Biomedical Data Science
2024, VOL. 4, NO. 4, 36-43
DOI: 10.47297/wspbdsWSP2752-630506.20240404

Age assessment from unlabeled hand bone X-ray images based on Xception

Minghao Liang, Hang Lu*

Data Science and Big Data Technology, Guangdong Medical University, Dongguan, 523000, China

ABSTRACT
Bone age assessment (BAA) through hand X-rays plays a vital role in diagnosing endocrine disorders in adolescents and assessing children's growth and development. Traditionally, clinical BAA has been a subjective estimation, with its accuracy largely depending on the expertise of pediatric specialists. Recently, several deep learning techniques have been introduced to automate bone age estimation, demonstrating encouraging outcomes. However, these approaches often lack the ability to utilize discriminative information adequately due to the absence of manual annotations for key bone regions, which are critical indicators of skeletal maturity, potentially limiting their practical effectiveness. In this research, we introduce a deep learning-based BAA approach that eliminates the need for manual region annotations. The proposed model leverages the Xception deep convolutional neural network, enhanced with the CBAM attention module, to estimate bone age from unlabeled hand X-ray images. Experiments conducted on the RSNA public dataset indicate that our method achieves a mean absolute error (MAE) of 7.93 months.

KEYWORDS
Bone Age Assessment (BAA); Deep Learning; CBAM Attention Module; Xception Network; RSNA

1 Introduction

Bone age assessment (BAA) based on hand X-rays usually refers to a method of assessing bone age in months by obtaining X-ray images of the child's left hand (non-dominant hand) through pediatric radiology[1]. The ossification level of the hand bones often reveals the maturity of the body skeleton[2]. Clinical studies have shown that many physiological indicators are highly correlated with bone age, including height growth rate, metabolic level, menarche, muscle mass and bone density[3]. It has been reported that severe thyroid dysfunction, pituitary changes secondary to deformities, tumors or invasive lesions, hypogonadism, malnutrition or some mental illnesses may affect bone age development so that the test value does not match the actual age. These indicate that bone age is important for the clinical diagnosis of adolescent endocrine disorders and the estimation of children's growth and development[4].

Traditional approaches for bone age assessment (BAA) primarily include the Greulich-Pyle (G&P) method and the Tanner-Whitehouse (TW) method[1]. The G&P technique requires comparison of an X-ray with a reference atlas containing annotated bone ages. The closest match between the X-ray and the atlas determines the estimated bone age. In contrast, the TW approach is more detailed, focusing on specific regions of interest (ROIs) like the radius, ulna, carpal bones, metacarpal bones, and phalanges (illustrated in Figure 1). Instead of analyzing the entire X-ray, this method assigns scores to individual ROIs, which are then averaged to calculate the final bone age. Due to the significance of anatomical features, a numerical scoring system is utilized to assess each ROI. However, these traditional assessments depend heavily on skilled radiologists or pediatricians for manual evaluation and annotation. This process is not only time-consuming and labor-intensive but also prone to inconsistencies stemming from variability among different practitioners. As a result, automated BAA techniques are increasingly being explored to address these limitations.

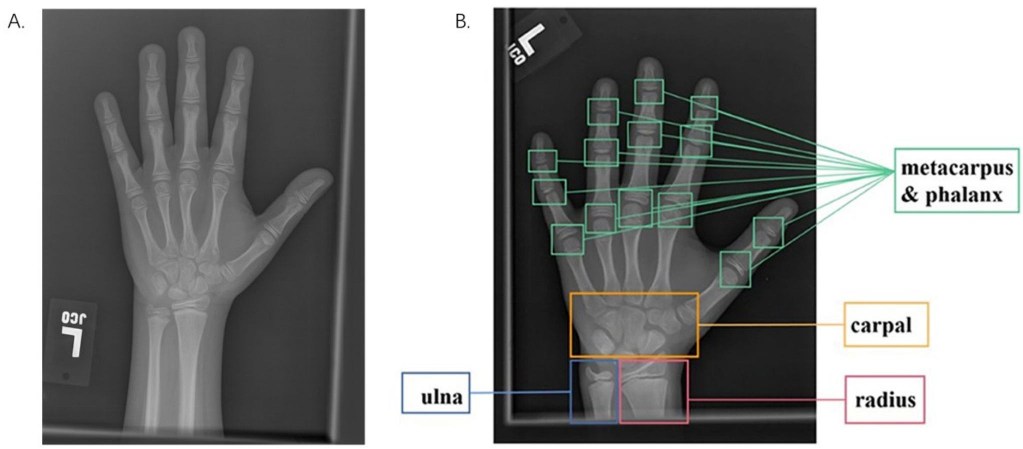

Figure 1 Left hand X-ray image and left hand X-ray image with annotations.(A. Left hand X-ray image. B. Left hand X-ray image with annotations using ROI evaluation and image transformation processing. C.)

The integration of deep learning and computer vision has revolutionized medical imaging, leading to significant advancements in automated bone age estimation. Neural networks, especially convolutional neural networks (CNNs), have shown great promise in automating the analysis of skeletal maturity from X-rays. Traditional manual assessments, while effective, are labor-intensive and prone to subjective errors. In contrast, deep learning algorithms offer improved accuracy and efficiency in bone age estimation. Research has shown that CNNs are highly effective in processing images and extracting critical features, often outperforming conventional methods [5]. Despite these advances, challenges remain in achieving precise automated assessments due to inherent errors in the algorithms.

Regions of interest (ROIs) significantly enhance bone age assessment by enabling a concentrated evaluation of specific bones, thus allowing the scoring of relevant areas while excluding non-essential details to improve accuracy. Automated bone age assessment (BAA) methods are generally categorized into two types: unsupervised approaches and supervised methods. In early unsupervised approaches, the entire hand X-ray image was processed directly using end-to-end models to estimate bone age. For instance, a deep residual network (ResNet50) incorporating the G&P mapping technique reported a mean absolute error (MAE) of 6.24 months when applied to the RSNA dataset[6]. This model not only quantified bone age but also highlighted the most responsive regions in each X-ray. Similarly, The work of Pan and colleagues (2020) employed a U-Net model to extract hand masks from raw X-rays, leveraging a deep active learning strategy to reduce the annotation workload. Using this approach, the mean absolute error (MAE) obtained on the RSNA dataset was 8.59 months[7]. and introduced an innovative active learning framework for segmenting hand images with minimal labeled data. However, annotation-free models often encounter difficulties in maintaining both accuracy and interpretability.

Conversely, annotation-based methods utilize images pre-processed with manually defined bounding boxes. These approaches rely on domain knowledge to target specific regions, leading to enhanced accuracy in bone age estimation. Iglovikov's research team (2017) applied an advanced U-Net framework to identify key points using manually labeled masks, achieving an MAE of 6.3 months for males and 6.49 months for females[8]. Escobar's group, in 2019, integrated additional bounding box annotations into their training data, resulting in improved performance with an error margin of 4.14 months when applied to the RSNA dataset[9]. Ren et al. (2018) implemented the Faster R-CNN framework to detect hand regions and utilized the Inception-V3 model to estimate age, reaching an MAE of 5.2 months[10]. Additionally, Son et al. (2019) focused on 13 individual bones using the TW3 method, resulting in a mean prediction error of 5.52 months when used for the RSNA dataset[11].

The introduction of the attention mechanism, which has demonstrated strong performance across diverse visual and multimodal tasks, presents an opportunity to enhance bone age estimation by focusing the model's attention on regions of interest (ROI), much like a physician does. Li et al. (2023) applied a convolutional block attention module to direct the model's focus to specific ROIs. They utilized Xception and ResNet50 architectures for feature extraction from carpal and metacarpal areas, respectively, and achieved modality fusion through dynamic feature map exchange. Using this method, the authors reported a mean absolute error of 5.45 months when adopted for the RSNA dataset[12]. In another study, B. He et al. (2023) implemented a multi-branch attention mechanism to evaluate bone age from ambiguous labels, achieving an MAE of 4.09 months on the same dataset[13].

Generally, methods utilizing manual annotations outperform unannotated approaches in terms of precision and accuracy. However, models that process raw images directly often fall short in leveraging discriminative local details, thus compromising both accuracy and interpretability due to a lack of fine-grained regional analysis. The manual annotation process, though effective, is labor-intensive, which limits the scalability of experimental techniques for clinical applications.

In the realm of bone age estimation using deep neural networks, convolutional neural networks (CNNs) have proven effective in image processing and feature extraction tasks. Xception, a deep CNN model developed by François Chollet at Google, builds upon the traditional Inception network by employing depthwise separable convolutions. This architecture has been broadly employed across different computer vision fields, including segmentation, image classification, and object detection [14].

To leverage these advancements, this project will integrate the Xception model with the convolutional block attention module (CBAM) for enhanced ROI focus and train the model using a large-scale hand bone dataset to improve prediction accuracy.

2 Resources and Techniques

2.1 Dataset

The RSNA bone age dataset, made publicly accessible by the Radiological Society of North America (RSNA), includes a total of 14,236 hand X-rays. This includes 12,611 images for training, 1, 425 for validation, and 200 for testing. The dataset was utilized in the 2017 RSNA challenge, where participants were tasked with estimating the age of children based on hand X-rays. The complete dataset is available for download on RSNA Bone Age

2.2 Bone Age Assessment Network Framework

This study introduces a novel two-stage framework, comprising a network for extracting key bone regions and a subsequent network for recognizing bone age. First, the full hand X-ray is processed through the Xception network, integrated with a CBAM (Convolutional Block Attention

Module) for enhanced feature representation. These features are then refined using the Xception model for deeper learning. Finally, the extracted feature vector is passed to a fully connected layer for bone age assessment (illustrated in Figure 2).

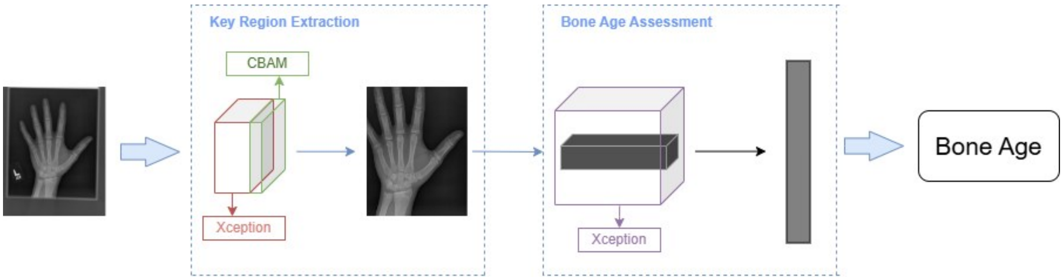

Figure 2 The BAA network framework

2.3 Integration of Xception network and CBAM

The Xception architecture(François Chollet, 2017)[14] was adjusted by removing its top layer and adding additional convolutional and max pooling layers. As illustrated in Figure 3, the architecture consists of three main components: an input module, an intermediate module repeated eight times, and an output module. Unlike the Inception network, Xception applies a filter to each depth map individually before performing convolutional operations across the entire depth dimension. This approach allows the network to process the spatial layout of each channel separately while effectively capturing inter-channel correlations. The Xception model leverages depthwise separable convolutions, an enhanced version of standard convolutions that significantly reduces computational cost while improving efficiency. It consists of 36 convolutional layers dedicated to feature extraction, organized into 14 distinct modules. Except for the first and last modules, all others are connected through linear residual links to mitigate issues like gradient vanishing during training. Additionally, Each convolutional layer and depthwise separable convolution layer utilize ReLU activation functions immediately after, enhancing the model's nonlinearity. The network's design emphasizes a series of depthwise separable convolutional layers stacked sequentially, maximizing parameter utilization for robust feature extraction while maintaining computational efficiency.

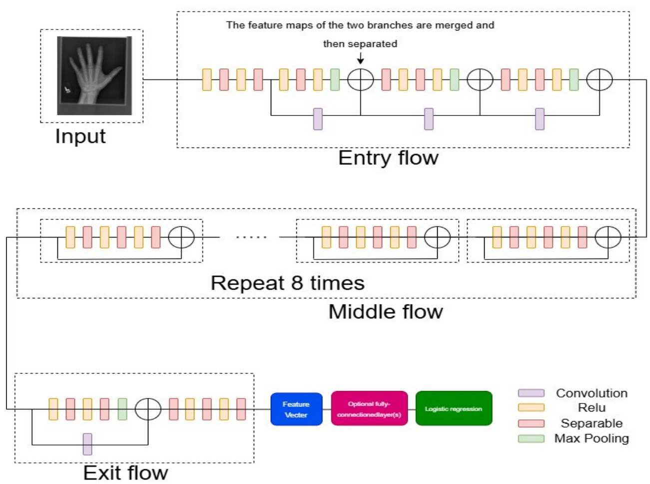

Figure 3 Xception network

To save computational power and quickly extract the most useful features, attention mechanisms have proven effective in enhancing feature extraction capabilities by emphasizing significant attributes while filtering out less relevant information. In particular, CBAM, which was introduced by Woo and colleagues in 2018 [15], is an advanced attention module that integrates two distinct components: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). These components collaborate to boost segmentation accuracy and direct the neural network's focus toward smaller objects. Once the feature maps are obtained from the Xception model, they are first fed into the CBAM module. Initially, the data passes through the channel attention submodule, which generates channel-refined feature data. This data is then element-wise multiplied with the original input features, resulting in enhanced feature representations. Subsequently, the refined output is directed to the spatial attention submodule, which further generates spatially refined features. This final set of features is multiplied once more with the previously refined data to produce the output. The structure of the CBAM module is illustrated in Figure 4.

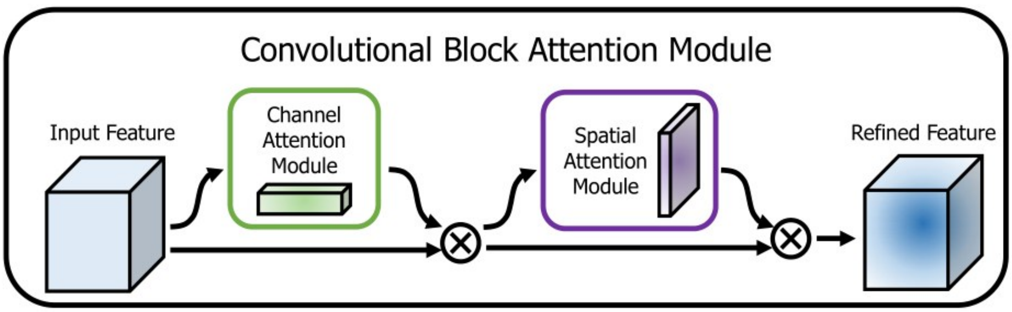

Figure 4 The structure of the CBAM module

Figure 5(A) shows the steps of the CAM architecture. After all the data is input, the CAM function processes the global processing data and the global processing data at the same time, and then divides the output of the two processing units through the common multi-part (MLP). Then the two received signals are combined and processed with the sigmoid process.

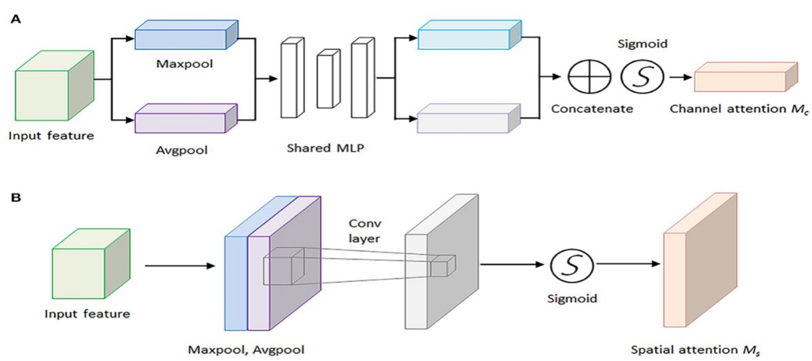

Figure 5 The structures of the channel attention mechanism and spatial attention mechanism

(A) Channel attention mechanism. (B) Spatial attention mechanism

Conversely, Figure 5(B) depicts the calculation process of SAM. Initially, the data outputted from CAM is directed to the SAM module. At this stage, global multi-task operations are successfully executed. Following this, a 7×7 transformation is applied to the combined data, and the resulting weights are processed using the sigmoid activation function.

The calculation method of Mc(F) and the affected area weight Ms(F) can be expressed as follows:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

$$= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}) \quad (1)$$

$$M_s(F) = \sigma(f^{7*7}([AvgPool(F); MaxPool(F)]))$$
$$= \sigma(f^{7*7}([F^s_{avg}; F^s_{max}])) \quad (2)$$

Let F denote the feature map output by each model layer, while MLP refers to a fully connected network layer. Additionally, AvgPool functions as a layer that applies global average pooling, whereas MaxPool operates as a layer performing global maximum pooling. The sigmoid activation function, represented by the symbol σ, plays a critical role in the architecture. To determine spatial attention weights, a method resembling the standard approach is utilized. However, instead of pooling along the spatial dimension as commonly practiced, the pooling process here is conducted across the channel dimension. Furthermore, the MLP layer is substituted with a convolutional layer. Specifically, f(7×7) indicates a convolutional layer using a kernel size of 7×7.

2.4 Network for Bone Age Prediction

In this evaluation module, we use the Xception network to perform overall feature extraction on all bone regions in the hand X-ray image. The Xception network directly performs deep feature extraction on the input image to generate a high-dimensional bone feature expression. These extracted bone features contain information on all bone regions such as carpal bones, metacarpal bones, and phalanges. Next, we combine the extracted bone features with gender information before feeding them into a fully connected network layer, which then generates the estimated age for the hand bones. The structure of the bone age assessment model is depicted in Figure 6.

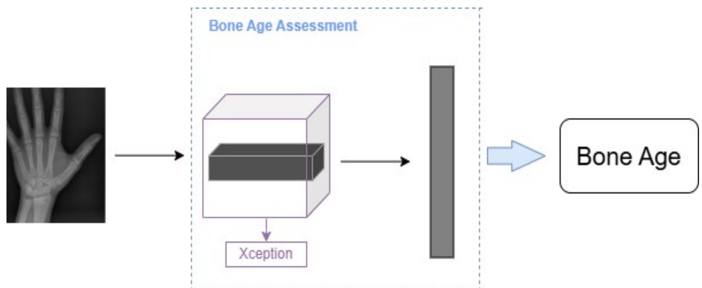

Figure 6 Bone age assessment network diagram

2.5 Loss function and evaluation metrics

In the bone age assessment task, since the goal is to predict a continuous age value, we regard this task as a regression problem. In order to measure the error between the model prediction and the true age, the mean squared error (MSE) is used as the loss function. MSE calculates the squared average of the error between the predicted value and the true value, and is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$$

Among them, y_i is the real age, y_i' is the age predicted by the model, and n is the total number of samples. By minimizing MSE, the model can better fit the data and reduce the prediction error.

To further evaluate the performance of the model on the test set, we use Mean Absolute Error (MAE) as the main evaluation indicator. Compared with MSE, MAE is less sensitive to outliers and can more intuitively reflect the deviation of predicted age. The calculation formula of MAE is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$$

In bone age assessment, MAE has practical significance, namely the average absolute difference between the predicted age and the true age, which makes this indicator very suitable for evaluating the accuracy of bone age prediction models.

During the training process, in order to prevent overfitting, data enhancement was performed on the input hand X-ray images. Data enhancement includes operations such as resampling, rotation, and adjusting contrast and brightness, so that the model can learn robust features under different image conditions. Additionally, the original image dimensions are specified as (300, 300, 3), with a learning rate configured at 0.001 and a total of 10 epochs used for training.

Through these settings and the selection of loss functions and evaluation indicators, the model can be effectively optimized during training and evaluation, and ultimately improve the accuracy in the bone age prediction task.

3 Results

We designed a deep learning model integrated with the CBAM module to reduce the high costs linked to the manual annotation of wrist bone images. The Xception model, enhanced with CBAM, was utilized for feature extraction from X-ray images, allowing the segment extraction module to perform robustly without the need for additional segmentation. This approach enables the model to flexibly and effectively capture large bone structures, enhancing the method's robustness and adaptability.

While being trained using the RSNA dataset, the model designed for predicting bone age exhibited remarkable convergence, with training and validation errors gradually decreasing as the number of epochs increased. In the initial epoch, the training error was close to 30, while the validation error was around 20. However, as training progressed, both errors dropped rapidly. By the second epoch, the training and validation errors were close to 15. Subsequently, the training error continued to decline steadily, while the validation error showed some fluctuations but maintained an overall downward trend. In the final epoch, both training and validation errors were below 10, indicating that the model achieved high accuracy and stability in the bone age prediction task. Upon testing on the RSNA dataset, the model achieved a mean absolute error (MAE) of 7.93 months, as shown in Figure 7.

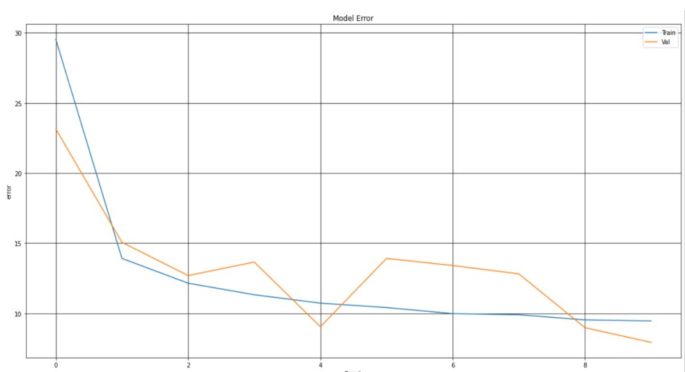

Figure 7 Training and validation error curve of the bone age prediction model

4 Conclusion

In this study, we developed a deep learning model based on the Xception network with the integration of CBAM to improve the accuracy and robustness of automated bone age assessment (BAA) from hand X-ray images. The model was trained and tested on the RSNA dataset, demonstrating strong convergence and a high level of accuracy in predicting bone age. By avoiding the need for manual annotation, our approach provides a cost-effective solution for clinical use, capable of efficiently capturing critical bone features and delivering reliable predictions.

Throughout the training process, both the training and validation errors steadily decreased, indicating effective model convergence. In the final assessment leveraging data from RSNA, the model achieved a MAE of 7.93 months, showcasing its potential for precise and stable bone age prediction. This automated approach significantly reduces the dependency on human expertise, thus lowering the workload for medical professionals and minimizing subjective variability in bone age assessments. The results highlight the model's suitability for clinical application, offering an accurate, robust, and annotation-free method for assessing skeletal maturity.

References

[1] Prokop-Piotrkowska, Monika, et al. "Traditional and new methods of bone age assessment-an overview." Journal of Clinical Research in Pediatric Endocrinology 13.3 (2021): 251.

[2] Cavallo, Federica, et al. "Evaluation of bone age in children: a mini-review." Frontiers in Pediatrics9 (2021): 580314.

[3] Wang, Qingju. Bone growth in pubertal girls: cross-sectional and longitudinal investigation of the association of sex hormones, physical activity, body composition, and muscle strength with bone mass and geometry. No. 110. University of Jyväskylä, 2005.

[4] Hochberg, Ze'ev. Endocrine control of skeletal maturation: annotation to bone age readings. Karger Medical and Scientific Publishers, 2002.

[5] Ren, Xuhua, et al. "Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph." IEEE journal of biomedical and health informatics 23.5 (2018): 2030-2038.

[6] Larson, David B., et al. "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs." Radiology 287.1 (2018): 313-322.

[7] Pan, Xiaoying, et al. "Fully Automated Bone Age Assessment on Large - Scale Hand X - Ray Dataset." International journal of biomedical imaging2020.1 (2020): 8460493.

[8] Iglovikov, Vladimir, and Alexey Shvets. "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation." arXiv preprint arXiv:1801.05746 (2018).

[9] Escobar, María, et al. "Hand pose estimation for pediatric bone age assessment." Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI 22. Springer International Publishing, 2019.

[10] Ren, Xuhua, et al. "Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph." IEEE journal of biomedical and health informatics 23.5 (2018): 2030-2038.

[11] Son, Sung Joon, et al. "TW3-based fully automated bone age assessment system using deep neural networks." IEEE Access7 (2019): 33346-33358.

[12] Li, Zhangyong, et al. "Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction." Frontiers in Artificial Intelligence6 (2023): 1142895.

[13] He, Bishi, et al. "Multi-Branch Attention Learning for Bone Age Assessment with Ambiguous Label." Sensors 23.10 (2023): 4834.

[14] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[15] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.