutional neural networks, joint learning

Introduction

X-rays and other medical imaging techniques have dramatically changed the landscapes of modern medical practice. One specialized application of the X-radiographs is to provide the evidence for radiologists to estimate the bone age of a child, based on the belief that the radiographs of hand skeletons reflect the maturity degree of the child's bones [1]. Although most children have bone ages identical to their chronological ones, an individual's bone growth may be affected by many other elements, such as genetics, hormone levels, dietary habits, and metabolic disorders, etc. In practice, accurately estimating bone age is critical in identifying many growth-related problems.

[1] Corresponding Author, Mail: gang_shen@mail.hust.edu.cn.

Conventionally, radiologists search in an atlas for a match by visually examining the similarities with the X-ray image, as established in the Greulich and Pyle (GP) method. To reduce the human bias in comparison, the Tanner and Whitehouse (TW) method requires the evaluation between more specific regions of the radiographs and introduces a set of detailed features for the systematic scoring [2] (see Fig. 1 for an example). Atlas-based bone age assessment relies on the standard atlas and personal judgment. Therefore, sometimes two radiologists may disagree on the prediction for the same radiograph. In the past decade, machine learning has made innovative progress in understanding and interpreting medical data. Especially, assorted deep neural networks have proven their potential in assisting the doctors to diagnose many diseases [3, 4]. Several factors contribute to the successes of deep learning applications, including the increasing computing power, more sophisticated network architecture, and the large annotated datasets. Nevertheless, labeling the large dataset is a nontrivial project for most medical tasks. In addition to the reluctance from the medical institutions to share their privacy-sensitive data, the data labeling process is both laborious and costly. As an alternative, transfer learning enables a new application to take advantage of the knowledge learned from other domains, and thus eases the burden of developing large data sets from scratch.

In 2017, the Radiological Society of North America (RSNA) organized a pediatric bone age machine learning challenge. Deep neural network methods won an impressive advantage over competitors in this challenge, with the few best results about 4 months in mean absolute difference (MAD) on test set [5]. Since the bone age only marks the average development phase of bones for the children at a certain age, it exhibits large variabilities across different economies, regions, and races. Therefore, model refinement is needed when the target group differs from RSNA sources.

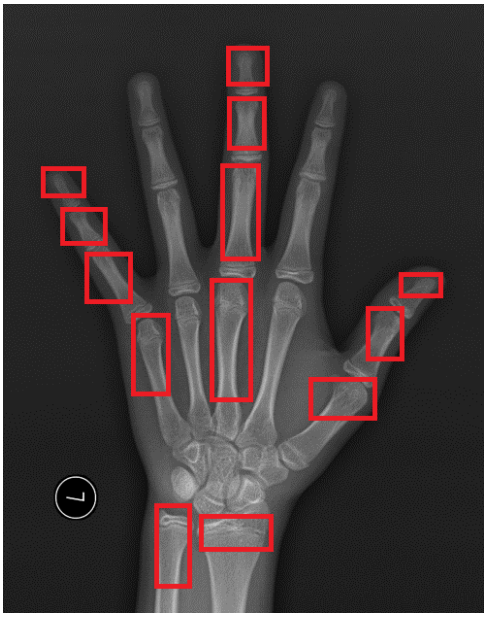

Figure 1. Some of the selected regions of interest (ROI) in specific bones of the wrist and hand used in TW method.

In this paper, we propose a multi-scale residual neural network (termed TJ-Net in this paper) for radiograph-based bone age prediction, introducing the attention mechanism to highlight the features important to the age group classification. Our earlier research reported that the features for bone age assessment could deliver a meaningful sex classification [6]. In this work, we extend this finding to a joint sex and bone age learning framework. The extensive experiments confirm that the learned features are in line with the traditional atlas-based assessment and transferrable to other datasets.

1. Related Work

The rich representation learning capabilities make the convolutional neural networks (CNN) and their variations one of the most successful deep learning tools in computer vision tasks [7, 8]. With CNN's ability to learn semantics-level features, [9] uses CNN to match the GP style atlas, and [10] uses CNN as a feature extractor for support vector machines to finish the classification of bone-ages. Most deep bone age assessments directly take the existing CNN models such as VGG16, Inception V3, and ResNet50 as the backbone, followed by a couple of dense layers to conduct regression or classification ([11], [12]}, and [13]). In [14], the authors take the pretained ImageNet and further fine-tune the classifier, generating an attention map similar to the ones used by human experts. [15] examines several bone age assessment DNNs consisting of convolutional and regressional layers. To focus on the specific regions in a radiograph that are believed to be critical, in the preprocessing stage, people either manually or automatically detect these regions in the hand skeleton ([16], [17], and [18]). Also, data augmentation can help alleviate the overfitting of some models, thus many proposals take random flips, crops, and contrast adjustment for image preprocessing [11]. [12] proposes an ensemble to integrate the estimates by three hand sections. The large dataset provided by RSNA laid the foundation for further improvement of applying DNNs in bone age assessment [5]. In addition to the pixels in radiographs, sex information input proved to help improve bone age assessment [5]. Furthermore, [19] explores the correlations between the models submitted to the RSNA challenge and achieves improved performance by combining the less-correlated ones into an ensemble.

2. Proposed Method

The proposed bone age classification network, called TJ-Net in this letter, consists of several functional blocks. Fig. 2 outlines the two-pronged architecture of TJ-Net. It maps an input image into one of the 77 bone age classes (corresponding to 0-19 years with a 3-months basic unit), as well as a binary gender label (the gender classifier, i.e. block 5, will be discarded in the test stage). The first 4 cascaded blocks in TJ-Net extract the crucial features from the input. Block 5 fuses the multi-scale features into the high-level ones for a dense layer to finish the sex classification. Block 6, the age classifier, matches the image features and the sex input into predefined age categories. The separate sex input supplies additional information for age classifier to adjust the result, and the sex classification helps to learn the bone features applying to both age estimation and sex labeling.

Each of the blocks 1-4 of the network comprises convolutional/pooling parts. The final output comes from a softmax function, and blocks 5 and 6 have fully connected layers. Fig. 2 also lists the network parameters along with the components.

We introduce the modules CBAM (Convolutional Block Attention Module), IncRes (Inception ResNet) into TJ-Net explores the attention mechanism, residual learning, and multi-scale features to empower the bone age-related feature extraction.

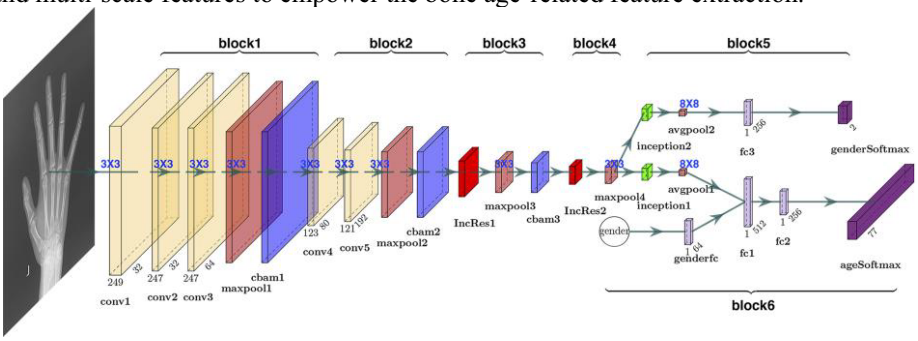

Figure 2. Architecture of TjNet.

2.1. Convolution-based Attention Modules

Among those channel and spatial features detected from the frontal convolutional blocks, the convolution-filtered features have varying importance in discriminating the bone age. The introduction of CBAMs is to enhance the features more relevant to the overall objective across the channel and spatial dimensions.

Following the practice in [20], each of three CBAMs in TJ-Net consists of two sequentially stacked components: channel attention and spatial attention modules. The input features first go through the average and maximum pooling layers in parallel, then two fully connected layers will produce a channel attention scale vector with sigmoid activation. The original input features weighted by the scale form a new tensor. This refined tensor subsequently walks through two parallel pooling layers the same as in the previous module (see Fig. 3). Then, the concatenated result gets processed by a $1 \times 7 \times 7$ convolution kernel. Finally, a sigmoid activation function outputs a spatial attention scale matrix. The output feature map is the multiplication of the scale matrix and the input features.

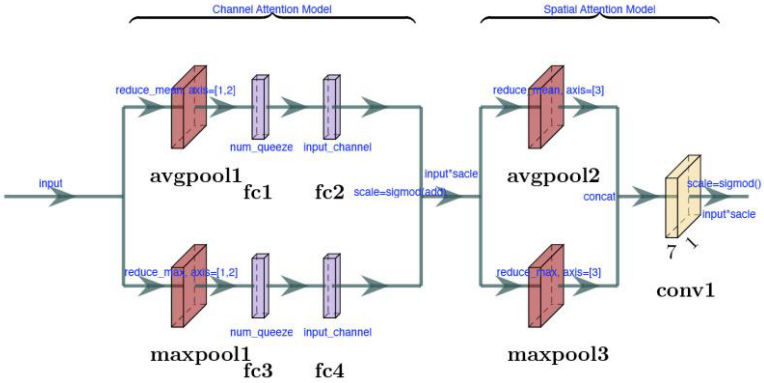

Figure 3. Internal structure of module CBAM1.

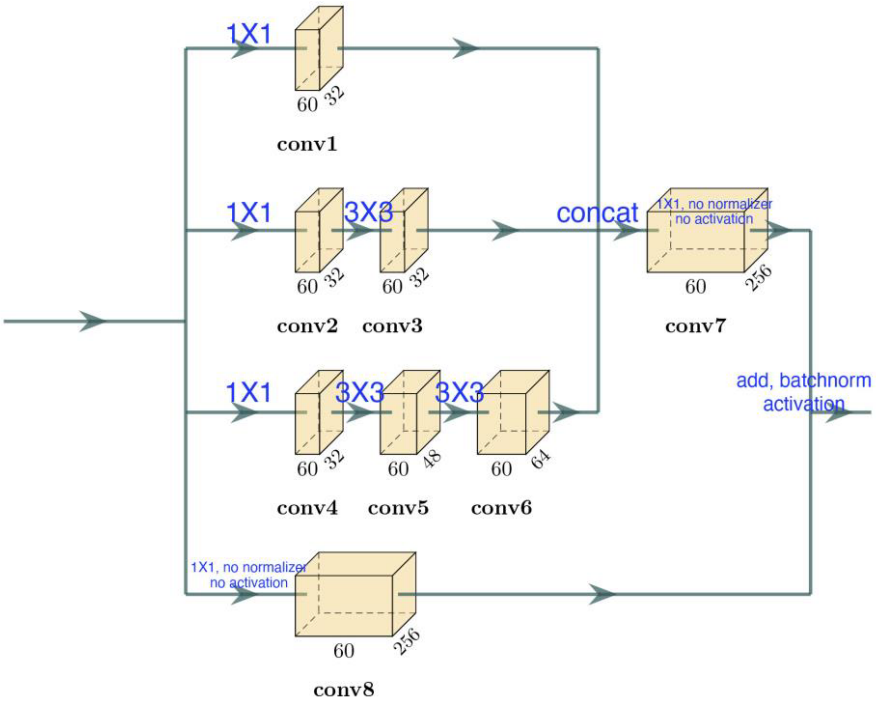

Figure 4. Internal structure of module IncRes1.

2.2. Inception Residual Modules

Local features in different scales may provide more and broader views to examine the relations between X-ray images and the bone ages.

There are two Inception Residual (IncRes) modules in TJ-Net, with slightly variant structures. Block IncRes1 consists of several branches with diverse convolutional kernel sizes (see Fig. 4). The multi-scale branches combine 1×1, and 3×3 convolution kernels in particular ways. Additionally, IncRes2 has four inception branches, namely 1×1, 3×3, 1×3, and 3×1 respectively. Similarly, each of the modules Inception1 and Inception2 has four multi-scale branches, the combinations of 1×1, 3×3, 1×3, 3×1, 5×5, 1×7, and 7×1. The convolutional outputs are concatenated in the later stage.

2.3. Loss Function for Joint Learning

The proposed loss function integrates the global classification errors over the training set, and the local sex/age batch-wise cross-entropies. Let n denote the size of dataset, g_i be the ground truth age label of sample i, and p_i be its predicted label. Let MAD be

$$e = \sum_{i=0}^{n} \frac{|g_i - p_i|}{n}.$$

We take an adaptive weight of $\alpha > 0$, to stress the importance of batch-wise errors in the early training phase. After the batch losses become stable, we shall let the global error to navigate the learning process by increasing α.

Denote the cross-entropy of two distributions p and q as

$$H(p,q) = \sum_{i=0}^{n} -p(x_i) \log q(x_i).$$

While at a certain age girls are roughly 2 years ahead of boys in bone development, it is not clear how the hand radiographs tell the difference between the two genders. However, gender labels can improve bone age assessment, as asserted in [5]. In other words, the conditional probability $P(age|gender, I)$ for a given X-ray image I helps accurately judge the age with the correct gender input. To better characterize this conditional probability, we shall learn the joint distribution $P(age, gender|I)$ from the training data by simultaneously minimizing the cross-entropies in a batch for both gender and bone age.

$$\mathcal{L} = \alpha e^2 + \sum_i (1 - \frac{|C_i|}{|C_t|}) H(Y_g, P_g) + \sum_j (1 - \frac{|C_j|}{|C_t|}) H(Y_a, P_a)$$

where Y_g, Y_a and P_g, P_a are the ground truth gender/age frequencies and the predicted ones in a batch, $|C_i|$ and $|C_j|$ are the batch-wise cardinality of gender class i and age class j, i.e. the number of samples labeled as the class i or j accordingly, and C_t is the size of a batch. Introducing the ratio weight is to level the influence of different classes within a batch.

3. Experiments

3.1. Datasets and Preprocessing

We evaluated the proposed TJ-Net on both the RSNA data and the much smaller Tongji dataset.

There are 14236 hand radiographs in the RSNA dataset, where 12611 are used for training, 1425 for validation, and 200 for test purposes. However, since the competition has been closed, the test data is no longer available. Hence, we partition the original training data into training, validation and test sets by the proportion 0.8:0.1:0.1, leaving 1261 samples for test. Tongji X-ray images were collected and manually analyzed by a group of experienced radiologists with Tongji Hospital, a major teaching hospital affiliated to Huazhong University of Science and Technology. This dataset contains the radiographs of the (mostly left) hands of subjects ranging from 0 to 22 years old. We masked the original X-ray images to remove the private information. Because the original Tongji annotations were in years, we modified the network to output class labels in years.

In the raw X-ray images, hands might orient to arbitrary directions. We rotated these images to make the hands' orientation consistently upwards. After giving up the radiographs with poor quality, we kept 1385 images for the experiments. Tongji data distributes almost evenly in gender distribution, with 768 samples for female, and 617 for male.

Because RSNA data and Tongji data had different resolutions, they were both scaled down to the 500×500 pixels in size. We took this size to balance two factors: first, small images would leave out the necessary detailed for radiologists to make an informed decision; second, high-resolution images might slow down the learning process with the limited computing resources in our lab. Also, regular data

augmentations were exploited in the online stage, including random crops, brightness changes, contrast variations, and flips.

Table 1. Comparison of Model Test Performances (in years) on RSNA Dataset.

Method	MAD	RMSE
VGG16	0.53	0.62
Inception V3[5]	0.57	0.58
Inception V4	0.59	0.97
ResNet50	0.60	0.72
TJ-Net	**0.41**	0.42

3.2. Implementation

We developed TJ-Net and other models in Python 3.6 with TensorFlow 1.7 framework. The hardware was a desktop computer equipped with one NVIDIA 2080Ti GPU card. All the models were trained using ADAM optimizer, with the batch size 16, and the initial learning rate 1^{-3}. When the training was close to being stabilized, we reduced the learning rate to 1/1000 of its initial value. It took about 15 hours for the training on the RSNA dataset, and 5 hours for fine-tuning with the Tongji data. To compare TJ-Net with other methods, we also developed the models using different backbones (e.g. VGG16, ResNet50, and Inception V4) with the separate sex input module and dense layers for classification [5]. These models were trained and tested on RSNA data. We followed the same protocol in all experiments: randomly choosing 80% of the data for training, leaving 10% for validation, and the rest 10% for test. To make the fair comparison, we tested Inception V3 with the RSNA data on the online server provided by 16bit, the top winner of the RSNA challenge (https://www.16bit.ai/bone-age).

In TJ-Net, the first two convolutional blocks feature small 3×3 kernels, to reduce the information loss of the images. In general, the features extracted from the first layers mainly capture the low-level vision details. Multi-scale deployment can extract multiple structural characteristics, and the residual link helps train the network. In the binary sex classification in block 5, we take only one fully connected layer of 256 neurons, to predict sex using the standard softmax function. The additional sex input block uses 64 neurons. Block 6, the age classification, has two dense layers, with 512 and 256 nodes respectively, followed by a normalized softmax function. Parameter α was set to 1.5 in the beginning, then boosted to 77 (the number of classes) after the loss became stable.

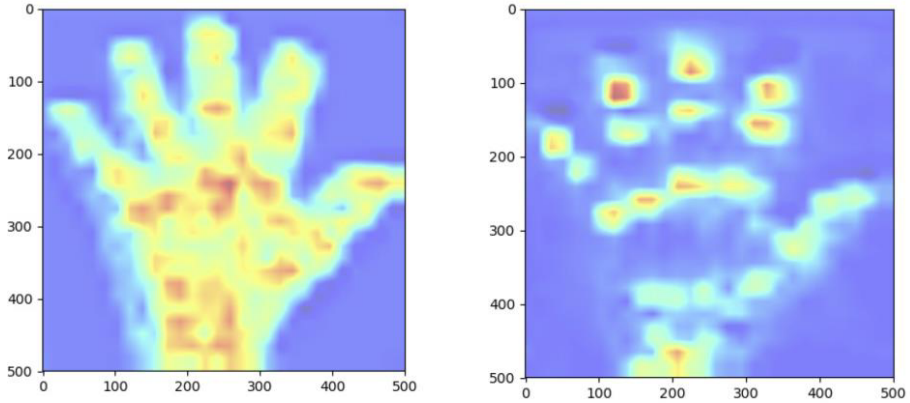

Figure 5. Comparison of the activation maps of block 4 learned without CBAMs (left) and with CBAMs (right), for a boy with the bone age labeled as 15 years old.

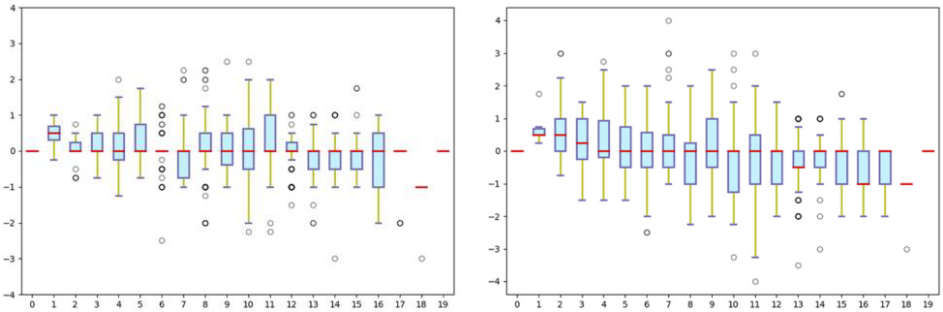

Figure 6. Comparison of prediction error distributions over age groups (in years) of RSNA data with (left) and without (right) joint sex learning.

3.3. Results and Analysis

Table 1 lists the comparison results of different models. We first trained TJ-Net with RSNA data and preserved the parameters of some blocks while fine-tuning the rest of the model with the Tongji training set. The MAD on the test set with blocks 1-3 frozen achieved the best outcome of 0.36 years (Table 2). It showed that our model TJ-Net was capable of capturing the critical features in assessing radiographic bone age and these low-level features applicable to different ethnic groups in distant regions.

Fig. 6 indicates that the predictions for both RSNA and Tongji data were basically unbiased. The errors for the two datasets were similar in distributions.

Table 2. Comparison of transfer learning results using different freezing strategies.

Pretrained Blocks	MAD	RMSE	Training Accuracy	Test Accuracy
Block 1	0.688	1.645	**0.996**	0.797
Blocks 1, 2	0.565	0.986	0.889	0.875

Pretrained Blocks	MAD	RMSE	Training Accuracy	Test Accuracy
Blocks 1-3	**0.362**	**0.551**	0.891	**0.922**
Blocks 1-4	0.638	1.072	0.879	0.883

To verify the effectiveness of different components in TJ-Net, we did the ablation analysis for the functional parts. We compared the resulting MAD and RMSE using RSNA dataset for the models, leaving one component out at each time. After eliminating the shortcuts in IncRes modules, the re-trained model obtained a MAD of 0.571 years. Similarly, without CBAMs, the simplified TJ-Net generated a MAD of 0.578 years, suggesting the positive role of attention modules. To visualize the impact of CBAMs on the trained model, we draw the heat maps of the activated features learned by block 4. In Fig. 5, the activation map with CBAMs pinpoints to the key locations similar to the ROIs examined by the TW method (Fig. 1), while the salient points found without attention mechanisms spread over a large area. Finally, if the gender classification and the associated loss were taken out, the MAD increased to 0.578 years. This indicates that joint learning improved the contribution of sex input in assessing bone age.

Figure 7. Comparison of the sex input weight matrices in block 6 without joint learning (top) and with joint learning (bottom), black represents 0.

The boxplots in Fig. 6 display the estimation error distributions for TJ-Net with and without block 5 and joint learning. Though the estimations are virtually unbiased, adding joint learning to TJ-Net allows more accurate predictions for most of the age classes. Moreover, in Fig.7 where the grey scales represent the strengths of the weights (black equals to 0), we see that joint learning makes the sex input weight matrix sparser, intensifying the influence pattern of sex input on bone age.

4. Conclusions

In this work, we proposed a specifically designed deep learning neural network, TJ-Net, for automatic radiographic bone age assessment. In TJ-Net, the attention modules helped find the features resembling the focal points explored by human experts, and the joint sex/age learning enhanced the predication of age conditional on sex labels. Experimental results demonstrated that the low-level features learned from the RSNA dataset could be transferred to the data acquired by a Chinese hospital from local subjects. With the MADs of 0.41 years and 0.36 years on RSNA and Tongji data, the proposed model performed better than other single model state-of-the-art methods.

References

[1] M. Satoh, Bone age: assessment methods and clinical applications, *Clin Pediatr Endocrinol*, vol. 24, no. 4, pp. 143--152, 2015.

[2] R.K. Bull, P.D. Edwards, P.M. Kemp, et al, Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods, *Arch Dis Child*, 1999, vol. 81, pp. 172-172.

[3] D.Shen, G. Wu, and H.-I. Suk, Deep Learning in Medical Image Analysis, *Annual Review of Biomedical Engineering*, 2017, Vol. 19, pp. 221-248.

[4] G. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, A survey on deep learning in medical image analysis, *Medical Image Analysis*, 2017, Vol. 42, pp. 60-88.

[5] S.S. Halabi, L.M. Prevedello, Jayashree Kalpathy-Cramer, et al, The RSNA Pediatric Bone Age Machine Learning Challenge, *Radiology*, Jan. 2019, Vol. 290, no. 2, pp. 1-6.

[6] W. Tang, G. Wu, G. Shen, Improved Automatic Radiographic Bone Age Prediction with Deep Transfer Learning, In: *CISP-BMEI 19*, 2019, DOI: 10.1109/CISP-BMEI48845.2019.8965906.

[7] P. He, X. Jiang, T. Sun, and H. Li, Computer Graphics Identification Combining Convolutional and Recurrent Neural Networks, *IEEE Signal Processing Letters*, 2018, Vol. 25, no. 9, pp. 1369-73.

[8] J. Hu, Z. Chen, M. Yang, et al, A Multiscale Fusion Convolutional Neural Network for Plant Leaf Recognition , *IEEE Signal Processing Letters*, 2018, Vol. 25, No. 6, pp. 853--357.

[9] J.R. Kim, W.H. Shim, H.M. Yoon, S.H. Hong, et al, Computerized Bone Age Estimation Using Deep Learning - Based Program: Evaluation of the Accuracy and Efficiency, *American Journal of Roentgenology*, December, 2017, Vol. 209, pp. 1374-1380.

[10] X. Chen, J. Li, Y. Zhang, et al, Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age, *Future Generation Computer Systems*, 2019. https://doi.org/10.1016/j.future.2019.10.032

[11] D.B. Larson, M.C. Chen, M.P. Lungren, et al, Performance of a Deep-learning neural network Model in assessing skeletal Maturity on Pediatric hand radiographs, *Radiology*, 2018, Vol. 287, No. 1, pp. 313-322.

[12] V. Iglovikov, A. Rakhlin, A. Kalinin and A. Shvets, Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks, *arXiv:1712.05053* [cs.CV], 2017.

[13] T. Van Steenkiste, J. Ruyssinck, O. Janssens, et al, Automated Assessment of Bone Age Using Deep Learning and Gaussian Process Regression, in: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018, DOI: 10.1109/EMBC.2018.8512334.

[14] H. Lee, S. Tajmir, J. Lee, et al, Fully Automated Deep Learning Systemfor Bone Age Assessment, *J Digit Imaging*, 2017, Vol. 30, pp. 427-441.

[15] C. Spampinato, S. Palazzo, D. Giordano, et al, Deep learning for automated skeletal bone age assessment in X-ray images, *Medical Image Analysis*, 2017, Vol. 36, pp. 41--51.

[16] J.H. Lee, K.G. Kim, Applying Deep Learning in Medical Images: The Case of Bone Age Estimation, *Healthcare Informatics Research*, January 2018, Vol. 24, No. 1, pp. 86-92.

[17] S. Lee, M. Choi, H.-s. Choi, et al, FingerNet: Deep Learning-Based Robust Finger Joint Detection from Radiographs, in: *IEEE Biomedical Circuits and Systems Conference: Engineering for Healthy Minds and Able Bodies*, 2015, DOI: 10.1109/BioCAS.2015.7348440.

[18] M. Escobar, C. Gonzalez, F. Torres, et al, Hand Pose Estimation for Pediatric Bone Age Assessment, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science, vol 11769*, 2019, pp. 531-539.

[19] I. Pan, H.H. Thodberg, S.S. Halabi, et al, Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge, *Radiology: Artificial Intelligence*, 2019, Vol. 1, No. 6, pp. 1-9.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, CBAM: Convolutional Block Attention Module, in: *ECCV 2018, Lecture Notes in Computer Science, vol 11211*, *arXiv:1807.06521 [cs.CV]* , 2018.