IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 2, 2018, accepted May 22, 2018, date of publication June 4, 2018, date of current version June 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2843392

Skeletal Maturity Recognition Using a Fully Automated System With Convolutional Neural Networks

SHUQIANG WANG[1,2], (Member, IEEE), YANYAN SHEN[1], (Member, IEEE), CHANGHONG SHI[3], PENG YIN[1], ZUHUI WANG[1], PRUDENCE WING-HANG CHEUNG[2], JASON PUI YIN CHEUNG[2], KEITH DIP-KEI LUK[2], AND YONG HU[2], (Senior Member, IEEE)

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2]Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong
[3]School of Public Health, Guangzhou Medical University, Guangzhou 511436, China

Corresponding author: Yong Hu (yhud@hku.hk)

This work was supported in part by the National Natural Science Foundations of China under Grant 61502473 and Grant U1613208, in part by the Shenzhen Basic Research Projects under Grant JCYJ 20160531184426303, in part by the Shenzhen Overseas High-level Talents Innovation Funds under Grant KQJSCX20170331162115349, in part by the Shenzhen Applied Demonstration Projects under Grant KJYY20160608154421217, and in part by the Natural Science Foundation of Guangdong Province under Grant 2016A030313176.

ABSTRACT In this paper, we present an automated skeletal maturity recognition system that takes a single hand radiograph as an input and finally output the bone age prediction. Unlike the conventional manually diagnostic methods, which are laborious, fallible, and time-consuming, the proposed system takes input images and generates classification results directly. It first accurately detects the distal radius and ulna areas from the hand and wrist X-ray images by a faster region-based convolutional neural network (CNN) model. Then, a well-tuned CNN classification model is applied to estimate the bone ages. In the experiment section, we employed a data set of 1101 hand and wrist radiographs and conducted comprehensive experiments on the proposed system. We discussed the model performance according to various network configurations, multiple optimization algorithms, and different training sample amounts. After parameter optimization, the proposed model is finally achieved 92% and 90% classification accuracies for radius and ulna grades, respectively.

INDEX TERMS Classification, convolutional neural network, detection, radiographs, skeletal maturity.

I. INTRODUCTION

Assessment of a child's skeletal maturity is crucial for management of skeletal disorders during growth [1]. Accurate prediction of growth spurts allow for precise implementation of growth guidance treatment [2]–[5]. Outside of scoliosis, endocrine and metabolic disorders also require good knowledge of skeletal maturity parameters [6]. There are many commonly used radiological parameters but bone age assessment using a hand and wrist radiograph is most accurate [1]. Several commonly used classification methods include the Greulich and Pyle [7] and Tanner-Whitehouse staging [8]–[10]. The GP method utilizes an atlas to match the child's X-ray image to estimate the bone age. This method is simple but is subjected to reliability problems. The TW3 method is more complex as it requires scoring of 20 bone complexes in the hand and wrist to determine the bone age. Due to its complexity, it is difficult to implement in a busy clinic [1]. In 2013, Luk *et al.* [6] simplified the TW3 system and proposed a novel bone age assessment scheme based on the distal radius and ulna (DRU). The DRU has been refined and validated in the adolescent idiopathic scoliosis population [11], [12]. Furthermore, with its wide range, it can detect the growth patterns of the infantile and juvenile population as well. It has also been shown to accurately predict the acceleration and deceleration phases of puberty and also used for prediction of risk of curve progression [13], [14]. With all radiographic parameters, interobserver reliability may be of concern especially for beginners and untrained eyes. Furthermore, standardizing an assessment method provides more reliable data for research purposes.

Deep learning which originates from artificial neural network (ANN) is a powerful kind of machine learning technique. It can learn higher level features from training data

and achieve more accurate. Deep learning technologies are a growing trend in data analysis around the world. Convolutional neural networks (CNN) have been proven to be powerful in image object detection and classification tasks. Some medical research groups have tried to apply CNN and other deep learning models into medical image analysis tasks [15]. For example, interstitial lung disease patterns can be detected and classified precisely by a deep CNN [16]. Also, the vascular network of human eye segmentation task outperforms the previous algorithms when using deep CNN models [17]. To detect and classify pulmonary nodules, Setio *et al.* [18] used multi-view CNN. Hence, deep learning techniques especially CNN have been applied in many medical imaging analysis projects with promising results [15].

Up to now, minimal studies have looked at the potential of using deep learning algorithms to process bone age assessment let alone the DRU classification [19]. In this study, we propose a completely automated deep learning based DRU assessment system. Radiographic images of the DRU will be detected and extracted directly by Faster R-CNN model [20]. Subsequently, this data will be fed into CNN models to predict bone ages. In the experiment section, we will examine several different model configurations of the DRU radiograph dataset and then choose the models that achieve the best performances. In conclusion, the main contributions of the study include: (1) to propose a totally automated and rapid skeletal maturity estimation system which is developed based on deep CNNs; and (2) to improve the models' performance with data sample balance, data augmentation and preferred optimization algorithms.

II. RELATED WORK

Classical methods of GP and TW3 are well known to be time-consuming and inaccurate in clinical management. In 2008, Tristan-Vega and Arribas [21] proposed an end-to-end automated system to estimate children's skeletal age. They adapted a clustering segmentation algorithm to segment the bones contour and then constructed a Generalized Softmax Perceptron (GSP) neural network to estimate bone ages. At the same year, Liu *et al.* [22] built an ANN which contained feed-forward multilayers to estimate bone age from digital left hand-wrist radiographs. Somkantha *et al.* [23] employed boundary information of carpal bone X-ray images to assess bone ages in young children. All these boundary features were processed by the Support Vector Regression (SVR) model to evaluate the bone ages. In 2012, another bone age cluster assessment system utilizing a fuzzy neural network (FNN) was presented by Lin *et al.* [24]. The phalangeal segmentation images were used for the system to predict bone ages. Seok *et al.* [25] constructed another automated bone age determination system using left hand-wrist radiographs in 2012. They first located the feature positions from input images using Scale Invariant Feature Transform (SIFT). Then, they proposed a Singular Value Decomposition based feature vectors to represent those features. After that, all these feature vectors were taken as input into a Neural Network classifier to predict bone ages. Davis *et al.* [26] also proposed a decision tree classifier to predict bone stages by salient radiographic features in 2012. Then, they recreated ages using the standard TW3 approach. Harmsen *et al.* [27] also built a SVM based model to predict bone age with epiphyseal regions of the hand. They evaluated their models comparing with k-nearest neighbor (k-NN) classifier on hand radiographs of 30 diagnostic classes. In 2014, Cunha *et al.* [28] used ensemble techniques to improve bone age assessment. They extracted feature descriptors from each finger joints. Then, these descriptors were combined using various ensemble schemes to obtain an estimated bone age. According to their reports, the combination of bagging and a rule-based regression achieved the best results. Bone age estimation also extended research attention from X-ray images to MRI. For example, Ebner *et al.* [29] proposed multiple random regression forests to locate the joints in a hand MRI via anatomical landmarks. There are still several more bone age assessment projects using conventional machine learning techniques. Although these methods achieved compelling results, they are still not completely automated systems. Moreover, only few deep learning based methods have been applied to solve bone age estimation [19]. Hence, more fully automated deep learning based approaches are necessary to investigate bone age assessment in current medical image areas.

III. DATA AND METHOD

A. DATA PREPARATION

In our experiments, all hand and wrist radiographs were obtained from patients with adolescent idiopathic scoliosis (AIS) undergoing treatment in a tertiary scoliosis clinic. Although all images were of the left hand and wrist, many have different resolutions and positions. A total of 400 radius images and 600 ulna ones were retrieved which also contain label information. The sample image is shown in Fig. 1.

Based on the refined DRU classification system [11], [12], the DRU were graded from R1–R11 and U1–U9. Sample images of radius and ulna at each stage are illustrated in Fig. 2. Data regarding standing height, sitting height, arm span, radius length, and tibia length during all these epiphysis maturity stages were collected. All the growth change information of the radius and ulna at different maturity stages are listed in Table 1.

It is not necessary to identify the bone age stages in such exhaustive division. In most clinical applications, identifying peak growth angrowth cessation is more important for AIS patients since it can be used to judge for initiating or terminating brace-wear. According to Table 1, the growth peak period for radius is around R7 stage and ulna is around U4 or U5 stage. Besides, the growth cessation period for radius is around R9 or R11 stage and ulna is around U9 stage. Then, we can re-define the maturity periods of the radius and ulna base on the above analysis in the following:

(1) The development of radius is graded as 4 periods:

- *growth early period* including R5 and R6 stages
- *growth peak period* including R7 stage

FIGURE 1. Hand radiograph samples of different resolutions and positions.

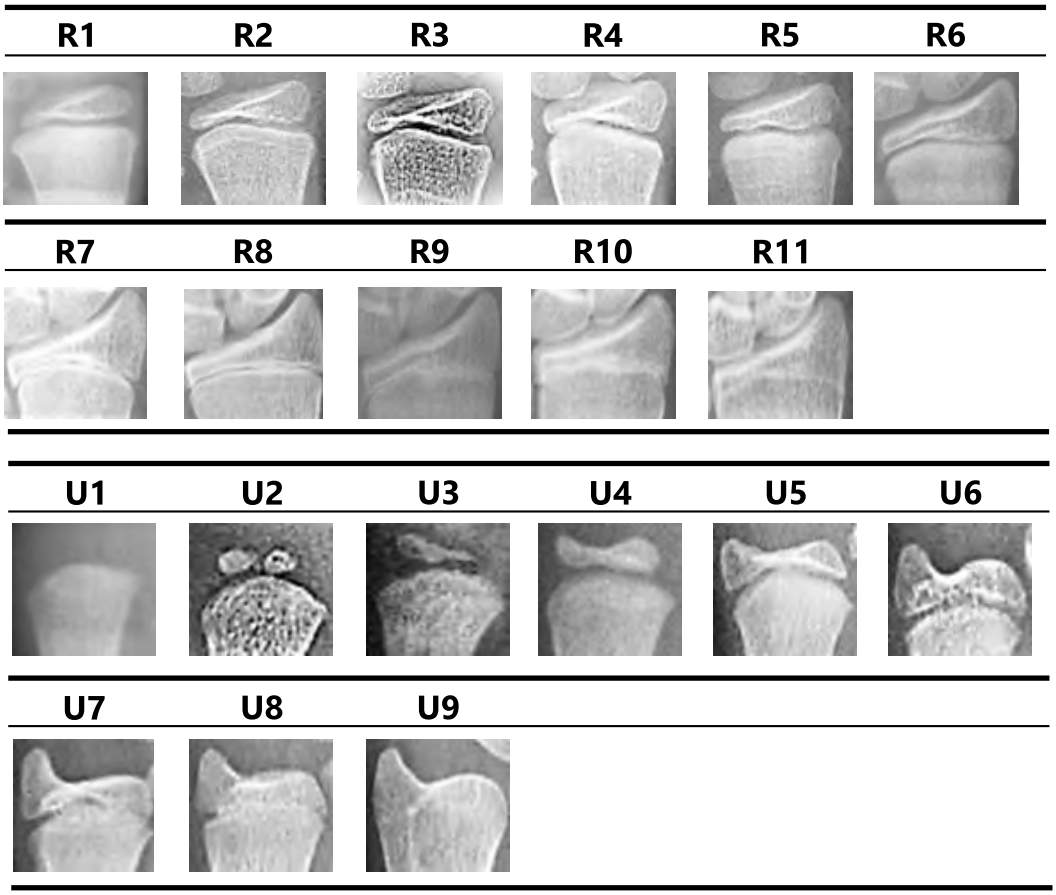


FIGURE 2. Each stage samples of radius and ulna.

- *growth cessation period* including R9 stage
- *growth maturation period* including R11 stage

(2) The development of ulna is graded as 3 periods:

- *growth peak period* including U4 and U5 stages
- *growth transition period* including U7 stage
- *growth maturation period* including U9 stage

The sample radiographs of each developing grade are displayed in Fig. 3.

For the radius radiographs dataset, 75 images of early growth period were gathered from 33 images of R5 stage and 42 images of R6 stage which closely resembled R5. We collected 75 images from the R7 stage and 175 images

TABLE 1. Growth changes of radius and ulna at different maturity stages.

Growth Change		Radius Stages	Ulna Stages
Initial Visible Stage		R5	U2
Chronological Age Interval Between Two Stages		1.1 years	1.1 years
Mean Bone Age Interval Between Two Stages		1.5 years	1.3 years
Growth Peak	Standing Height	R7	U4
	Sitting Height	R7	U5
	Arm Span	R7	U4
	Radius	R7	U5
	Tibia	R6	U5
Growth Cessation	Standing Height	R9	U9
	Sitting Height	R9	U9
	Arm Span	R11	U9
	Radius	R11	U9
	Tibia	R9	U9

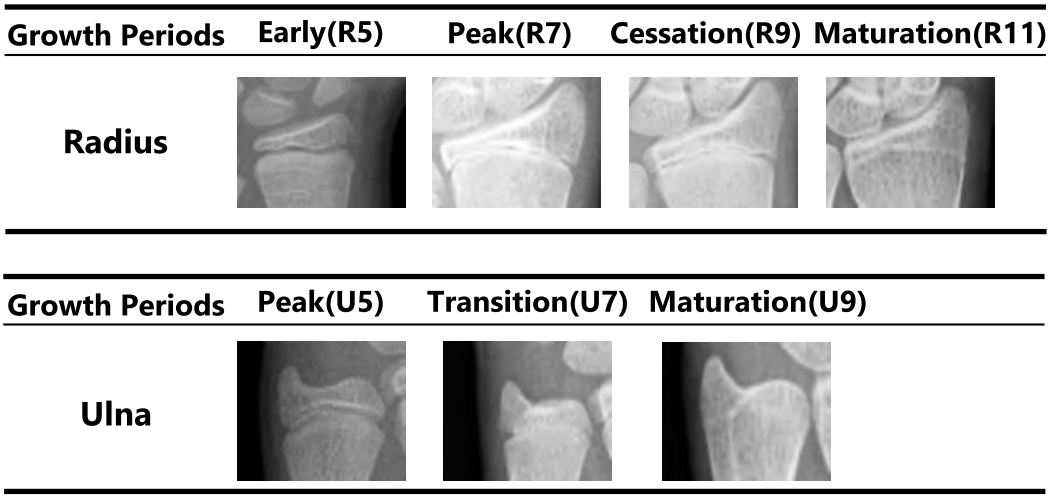


FIGURE 3. New stages of radius and ulna epiphysis maturity grades.

from R9 to create the peak growth period and growth cessation period respectively. Also, 75 images were selected from the original R11 stage to build the maturation period. There were 400 radius X-ray images in total. They were separated as: 100 images which were randomly selected equally from the above four periods as testing samples and 300 images treated as training samples. The separation details are listed in Table 2 and Fig. 4(a).

TABLE 2. Samples distribution of radius developing grades.

Growth Periods	Early	Peak	Cessation	Maturation
Training Samples (300)	50	50	150	50
Testing Samples (100)	25	25	25	25
Total (400)	75	75	175	75

For the ulna radiographs dataset, we collected 191 images from the U4 and U5 stages and selected 9 images from the U6 stage which resembled the U5 stage. There were finally 200 images of early growth period. 250 images were chosen from the U7 stage to be composed as growth transition period. Also, we combined 138 images from the U9 stage and 12 images from the U8 stages to construct the growth maturation period samples. We collected 600 ulna X-ray images in total. Then, we randomly selected 150 images equally from the three periods to establish as testing samples. The remaining 450 images were used as training samples. The details are showed in Table 3 and Fig. 4(b).

TABLE 3. Samples distribution of ulna developing grades.

Growth Periods	Peak	Transition	Maturation
Training Samples (450)	150	200	100
Testing Samples (150)	50	50	50
Total (600)	200	250	150

B. SAMPLE BALANCE AND DATA AUGMENTATION

For our original clinical DRU dataset, we collected around 1000 image samples in total. However, these data samples

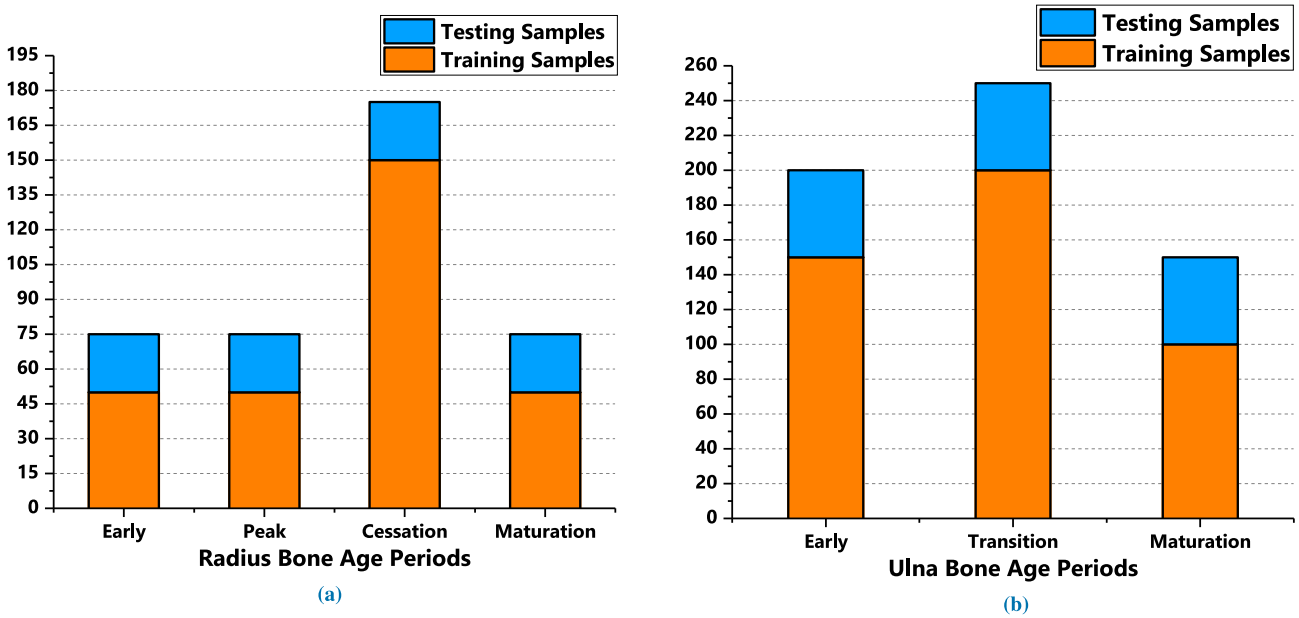


FIGURE 4. Radius and ulna samples distribution in each growth period. (a) Radius samples distribution. (b) Ulna samples distribution.

were not well balanced between each category (growth period). Networks cannot learn useful enough features from categories with less samples. It will affect networks performance negatively. To solve this problem, oversampling the less prevalent data can balance the dataset samples. In our training dataset, radius data distribution was 50, 50, 150 and 50 of each group. This data was balanced to be 150 in each maturity period. The ulna dataset contained 150, 200 and 100 of each period which was balanced with all data as 200 images.

In our experiment section, we explored numerous deep neural networks with various configurations to evaluate their performance. When training deep networks, we needed sufficient training data in order to prevent overfitting issues. More specifically, some data augmentation transformations such as translations and rescale was used to enlarge the training dataset in the experiment section. Finally, for the radius training dataset, we obtained 750 radius images of each bone age period. Also, 1000 ulna images of each period were collected in the ulna training dataset. The details of radius and ulna data distribution are shown in the Tables 4 and 5 and Fig. 5.

C. METHOD

The original X-ray images were taken on the whole hand and wrist area as illustrated in Fig. 1. Besides, images were of different resolutions and sizes. It was too large to use the original radiographs as input for the CNN models directly. For example, there were two representative resolutions: 1400 × 900, 989 × 1302. However, the radius and ulna information were assessed within a small region of original X-ray images only. Furthermore, these regions have almost identical sizes and fixed locations. To reduce other irrelevant regions' negative influences, we first decided to use object detection tools to

TABLE 4. Radius training data distribution after balance and data augmentation.

Bone Age Periods (Radius)	Early	Peak	Cessation	Maturation	Total
Original Training Data	50	50	150	50	300
After Data Balance	150	150	150	150	600
After Data Augmentation	750	750	750	750	3000

TABLE 5. Ulna training data distribution after data balance and augmentation.

Bone Age Periods (Ulna)	Peak	Transition	Maturation	Total
Original Training Data	150	200	100	450
After Data Balance	200	200	200	600
After Data Augmentation	1000	1000	1000	3000

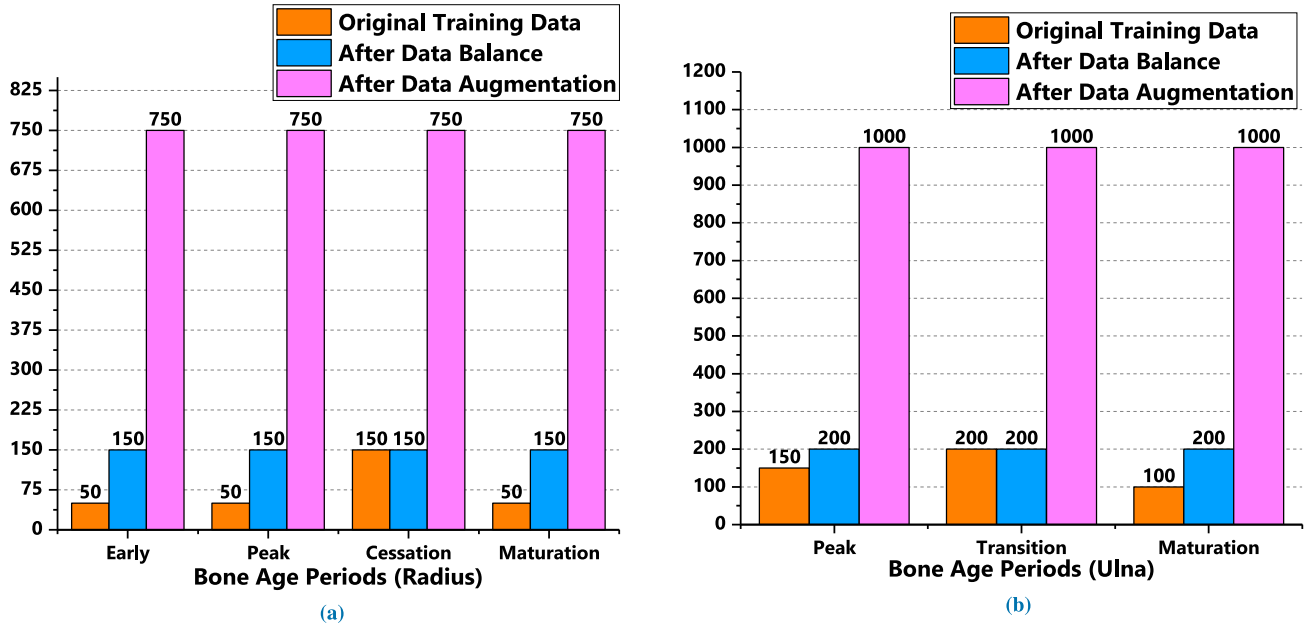


FIGURE 5. Radius and ulna training data distribution after data balance and augmentation. (a) Radius training data distribution. (b) Ulna training data distribution.

extract distal radius and ulna regions. Then, we used the detected image segments to cut out the radius and ulna regions separately and resized them into identical resolutions. After obtaining the radius and ulna regions images, new labels were assigned to them and saved in a file. Finally, we used the data to train CNN to predict the bone ages and visualize the final outputs. The whole architecture of the proposed system is illustrated in Fig. 6.

The proposed bone age assessment system mainly included the following key steps:

1) Data preparation
2) Object detection model which extracted distal radius and ulna regions from hand radiographs
3) Automatic clipping distal radius and ulna regions and to resize them
4) CNN-based classifier to predict bone age stages
5) Final output visualization

The above steps can be summarized into one flow diagram displayed in Fig. 7.

1) ROI DETECTION

In our proposed bone age assessment system, distal radius and ulna regions were extracted separately from original hand and wrist X-ray images. These regions were inputted into a CNN to do bone age prediction. Therefore, their resolutions were resized identically i.e. 128 × 96 in our experiment. To detect these distal radius and ulna regions, we chose to use the Faster R-CNN algorithm [20]. The basic network was ZFNet [30]. It had 5 shareable convolutional layers. The first convolutional layer had 96 filters with 7 × 7 size. Then it was followed by a 3 × 3 max pooling layer. The second convolutional layer contained 256 filters with 5 × 5 size. A 3 × 3 max pooling layer was attached to the preceding layer. The strides were 2. Then, three 3 × 3 convolutional layers were added to

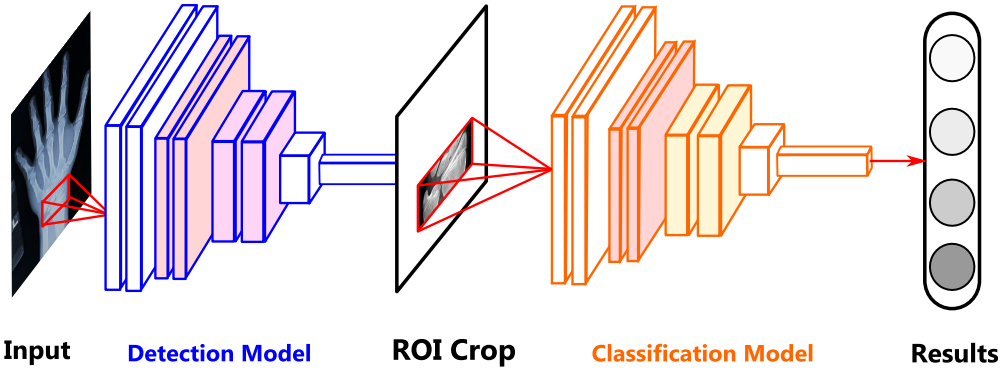


FIGURE 6. The whole architecture of our proposed deep learning system.

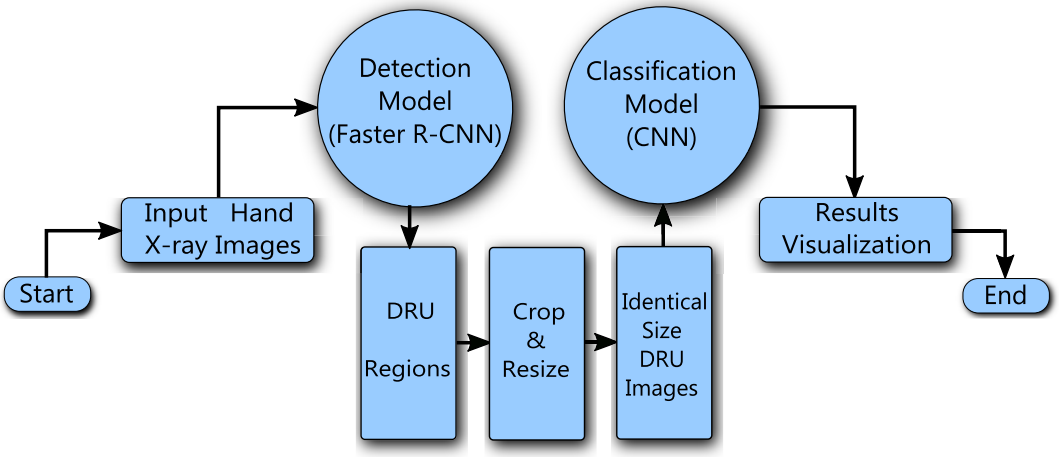


FIGURE 7. Framework of bone age assessment system.

the model. They have 384, 384 and 256 filters respectively. The strides were all 1. After this, a 3×3 convolutional layer was attached to these shared convolutional layers. Then it was followed by two sibling 1×1 convolutional layers. The score of proposed regions and the bounding box was presented as output in the end. The above network structure was the so-called Region Proposal Network (RPN) which was used to extract regions of interest (ROI). The shareable convolutional layers and ROI information were passed into the ROI pooling layer. Then, it was followed by two fully connected layers which have 4096 neurons on each layer. Finally, the network showed the classified score and bounding box coordinates as output. The above network structure is known as Fast R-CNN [31]. It predicted the exact ROI bounding box position and corresponding categories. In our proposed example, we set 3 categories which included radius, ulna and background. The proposed object detection model structure is shown in Fig. 8. The detection model was trained by back-propagation and stochastic gradient descent (SGD) with mini-batch data. The initial learning rate was set as 0.01 and all layer weights were randomly initialized from a Gaussian distribution. The first step was to train RPN with 8000 iterations and to train Fast R-CNN with 4000 iterations as the second step. Then, the third step was to train RPN with another 8000 iterations and finally the fourth step to train Fast R-CNN with another 4000 iterations.

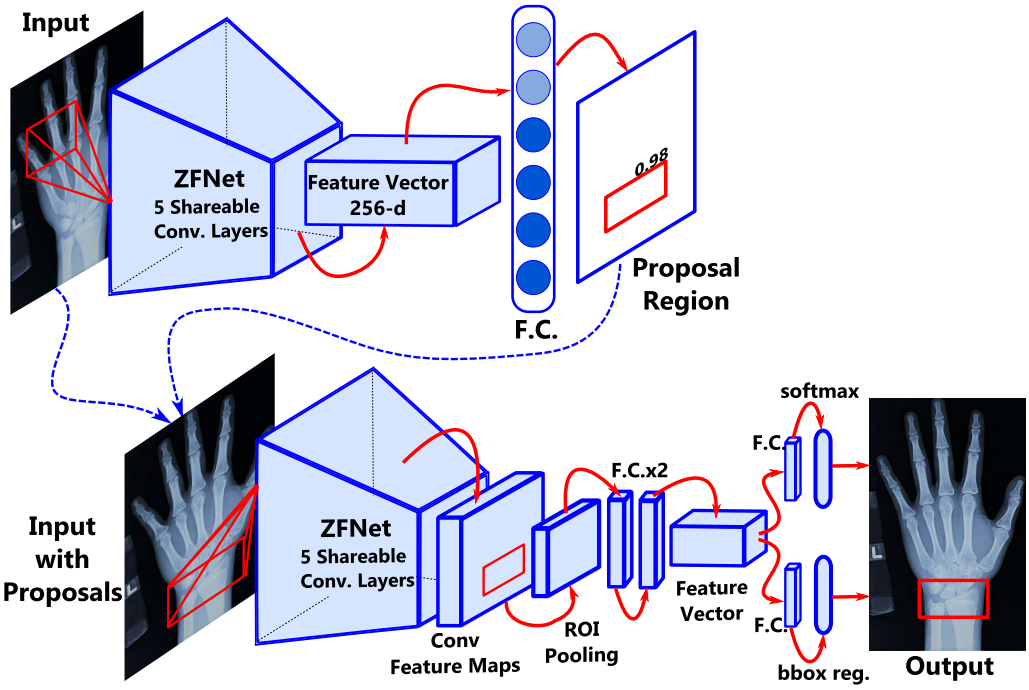


FIGURE 8. The proposed object detection model structure (Faster R-CNN).

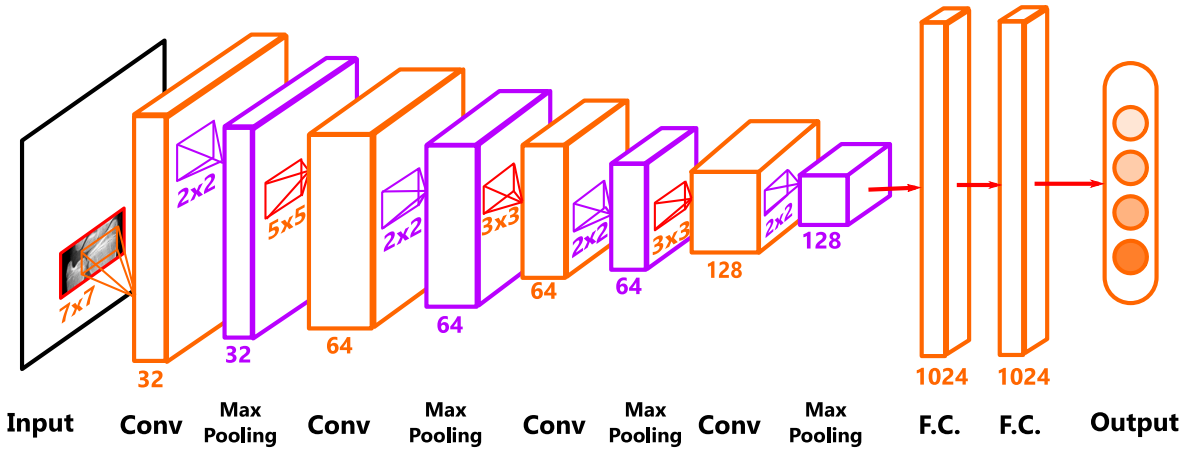


FIGURE 9. The proposed CNN classification model structure.

2) SKELETAL MATURITY CLASSIFICATION

CNN is a multiple-layer neural network model. It includes an input layer, hidden layers and an output layer. Normally hidden layers consist of one or several convolutional layers and pooling layers alternately followed by one or several fully connected layers. The features of input images can be extracted by convolutional layers and pooling layers can reduce data dimensions and improve feature invariants. Fully connected layers are used to choose useful features to construct mapping relations between previous layers and final outputs. The proposed CNN classification model configuration is illustrated in Fig. 9. In our experiment, the proposed network configuration included: input data with the DRU image size of 128 × 96, the number of convolutional layers was 4 or 5, convolutional kernel sizes were chosen as 7 × 7, 5 × 5, 3 × 3 (ZFNet) or all 3 × 3 (VGGNet) [32]. The number of convolutional kernels were selected as 32, 64 and even 128. Each convolutional layer was followed by a 2 × 2 max pooling layer. Two more fully-connected layers in which the size was set as 4096 or 1024 with or without Dropout [33] were added into the model. ReLU was used as active functions. In this work, we employ the index, accuracy, to evaluated the performance of the proposed model. Accuracy is calculated by (TP+TN)/(TP+TN+FP+FN), where TP=True Positive, TN=True Negative, FP=False Positive, and FN=False Negative. False positive is a classification result that indicates a given condition exists, when it does not, and false negative is a classification result that indicates that a condition does not hold, while in fact it does. The best performance parameters configuration for our proposed system are described in the following section.

IV. EXPERIMENT AND RESULTS

A. EXPERIMENT CONFIGURATION

Our experiment was implemented on Ubuntu 16.04 system. CPU is Intel® Xeon® CPU E5–1620 v3. Its CPU frequency is 3.5GHz, GPU is NVIDIA Quadro M4000, CUDA8.0, cuDNN5.0. Detection task was based on Caffe framework, and classification task was based on Theano, Lasagne Deep Learning framework. In the training process, we used the Adam optimization algorithm [34] and minimized the multiple classification cross-entropy loss function. Adam is a kind of self-adapted parameters update algorithm which includes 3 parameters: learning rate and two decay parameters. They were initialized as 0.0001 and 0.9, 0.999 in our experiment. The weight matrix $\boldsymbol{W}$ initialized with Xavier uniform distribution, and the bias $\boldsymbol{b}$ uses 0 as the initialization value. These parameters were updated using mini-batch with size of 16.

B. CLASSIFICATION RESULTS

1) MODEL PERFORMANCE WITH SAMPLE BALANCE

In this subsection, we used the above two different datasets (balanced & unbalanced) to verify the performance in some basic neural network models. The basic network structure in our experiment was as follows: 4 convolutional layers and 2 fully-connected layers. Our network detailed configuration was: the first convolutional layer with 32 filters with size 7 × 7. The second convolutional layer contained 64 filters with size 5 × 5. The third one was 64 filters with size 3 × 3, and the fourth one was 32 filters with size 3 × 3. The sizes of two fully-connected layers were 4096 and 1024. The output size was set as 3. Each convolutional layer was followed by a 2 × 2 max pooling layer. The experiment results are displayed in Fig. 10.

From the above experiment results, there was only tiny accuracy improvement after data balanced. However, during the training procedure, the balanced samples convergence speed was faster than unbalanced samples convergence speed. In Fig. 11(a), the unbalanced data experiment started to converge around 40 iterations. But, according to Fig. 11(b), the balanced data experiment began to converge after 20 iterations. This indicated that data balance was effective and time-saving when training neural networks.

2) MODEL PERFORMANCE WITH DATA AUGMENTATION

In our experiment, we used translations and rescale tricks to enlarge the radius and ulna training dataset 5 times more. This led to 1000 samples in each category and 3000 samples

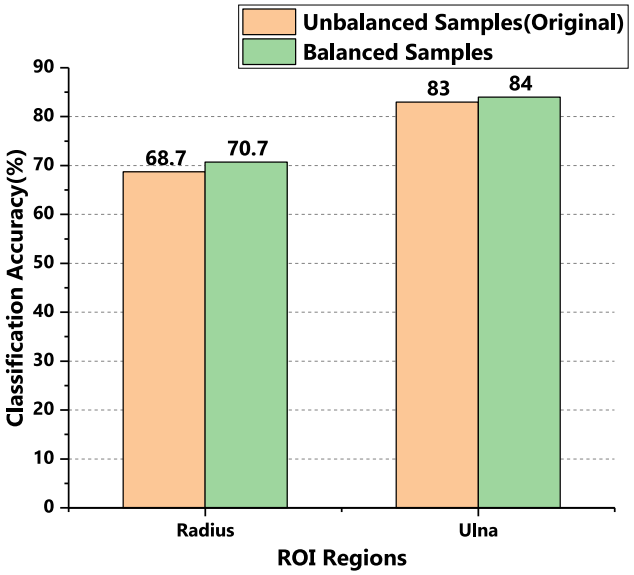


FIGURE 10. The network performance with different data sample distributions.

in total. Also, we compared the performance of the above two datasets with a basic neural network model. The model contained 4 convolutional layers and 2 fully connected layers. The comparison result is posted in Fig. 12.

As shown in Fig. 12, data augmentation helped to improve the model's performance significantly. In the rest of the experiment section, we composed new experiments with the dataset after data augmentation. Even more, data augmentation also accelerated the models' convergence speed during training. We used the same network configurations with identical parameters only except for the training data (with/without augmentation). As shown in Fig. 13, the training loss of models which fed with augmented data decreased dramatically after about 10 iterations. On the other hand, the training loss of models which trained with original data reduced towards zero after 30 iterations.

3) MODEL PERFORMANCE WITH DIFFERENT CONFIGURATIONS

For this bone age assessment task, we wanted to determine the best performance network configuration in the subsection. There were several factors which influenced the model performance such as the number of neurons in fully connected layers, the size of convolutional kernels, the number of convolutional layers and Dropout techniques. We compared seven different network configurations of prediction accuracy and chose the best one. The experiment results are recorded in Table 6.

In the above table, the column of *Convolutional Layer* stands for the convolutional layers configuration.

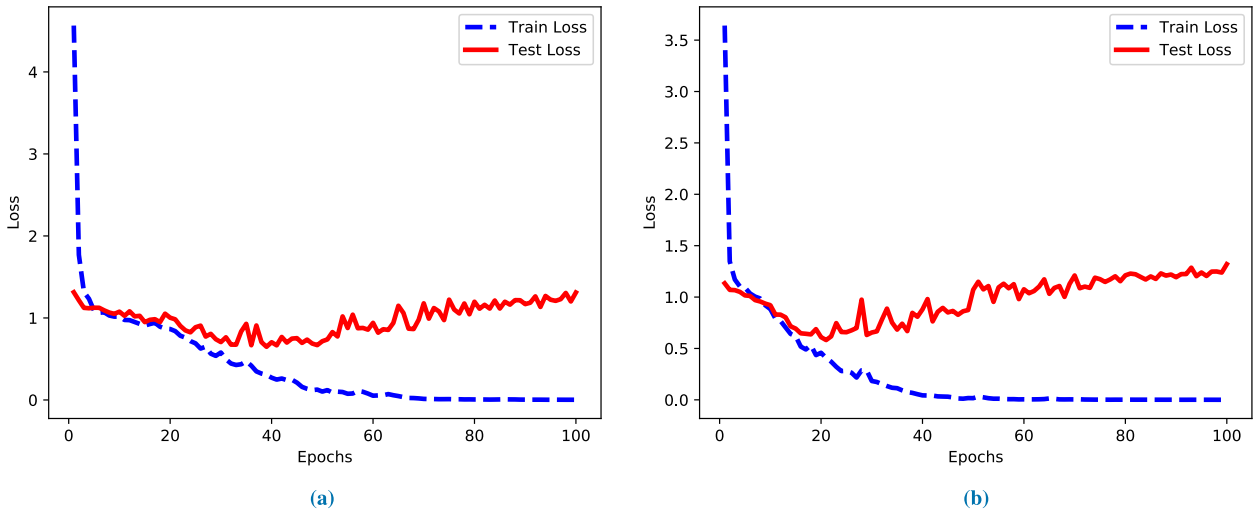


FIGURE 11. Convergence speed comparison of unbalanced samples and balanced samples. (a) Unbalanced Data. (b) Balanced Data.

TABLE 6. Various network configurations in the experiment.

Net No.	Network Configurations	
	Convolutional Layer	Fully-Connected Layer
1	32c7–64c5–64c3–32c3	1024–1024
2	32c3–64c3–64c3–32c3	1024–1024
3	32c3–64c3–64c3–128c3–128c3	4096–1024
4	32c7–64c5–64c3–128c3–128c3	1024–1024
5	32c5–64c5–64c5–128c3–128c3	1024–1024
6	32c5–64c5–64c5–128c3–128c3–128c3	4096–1024
7	32c3-64c3-64c3-128c3-128c3-128c3	1024–1024

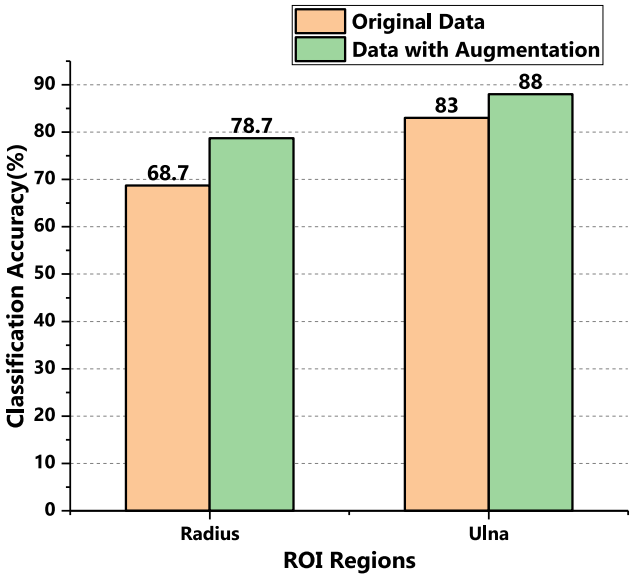


FIGURE 12. The network performance with/without data augmentation.

For example, "32c7–64c5–64c3–32c3" means it has 4 convolutional layers. The first layer (32c7) represents 32 filters with size 7 × 7. The stride is 1. Each convolutional layer is followed by a 2 × 2 max pooling layer. Besides, the column of *Fully–Connected Layer* represents a different combination of fully–connected layers. For instance, "1024–256" means there are two fully-connected layers with the size of 1024 and 256 respectively. During training procedure, we used the Adam optimization algorithm to minimize the loss function. The learning rate and two decay parameters of Adam algorithm were set at 0.0001 and 0.9, 0.999 respectively. The weight matrix $\boldsymbol{W}$ was initialized with Xavier uniform distribution. Also, bias $\boldsymbol{b}$ was set equal to 0 at the beginning. The classification results of the radius and ulna are displayed in Fig. 14.

From the above experiment results, we can observe that the radius classification result has the best performance of 0.92 on No. 3 network classification and ulna classification performs the best which is 0.88 on No. 4 network classification. The results illustrate that deeper networks cannot guarantee higher classification accuracies. It also explains that the appropriate network configuration for DRU classification should be decided after various experiments with different network structures. Additionally, networks that use Dropout will also help improve the classification accuracy.

4) MODEL PERFORMANCE WITH VARIOUS OPTIMIZATION ALGORITHMS

In this subsection, we compared classification accuracies with several optimization algorithms including: *Stochastic Gradient Decent (SGD)*, *Adam*, *Nesterov's Accelerated Gradient (Nesterov)*, *Adaptive Gradient (AdaGrad)*, *RMSprop* and *AdaDelta* with different learning rates. The No. 3 and No. 4 network configurations for the radius and ulna were used in this part. The experiment results are displayed in Fig. 15. Based on the experiment results, Adam performed better than any other optimization algorithm in our bone age classification task.

We also verified the network (No.3 and No.4) performances with Dropout techniques on the selected optimization techniques. Adam, SGD and RMSprop were chosen as the top 3 optimization algorithms to be tested in the following experiments. The results are displayed in Fig. 16. We can say that Dropout will help the network achieve higher classification accuracy. Moreover, Dropout technique helps the proposed model achieve a stable loss value during training. Experiment results are displayed in Fig. 17. In a word, Adam with Dropout performs slightly better than the other two

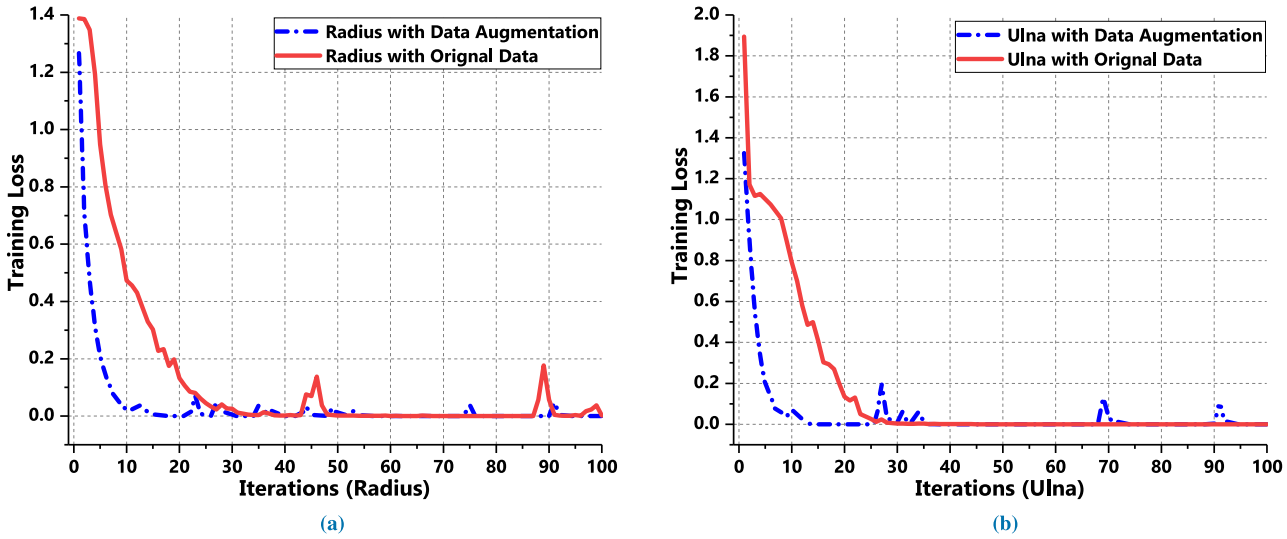


FIGURE 13. Training loss of models trained with or without data augmentation. (a) Radius. (b) Ulna.

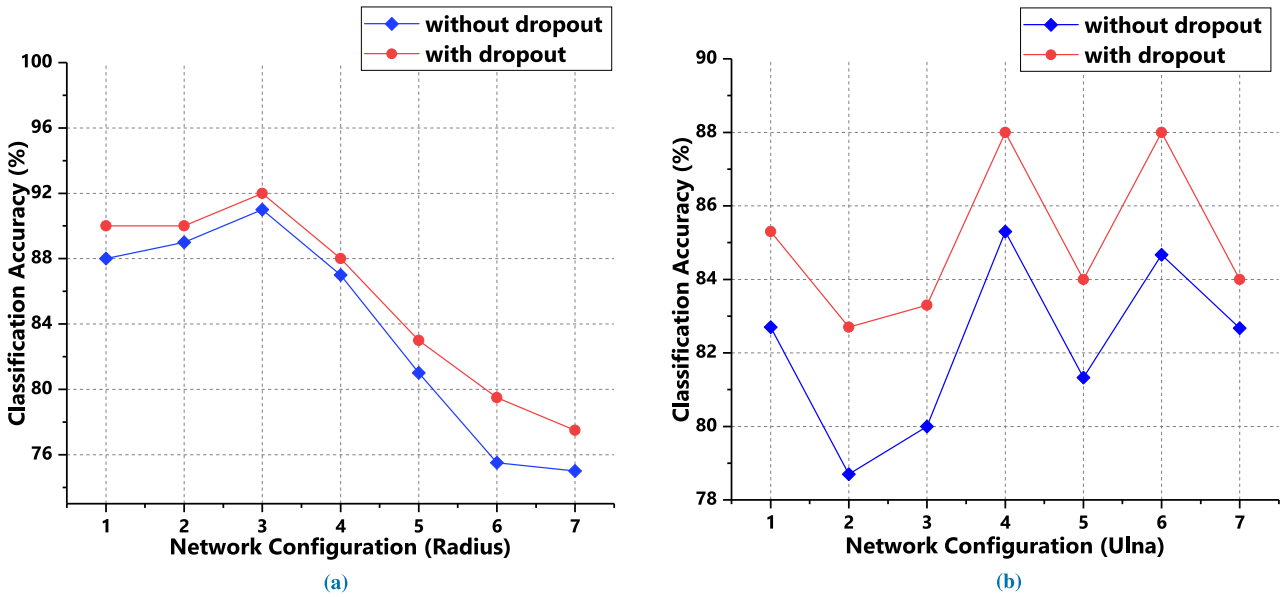


FIGURE 14. Radius and ulna classification results of different network configurations. (a) Radius. (b) Ulna.

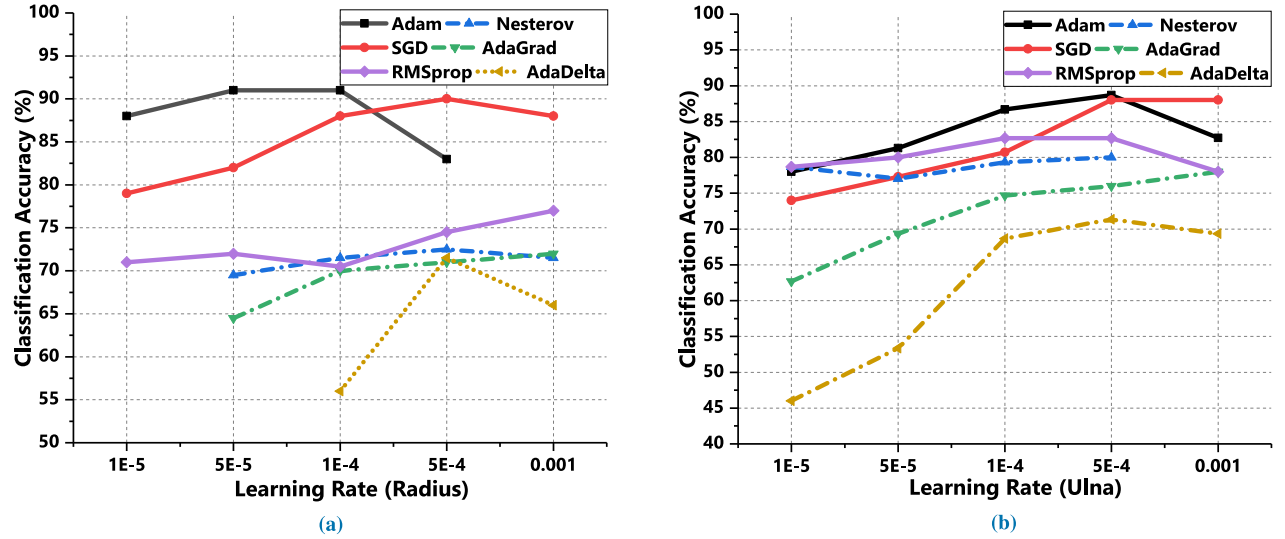


FIGURE 15. Radius and ulna classification accuracy with diverse optimization algorithms and learning rates. (a) Radius. (b) Ulna.

algorithms with Dropout. Radius classification accuracy can achieve 92% and ulna classification accuracy finally reaches the maximum value of 90%.

Besides all the experiments described above, we also want to explore the proposed model performance according to different number of training samples. We collected different training set sizes including 800, 1500, 2000, 2500 and 3000. All other model configurations were kept identical except training samples. The classification accuracies are reported in Fig. 18. We can make a conclusion that the highest accuracies are always achieved by models with 3000 training samples.

To summarize the performances of different experiment configurations in this section, we can conclude that the best network configurations for radius and ulna classification tasks are listed in Table 7.

TABLE 7. The best radius and ulna classification accuracy results.

	Net No.	Optimization Algorithms	Learning Rate	Training Sample Amounts	Classification Accuracy
Radius	3	Adam + Dropout	0.0001	3000	92%
Ulna	4	Adam + Dropout	0.001	3000	90%

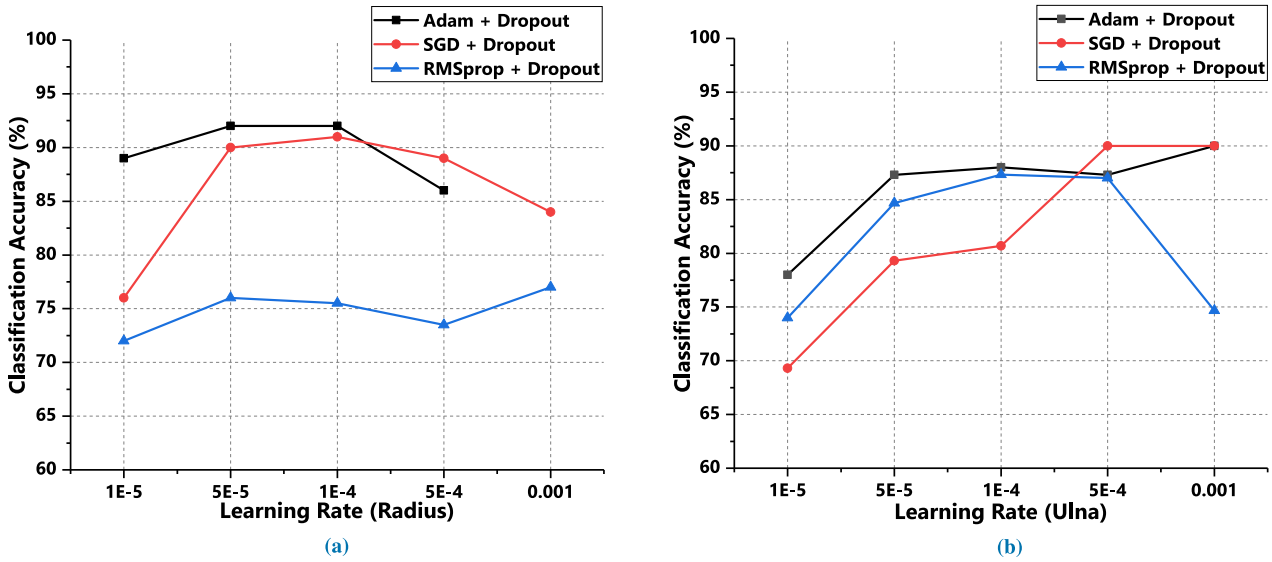


FIGURE 16. Three selected optimization algorithms (Adam, SGD and RMSprop) performance with dropout. (a) Radius. (b) Ulna.

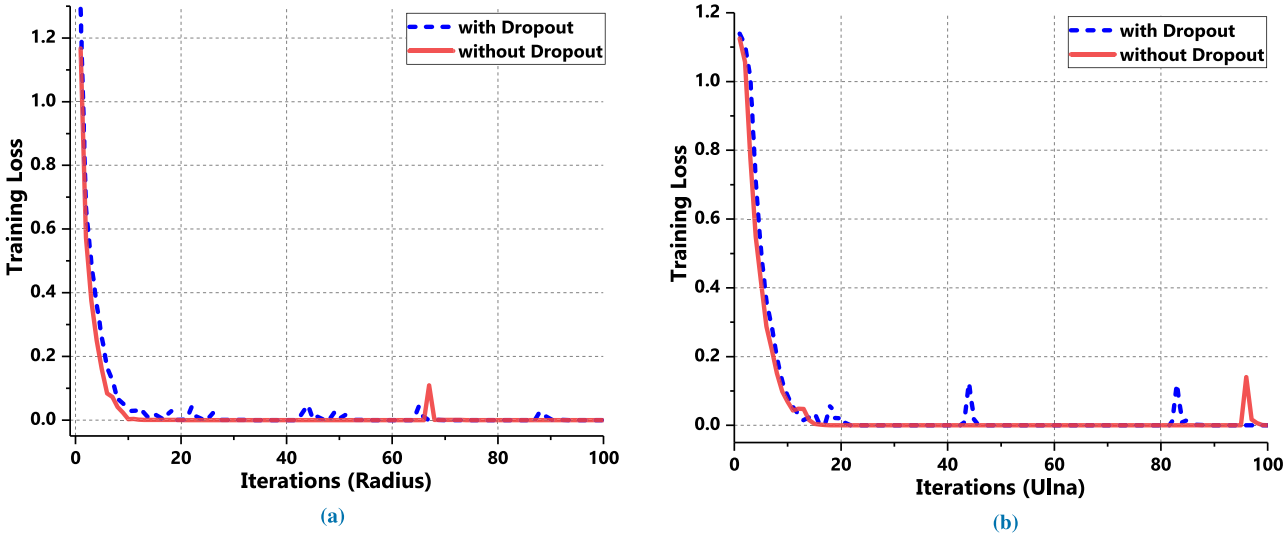


FIGURE 17. Training loss value of model training with/without dropout. (a) Radius. (b) Ulna.

V. DISCUSSION

A. ERROR ANALYSIS

In this section, we used confusion matrix to perform error analysis and then computed recall rate and accuracy of each bone maturity stage. For example, two confusion matrices have been created for the experiment accuracies of radius (0.90) and ulna (0.87) classification in Fig. 19. In the confusion matrix, the rows stand for the accurate labels. The predicted labels are denoted by the columns. The value at row i and column j represents the true label is i which is predicted as j. The diagonal line in the matrix contains the number of correct classification results.

The recall and precision rate of radius and ulna classification results can be calculated from the above confusion

TABLE 8. Recall and precision rate of radius classification results at each stage.

Growth Periods	0–Early	1–Peak	2–Cessation	3–Maturation
Recall	0.88 (22/25)	0.76 (19/25)	0.96 (24/25)	1.00 (25/25)
Precision	0.85 (22/26)	0.86 (19/22)	0.92 (24/26)	0.96 (25/26)

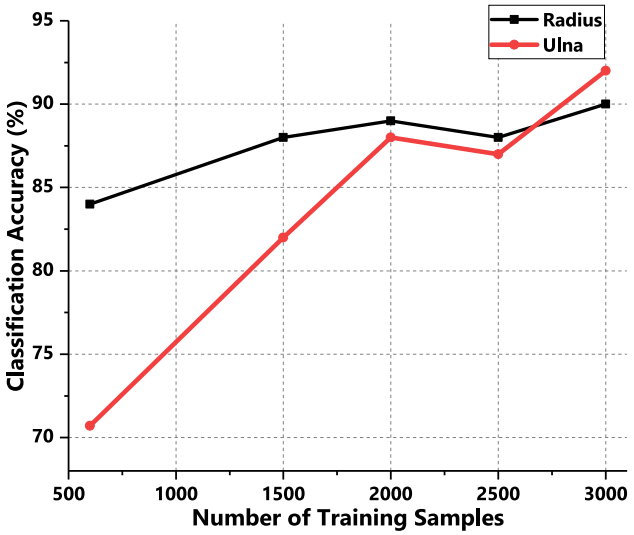


FIGURE 18. Classification accuracy of the proposed model with different size training datasets.

matrices. The classification errors are mainly caused by adjacent maturity stages which have high similarity in the X-ray images. The recall and precision rate are summarized in Tables 8 and 9.

B. COMPARISON WITH OTHER METHODS

The results of comparison with other methods are shown in Table 10. As for the test accuracy, our method is significantly superior to other methods. Comparing with the traditional methods [7], [9], [10], the greatest advantages of our method is its high efficiency and time saving. Once the proposed automated skeletal maturity recognition system is established, it can work without involving manual work. The

TABLE 9. Recall and precision rate of ulna classification results at each stage.

Growth Periods	0–Peak	1–Transition	2–Maturation
Recall	0.92 (33/36)	0.88 (28/32)	0.81 (26/32)
Precision	0.92 (33/36)	0.76 (28/37)	0.96 (26/27)

TABLE 10. Comparison of accuracy with others methods.

Growth Periods	Our methods	Ref.[19]	Ref.[35]
Radius	92%	88%	75%
Ulna	90%	87%	76%

disadvantage of the proposed model is that its performance relies heavily on training samples with high quality. However, a large of labeled samples with high quality are always difficult to collect. Our future work will focus on the optimised model which can work with limited training samples.

VI. CONCLUSION

In this study, we utilized hand and wrist radiographs and DRU data to train CNN for automatic analysis of skeletal maturity. Our proposed system can improve the bone age assessment efficiency, reduce doctor's workload and assist physicians' clinical decisions. The best performance of the system achieves a radius classification accuracy of 92% and an ulna classification accuracy of 90%. Moreover, the classification errors by confusion matrices were analyzed. Although performing well, our system still has some limitations. For example, early growth stage data were limited for training our model. Also, the detected distal radius and ulna regions contain noise which influences the final prediction accuracy of our model. In future work, we will consider

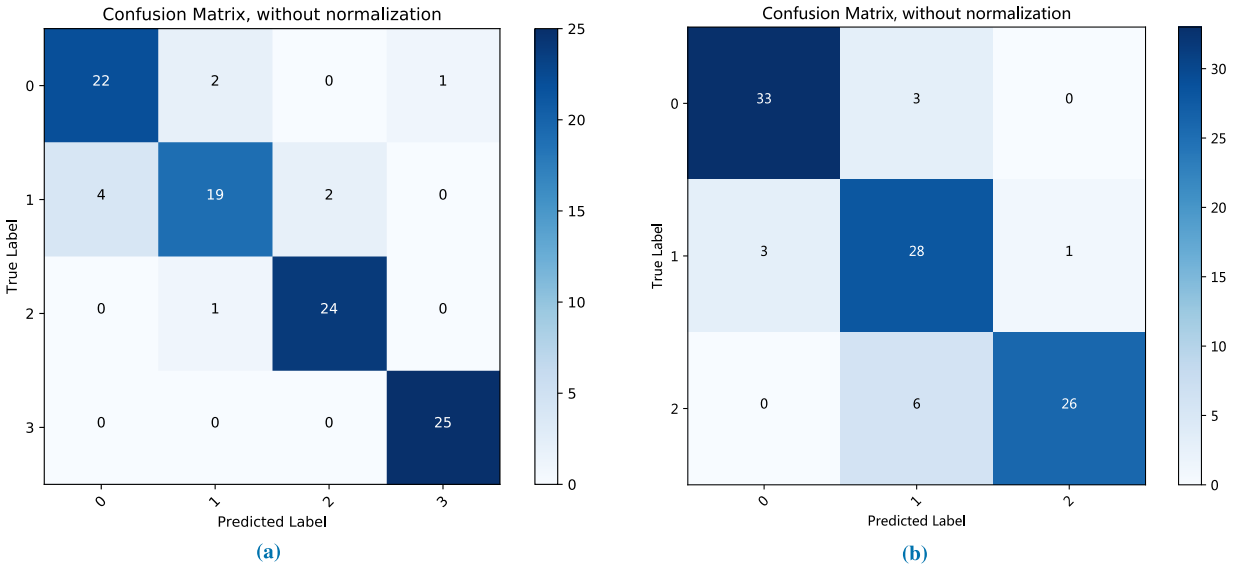


FIGURE 19. Confusion matrices of radius and ulna classification results. (a) Radius. (b) Ulna.

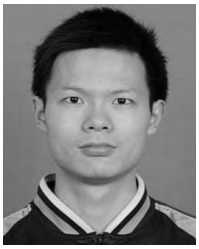
implementing more robust deep learning based systems to overcome the above issues. For example, more patients' features (e.g. standing height, sitting height and arm span) can be involved in the classification procedure. Even deeper and more complicated network models can be chosen to perform skeletal maturity classification tasks. This has high likelihood of achieving better bone age assessment results.

REFERENCES

[1] J. P. Y. Cheung and K. D.-K. Luk, "Managing the pediatric spine: Growth assessment," *Asian Spine J.*, vol. 11, no. 5, pp. 804–816, 2017.
[2] K. M. C. Cheung *et al.*, "Magnetically controlled growing rods for severe spinal curvature in young children: A prospective case series," *Lancet*, vol. 379, no. 9830, pp. 1967–1974, 2012.
[3] M. Dominkus, P. Krepler, E. Schwameis, R. Windhager, and R. Kotz, "Growth prediction in extendable tumor prostheses in children," *Clin. Orthopaedics Rel. Res.*, vol. 390, pp. 212–220, Sep. 2001.
[4] R. B. Duthie, "The significance of growth in orthopaedic surgery," *Clin. Orthopaedics Rel. Res.*, vol. 14, pp. 7–19, Jun. 1959.
[5] G. H. Thompson, B. A. Akbarnia, and R. M. Campbell, Jr., "Growing rod techniques in early-onset scoliosis," *J. Pediatric Orthopaedics*, vol. 27, no. 3, pp. 354–361, 2007.
[6] K. D.-K. Luk, L. B. Saw, S. Grozman, K. M. Cheung, and D. Samartzis, "Assessment of skeletal maturity in scoliosis patients to determine clinical management: A new classification scheme using distal radius and ulna radiographs," *Spine J.*, vol. 14, no. 2, pp. 315–325, 2014.
[7] W. W. Greulich and S. I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist," *Amer. J. Med. Sci.*, vol. 238, no. 3, p. 393, 1959.
[8] J. M. Tanner and R. H. Whitehouse, "Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty," *Arch. Disease Childhood*, vol. 51, no. 3, pp. 170–179, 1976.
[9] J. M. Tanner, R. H. Whitehouse, P. C. R. Hughes, and B. S. Carter, "Relative importance of growth hormone and sex steroids for the growth at puberty of trunk length, limb length, and muscle width in growth hormone-deficient children," *J. Pediatrics*, vol. 89, no. 6, pp. 1000–1008, 1976.
[10] J. M. Tanner, R. H. Whitehouse, E. Marubini, and L. F. Resele, "The adolescent growth spurt of boys and girls of the Harpenden growth study," *Ann. Hum. Biol.*, vol. 3, no. 2, pp. 109–126, 1976.
[11] J. P. Y. Cheung, D. Samartzis, P. W. H. Cheung, K. M. Cheung, and K. D.-K. Luk, "Reliability analysis of the distal radius and ulna classification for assessing skeletal maturity for patients with adolescent idiopathic scoliosis," *Global Spine J.*, vol. 6, no. 2, pp. 164–168, 2016.
[12] J. P. Y. Cheung, D. Samartzis, P. W. H. Cheung, K. H. Leung, K. M. C. Cheung, and K. D.-K. Luk, "The distal radius and ulna classification in assessing skeletal maturity: A simplified scheme and reliability analysis," *J. Pediatric Orthopaedics B*, vol. 24, no. 6, pp. 546–551, 2015.
[13] J. P. Y. Cheung, P. W. H. Cheung, D. Samartzis, K. M. C. Cheung, and K. D.-K. Luk, "The use of the distal radius and ulna classification for the prediction of growth: Peak growth spurt and growth cessation," *Bone Joint J.*, vol. 98, no. 12, pp. 1689–1696, 2016.
[14] J. P. Y. Cheung, P. W. H. Cheung, D. Samartzis, and K. D.-K. Luk, "Curve progression in adolescent idiopathic scoliosis does not match skeletal growth," *Clin. Orthopaedics Rel. Res.*, vol. 476, no. 2, pp. 429–436, 2017.
[15] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, Mar. 2016.
[16] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016.
[17] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.
[18] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.
[19] H. Lee *et al.*, "Fully automated deep learning system for bone age assessment," *J. Digit. Imag.*, vol. 30, no. 4, pp. 427–441, 2017.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[21] A. Tristán-Vega and J. I. Arribas, "A radius and ulna TW3 bone age assessment system," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 5, pp. 1463–1476, May 2008.
[22] J. Liu, J. Qi, Z. Liu, Q. Ning, and X. Luo, "Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method," *Comput. Med. Imag. Graph.*, vol. 32, no. 8, pp. 678–684, 2008.
[23] K. Somkantha, N. Theera-Umpon, and S. Auephanwiriyakul, "Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression," *J. Digit. Imag.*, vol. 24, no. 6, pp. 1044–1058, 2011.
[24] H.-H. Lin, S.-G. Shu, Y.-H. Lin, and S.-S. Yu, "Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation," *Pattern Recognit.*, vol. 45, no. 1, pp. 322–332, 2012.
[25] J. Seok, B. Hyun, J. Kasa-Vubu, and A. Girard, "Automated classification system for bone age X-ray images," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2012, pp. 208–213.
[26] L. M. Davis, B-J. Theobald, and A. Bagnall, "Automated bone age assessment using feature extraction," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.*, 2012, pp. 43–51.
[27] M. Harmsen, B. Fischer, H. Schramm, T. Seidl, and T. M. Deserno, "Support vector machine classification based on correlation prototypes applied to bone age assessment," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 190–197, Jan. 2013.
[28] P. Cunha, D. C. Moura, M. A. G. López, C. Guerra, D. Pinto, and I. Ramos, "Impact of ensemble learning in the assessment of skeletal maturity," *J. Med. Syst.*, vol. 38, no. 9, p. 87, 2014.
[29] T. Ebner, D. Stern, R. Donner, H. Bischof, and M. Urschler, "Towards automatic bone age estimation from MRI: Localization of 3D anatomical landmarks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2014, pp. 421–428.
[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[31] R. Girshick. (2015). "Fast R-CNN." [Online]. Available: https://arxiv.org/abs/1504.08083
[32] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[34] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[35] G. E. Güraksın, H. Haklı, and H. Uğuz, "Support vector machines classification based on particle swarm optimization for bone age determination," *Appl. Soft Comput.*, vol. 24, pp. 597–602, Nov. 2014.

SHUQIANG WANG (M'15) received the Ph.D. degree in system engineering and engineering management from the City University of Hong Kong in 2012. From 2012 to 2013, he was a Research Scientist with Huawei Technologies Noah's Ark Lab. From 2013 to 2014, he was Post-Doctoral Fellow with The University of Hong Kong. He is currently an Associate Professor with the Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. His current research interests include machine learning, medical image computing, and optimization theory.

YANYAN SHEN (M'12) received the B.S. and M.Eng. degrees in electrical engineering from Yanshan University, Qinhuangdao, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong, in 2012. From 2013 to 2014, she was a Post-Doctoral Research Fellow with the School of Information and Communication Engineering, Inha University, South Korea. She has been the principle/co-investigator in several research projects funded by NSFC and Shenzhen Basic Research Foundation. She is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her current research interests include optimization methods and machine learning in wireless networks.



CHANGHONG SHI received the Ph.D. degree from the School of Mathematics, Sun Yat-sen University, in 2013. He is currently an Assistant Professor with the Department of Statistical Science, Guangzhou Medical University. His current research interests include machine learning, bioinformatics, and biostatistics.

PENG YIN received the B.Sc. degree from the Department of Statistics and Finance, University of Science and Technology of China, in 2009, and the Ph.D. degree from the Department of Mathematics and Statistics, Newcastle University, U.K., in 2014. From 2014 to 2017, he was a Post-Doctoral Fellow with the University of Liverpool. He is currently an Associate Professor with the Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. His current research interests include big data, machine learning, medical statistics, and bioinformatics.

ZUHUI WANG received the B.E. degree in computer science and technology from Southern Medical University, China, in 2009. He is currently a Research Member with the Center for High Performance Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. His research interests include computational photography, computer vision, and machine learning on medical imaging.

PRUDENCE WING-HANG CHEUNG received the B.D.Sc. degree from The University of Melbourne, Australia, in 2001. She is currently a Senior Research Assistant with the Department of Orthopaedics and Traumatology, The University of Hong Kong. Her main research interests include skeletal maturity in scoliosis patients, developmental spinal stenosis, and health-related quality of life of patients.

JASON PUI YIN CHEUNG received the MBBS degree and the master's degree in medical sciences from The University of Hong Kong in 2007 and 2012, respectively. He completed the membership examination in 2009 and the fellowship examination in 2014. After completing his MBBS degree, he received his training in orthopedics at Queen Mary Hospital. He then joined the Department of Orthopedics and Traumatology, The University of Hong Kong, as a Clinical Assistant Professor, in 2012. His main interests in research are pediatric growth and spinal deformity, lumbar spinal stenosis, and novel imaging.

KEITH DIP-KEI LUK is currently the Chair Professor at the Department of Orthopaedics and Traumatology, The University of Hong Kong. His research interests include intraoperative spinal cord monitoring, spinal biomechanics, spinal deformity correction, genetics of scoliosis, and intervertebral disc transplantation.

YONG HU (SM'12) received the B.Sc. and M.Sc. degrees in biomedical engineering from Tianjin University, Tianjin, China, in 1985 and 1988, respectively, and the Ph.D. degree from The University of Hong Kong in 1999. He is currently an Associate Professor and the Director of the Neural Engineering and Clinical Electrophysiology Laboratory, Department of Orthopedics and Traumatology, The University of Hong Kong. His research interests include neural engineering, clinical electrophysiology, and biomedical signal measurement and processing.