Performance of Convolutional Neural Networks and Transfer Learning for Skeletal Bone Age Assessment

Chadaporn Keatmanee, PhD[1]; Songphon Klabwong, MSc[2]; Kamolphong Osatavanichvong, MD[3]; Chirotchana Suchato, MD[4]

Chadaporn Keatmanee, PhD

Abstract

OBJECTIVES: Bone age assessment is used by clinicians for estimating the maturity of a child's skeletal system. Traditionally, physicians use template matching methods (GP and/or TW2). Time and accuracy of the evaluation rely on a physician's experience. Therefore, this research proposes a fully automatic system for bone age assessment with cutting edge artificial Intelligence (AI) technology.

MATERIAL AND METHODS: Convolutional Neural Network (CNNs), a Deep Learning (DL) technique is applied to skeletal bone age prediction combined with transfer learning algorithm. Hence, various kinds of transfer learning algorithms (ResNet-50, Inception-V3, and VGG-16) are investigated in training in the proposed model fed by a number of x-ray images (12,000 image approximately—imbalanced data).

RESULT: VGG-16 shows significant accuracy compared to ResNet-50 and Inception-V3 (*mae* = 6.53, 20.52 and 43.11 months respectively)

CONCLUSION: The most effective pre-trained layer for CNNs in bone age assessment is VGG-16 according to the accuracy of its prediction.

Keywords: deep learning, convolutional neural network, bone age, growth disorder, maturity estimation, transfer learning

The skeletal bone development during an organism's changes in shape and size show a difference between chronological ages and a child's assigned bone (bone ages). Therefore, physicians use a bone age assessment to estimate the maturity of a child's skeletal system. The evaluation might indicate a growth disorder, endocrine diseases, neuro diseases, and newborn malnutrition. Primarily, the evaluation methods start with taking an x-ray image of the left hand covering bones from wrist to fingertips. Later, the bones on the x-ray image are compared with radiographs in a standardized atlas of bone development collected from children of the same sex and age, ranging from 0-228 months.

Generally, bone age assessment has been performed manually over the past decades using either Greulich and Pyle (GP)[1] or Tanner-Whitehouse (TW2)[2] methods. In both cases, the evaluation requires considerable time and its accuracy may have to rely on a clinician's experience. Therefore, a fully automatic bone age assessment system is strongly recommended. It would not replace the physicians but rather it would support their decision with AI technology.

In medical image processing, among various techniques in AI, Machine Learning (ML) is important[3]. Apart from ML, Deep Learning (DL) is one of the cutting edge technologies that applies ML to large data, which is a dominant approach for medical imaging, especially, applied in segmentation[4] and classification[5]. Therefore, this research aims to develop a bone age assessment system that applies a DL based method, convolutional neural networks (CNNs). Besides, the designing of the CNNs model architecture for bone age prediction, this research intends to evaluate the performance of various pre-trained models including ResNet-50, Inception-V3, and VGG-16 in order to create an appropriate design of pre-trained layer for CNNs in bone age prediction.

[1] Digital engineering, Faculty of Engineering, Thai-Nichi Institute of Technology, Bangkok, Thailand.
[2] Divertise Asia Co.Ltd., Bangkok, Thailand.
[3] Urupong Medical Center, Bangkok, Thailand.
[4] Radiology Department, Bangkok Hospital Headquarters, Bangkok, Thailnad.

**Address Correspondence to author:*
Chadaporn Keatmanee, PhD
Digital engineering, Faculty of Engineering,
Thai-Nichi Institute of Technology,
1771/1 Pattanakarn Road, Suanluang,
Bangkok, 10250, Thailand.
email:chadaporn@tni.ac.th

Received: January 11, 2019
Revision received: January 13, 2019
Accepted after revision: February 11, 2019
BKK Med J 2019;15(1): 1-6.
DOI: 10.31524/bkkmedj.2019.02.001
www.bangkokmedjournal.com

Background

The convolutional neural network (CNNs or ConvNets) is one of the most effective algorithms for image classifications[6] including x-ray images for bone age assessment. There are numerous researches that have been conducted for bone age prediction. CNNs, for example, were applied by Tom Van Steenkiste and others[7] to evaluate the effectiveness of data augmentation in CNNs. Not only the performance of various methods applied in CNNs[8] were investigated but also the different architectures of CNNs[9,10] were examined. Moreover, successful CNNs were compared to results reached by humans,[9-11] and they showed promising results.

That said, the CNNs model generally performs well if it is given balanced data. The dataset would have almost the same number of images in each category in order to train the model. Imbalanced data can impede generalization and this may cause the model to make grave mistakes after training. Therefore, solving an imbalanced dataset is mandatory and this can be achieved by resampling techniques[12]; oversampling and undersampling.

CNNs take an input x-ray image, process it and classify it under certain categories (the bone age, 0-240 months for this paper). CNN has two main parts including feature learning (Convolution block(s)) and classification (fully connected layer). CNNs works as an image recognition by transforming the x-ray image through layers to a class score as shown in Figure 1.

Besides the modification of CNNs' architecture for improving the model's performance, transfer learning[13,14] is widely utilized in CNNs. It could improve accuracy of CNNs in a timesaving way because transfer learning is built as a pre-trained model. The model was trained on a large benchmark dataset (a variety of images) to solve a problem similar to itself (the image classification in this paper)

There are numerous pre-trained models that have been used in CNNs, however, the investigation of transfer learning algorithms in this research focuses on their characteristics, widely applied in image classification, which are ResNet-50, Inception-V3, and VGG-16.

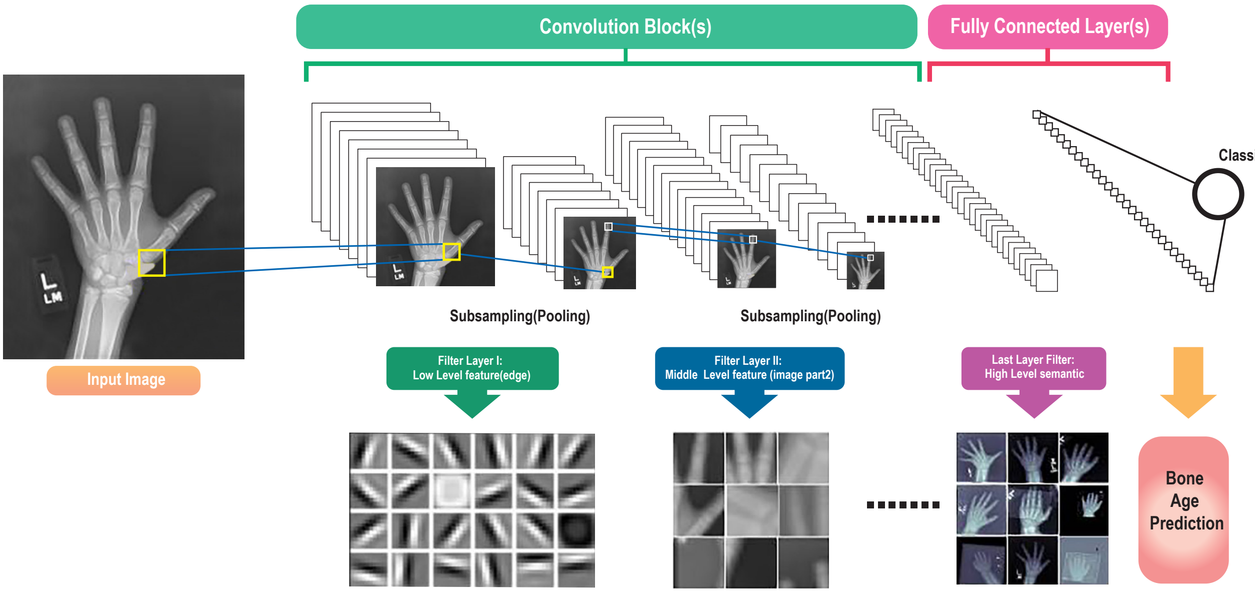

Figure 1: The proposed architecture of CNNs model for bone age assessment.

- **ResNet-50:** The pre-trained model is built from training on more than a million images from the ImageNet database.[15] It has 50 layers and can classify images into 1,000 categories.
- **Inception-V3:** The model is trained on a million data from 2012 for ImageNet Large Visual Recognition Challenge. It is 42-layer deep and can be categorized into 1,000 classes.[16]
- **VGG-16:** The model is also trained on ImageNet with 16 layers deep having 1000 outputs for 1000 classes.[17]

Materials and Methods

Materials

- **Tools:** The online Graphical Processing Unit, Google Colab (K80 GPU) is used for training and testing the proposed CNN model.
- **Library:** The proposed architecture of the CNNs model is developed and verified with DL open source libraries based on python programming language including: Keras 2.2.4 and Tensorflow 1.12.0.

- **Dataset:** The online access dataset is provided by the Center of AI in Medicine & Imaging (Stanford University).[18] They are x-ray images containing 12,611 images in total with two labels including bone age and gender.

Method

The procedure for designing bone age prediction model based CNNs are explained in detail as follows:

1. Inspect and verify image dataset. This is conducted primarily to see data distribution including age range (0-240) and gender shown in Figure 2.
2. Resampling data to equate data distribution. This is done to avoid an imbalanced dataset that would cause ineffective learning of the CNNs model. The training dataset is divided into 20 non-overlapped classes (10 age categories × 2 genders) AG_{train} as follows in Table 1.

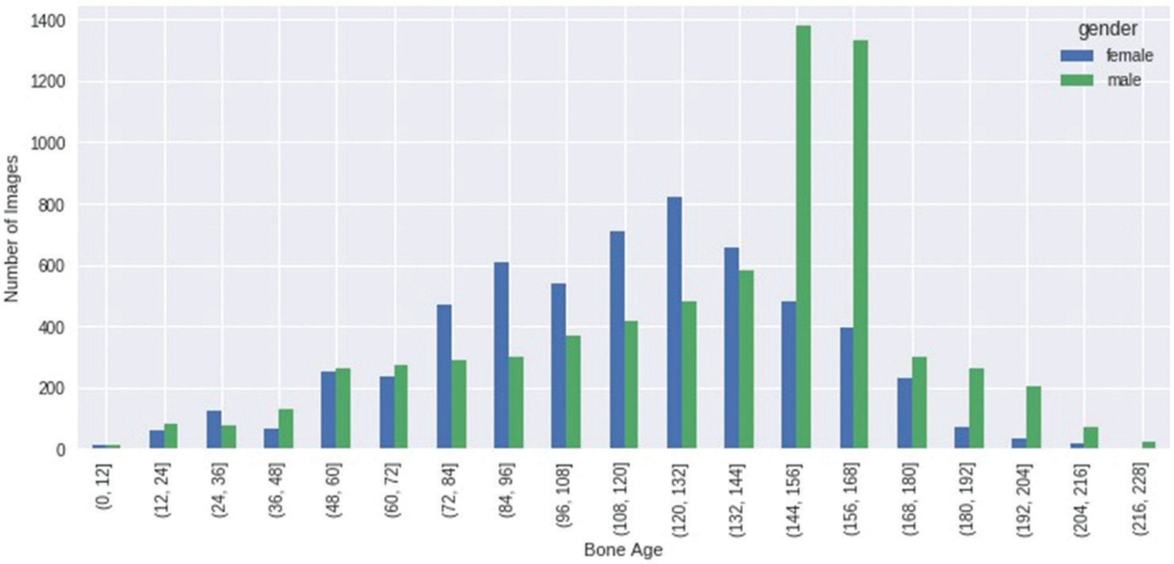

Figure 2: The X-Ray images distribution for bone age (0-240 months) of male (green) and female (blue).

Table 1: The 20 non-overlapped classes for the resampling.

Bone age category	Gender	Count
(0.773, 23.7]	Female	29
	Male	43
(23.7, 46.4]	Female	175
	Male	147
(46.4, 69.1]	Female	363
	Male	296
(69.1, 91.8]	Female	453
	Male	441
(91.8, 114.5]	Female	911
	Male	535
(114.5, 137.2]	Female	1161
	Male	698
(137.2, 159.9]	Female	915
	Male	1564
(159.9, 182.6]	Female	469
	Male	1261
(182.6, 205.3]	Female	75
	Male	378
(205.3, 228.0]	Female	13
	Male	73

The target dataset size of each class is set to $n = 1{,}500$. We perform oversampling on all categories except in the age range (137.2, 159.9] months of male images, which has the original sample size of 1,564 images. We perform undersampling on this particular class to match sample size of n. With the oversampling method, we ensure that the resampled dataset is a strict superset of original dataset by following Algorithm 1.

Algorithm 1: let $r \in AG_{train}$ be the group of age (range in months) and gender, and D (r) be a function return a set of image in group r

```
1. Set $n = 1{,}500$
2. For each age and gender group, count total number x of D($r$)
3. If $x < n$:
      a. Include $D(r)$ into D′(r)
      b. Randomly select image $I \in D(r)$
         for $n - x$ images
4. Else:
      a. Randomly select image $I \in D(r)$
         for $n$ images
```

3. Image Data Augmentation is applied to the dataset not only to enlarge the dataset but also to replace redundant images created in the resampling process. Despite the inherent translation invariance of CNN,[19,20] Image Data Augmentation can help improve scaling and the rotation invariant. The augmented images are performed by utilizing image translation, rotation, scaling techniques as shown in Figure 3. It randomly performed image translation on the x and y axis in the range of up to 5%, scaling up to 2% and rotation of up to 10 degrees in respect of the original image.

4. The proposed model based CNNs is designed and developed by applying a pretrained model and attention mechanism as shown in Figure 4. This paper focuses on evaluating well-known pretrained models, which are ResNet-50, Inception-V3, and VGG-16. The other parts such as model hyper- parameters adjustment, attention mechanism, and the comparison of the proposed model with prior methods are recommended studies in the future.

5. Model training is conducted several times with different pretrained layers as shown in Table 2. Every pre-trained model (VGG-16, ResNet-50, and Inception-V3) is evaluated under the same conditions as input- bone age x-ray image and gender, the hyper-parameters, as well as the number of test datasets.

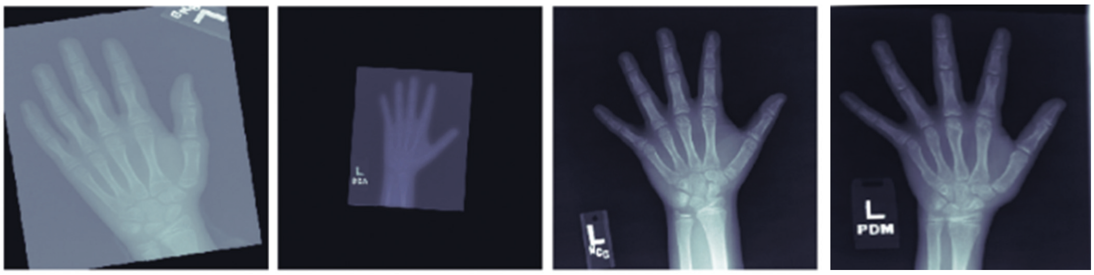

Figure 3: Examples of image augmentation including rotation, scaling, and translation.

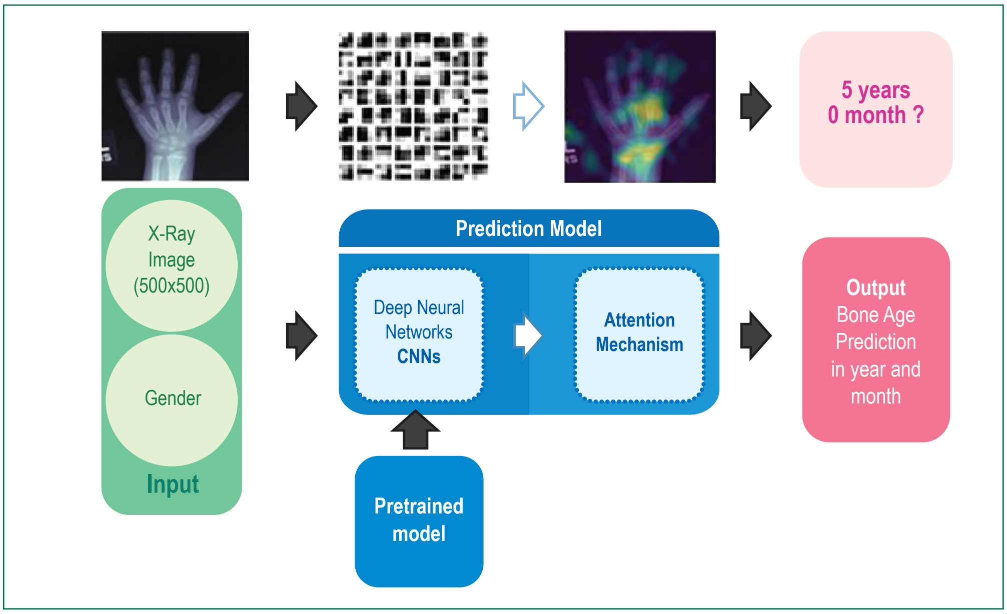

Figure 4: The proposed bone age assessment model with CNNs utilizing transfer learning, as well as, attention mechanism.

Table 2: The evaluation of different pretrained models

Architect	Input	Image size	Age range	Training size (Original/Augmented)	Step / Epoch	#Epoch	Test size	mae (months)	Time (s)
VGG-16	Gender ImageAugmented	500 x 500	0 - 228	10,000/30,000	3750	20	1,000	6.53	60.70
ResNet-50	Gender ImageAugmented	500 x 500	0 - 228	10,000/30,000	3750	20	1,000	20.52	50.07
Inception-V3	Gender ImageAugmented	500 x 500	0 - 228	10,000/30,000	3750	20	1,000	43.11	34.97

6. The model performance evaluation is examined using a fixed test dataset of 1,000 images. There are two criteria for testing the proposed model:
   - Mean Absolute Error (*mae*) is used for evaluating the model / month. We define *mae* over test dataset as:

$$mae = \frac{\sum_{i=1}^{n} |y_i - p(x_i, g_i)|}{n}$$

Where

n is total number of testing data

y_i is true bone age

$p(x_i, g_i)$ is a predicted bone age on image x_i and g_i gender.

   - Prediction time in second is measured averagely (3-trial).

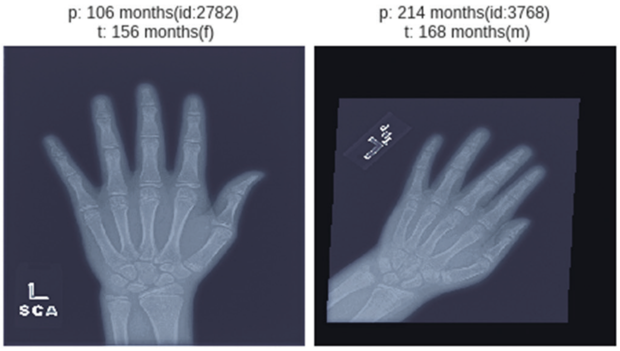

Figure 5: Example of the low quality images.

Results and Discussion

The results suggest VGG-16 as the best pre-trained model for the proposed CNNs model after 20 epochs training of 30,000 augmented image dataset. The VGG-16 yields *mae* of 6.53 months on the test set is shown in Figure 5. The figure depicts a clear correlation between the predicted age and the bone age dataset. However, some outliers make prediction results differ to the actual bone age by a large margin as shown in Figure 5. The variation is due to the low quality images in the dataset shown in Figure 6.

The results also demonstrate that prediction accuracy is higher around the middle age range, when we have more image data. This suggests that a large set of original images impact on model accuracy more than augmented images data. In the pre-trained ResNet-50 converses at epoch 20, the *mae* is 20.52 months. Whereas, Inception-V3 shows the *mae* result of 43.11 months. The prediction results for RestNet-50 and Inception-V3 are shown in Figure 7 and 8 respectively.

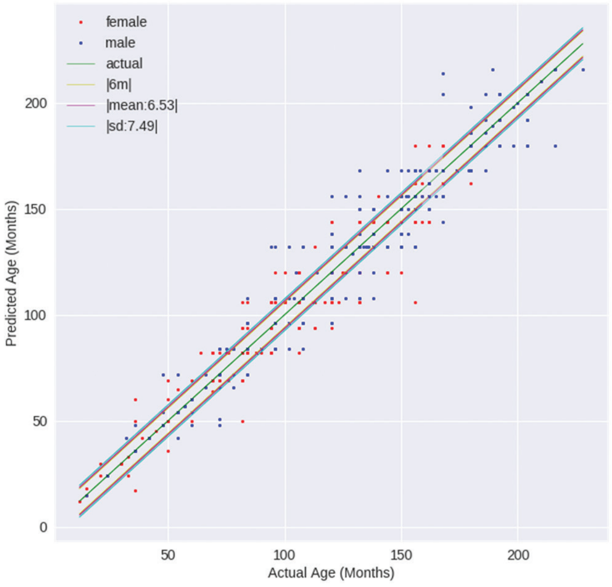

Figure 6: Evaluation of VGG-16.

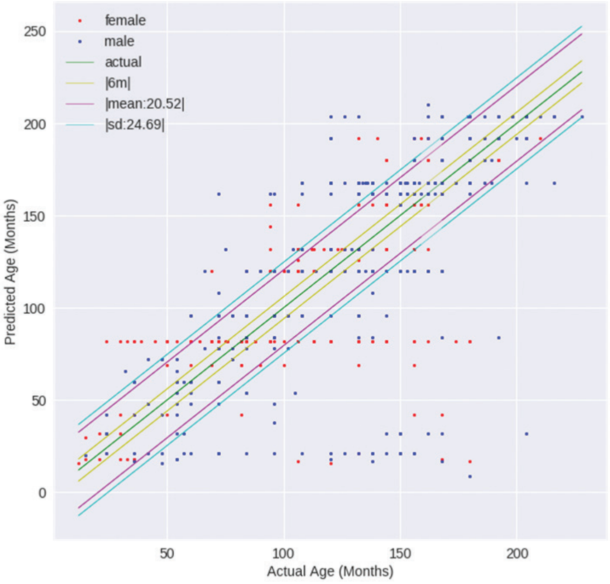

Figure 7: Evaluation of ResNet-50.

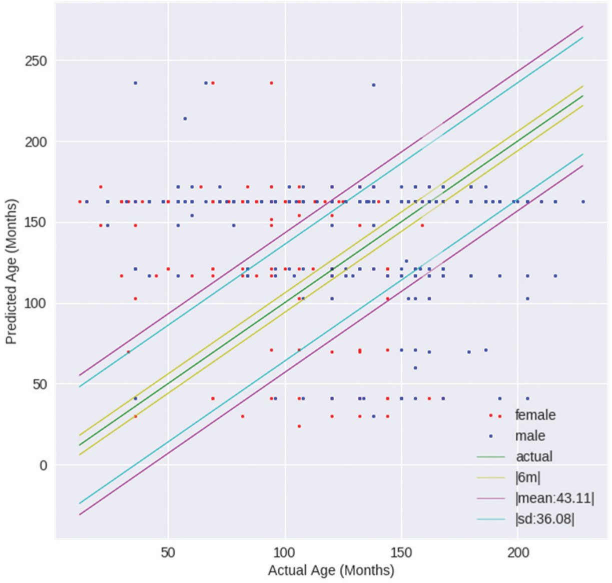

Figure 8: Evaluation of Inception-V3.

Despite the performance of the evaluation, InceptionV3 performs well during the training period. The large mae on the test dataset suggests that the model is overfitting. There are several ways for model improvement such as increasing the image augmentation and training on more epochs. The time evaluation is set out in Table 2, and the forecasting time for predicting 1,000 images, is the average number calculated from 3 trials. The Inception-V3 shows the fastest rate of prediction when compared to the speed of ResNet-50 and VGG-16, at 34.97, 50.07, and 60.70 respectively.

Although VGG-16 spends a considerable amount of time to make a prediction from 1,000 images, the accuracy rate (*mae*) of the model performance is significantly lower than RestNet-50 and Inception-V3. Therefore, VGG-16 is widely recommended to be applied as a pre-trained layer in CNNs for the proposed bone age assessment model in order to improve the model accuracy drawing on prior knowledge from the transfer learning.

Conclusion & Future work

Bone age assessment aims to examine a child's skeletal system bone age compared to a chronological age. Generally, clinicians perform a manual examination that is quite time consuming and the quality of evaluation is based on individual knowledge and skills. Therefore, we propose that an automatic bone age assessment system is created based on DL technique using CNNs. The contribution of this paper concentrates on evaluating various well-known pre-trained models including VGG-16, ResNet-50, and Inception-V3 under the same environment while training. The results of the evaluations indicated that VGG-16 could improve model accuracy significantly (*mae* = 6.53 months) whereas *mae* of ResNet-50 and Inception-V3 are 20.52 and 43.11 respectively.

However, the performance of the proposed bone age assessment system still suffered from low quality image, imbalanced dataset, as well as complex hyper-parameter adjustment. Hence, a deeper investigation for designing the model is strongly required. In addition, an examination of the age range for designing a model based on growing rate related to gender and attention mechanism will be performed in future work. This will be combined with an evaluation of the proposed bone age assessment system with a prior successful bone age prediction platform based on Deep Leaning technique.

References

1. Greulich, W. W., Pyle, S. I., and Todd, T. W. Radiographic atlas of skeletal development of the hand and wrist. Stanford: *Stanford university press*, 1959;2:150-159.
2. Tanner, J. M., Whitehouse, R. H., Cameron, N., Marshall, et al. Assessment of skeletal maturity and prediction of adult height (TW2 method). *London: Academic press*, 1975;16.
3. Wang S., Summers RM. Machine Learning and radiology. *Med Image Anal*, 2012;16:933-51.
4. Fausto M., Seyed-Ahmad A., Christine K., et al. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding,* 2017;164:92-102
5. Luis H.S. V., Rodrigo M.S. V., Flavio. H.D., et al. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. *Engineering Applications of Artificial Intelligence,* 2018;72:415-422
6. Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012:1097-1105.
7. Van Steenkiste T, Ruyssinck J, Janssens O, et al. Automated assessment of bone age using deep learning and Gaussian process regression. *IEEE Engineering in Medicine and Biology Society Conference Proceedings*.2018:674–7.
8. Yagang W., Qianni Z., Jungang H., and Yang J. Application of Deep learning in Bone age assessment. IOP Conference Series: *Earth and Environmental Science.* 2018;199(3):032012.
9. Matthew C., David B. L., Matthew P. L., et al. Deep Neural Nets: Pediatric Hand Radiographs. *Radiology informatics* [online]. 2019 [cited 2019 Jan 1]. Available from:http://langlotzlab.stanford.edu/projects/pediatric-hand-radiographs/
10. Matthew C. Automated Bone Age Classification with Deep Neural Networks [online]. 2019 [cited 2019 Jan 1]. Available from: http://cs231n.stanford.edu/reports/2016/pdfs/310_Report.pdf
11. Larson DB., Chen MC., Lungren MP., et al. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Prediatric Hand Radiographs. *Radiology.* 2018;287(1):313-322.
12. Kotsiantis S., Kanellopoulos D. and Pintelas, P. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering.* 2015;30:25-36.
13. Rawat, W. and Wang, Z. Deep Convolutional Neural Networks for Image Classification. A Comprehensive Review. *Neural computation.* 2017;29(9):2352-2449.
14. Sinno J. P. and Qiang Y. A Survery on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering [online]. [cited 2019 Jan 1]. Available from: https://www.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf
15. He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770-778.
16. Christian S., Vincent V., Sergey I., et al. Rethinking the Inception Architecture for Computer Vision. *The CVPR paper provided by the Computer Vision foundation* [online]. [cited 2019 Jan 1]. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf
17. Karen S. and Andrew Z. Very Deep Convolutional Networks for Large-Scale Image Recognition [online]. [cited 2019 Jan 1]. Available from: https://arxiv.org/abs/1409.1556
18. RSNA Pediatric Bone age challenge [online]. 2017 [cited 2017 Oct 7]. Available from: http://rsnachallenges.cloudapp.net/competitions/4
19. LeCun Y. Learning invariant feature hierarchies. Computer vision ECCV 2012. Workshops and demonstrations. Springer Berlin Heidelberg, 2012.
20. Jaderberg M., Karen S., and Andrew Z. Spatial transformer networks. *Advances in Neural Information Processing Systems.* 2015.