An Ensemble based Densely-Connected Deep Learning System for Assessment of Skeletal Maturity

Shuqiang Wang*, Xiangyu Wang*, Yanyan Shen, Bing He, Xinyan Zhao,
Prudence Wing-Hang Cheung, Jason Pui Yin Cheung, Keith Dip-Kei Luk and Yong Hu

***Abstract*—Assessment of skeletal maturity is important for a clinician to make decision of the most appropriate treatment on various skeletal disorders. This task is very challenging when using machine learning method due to the limited data and large anatomical variations among different subjects. In this paper, we propose an ensemble based deep learning pipeline to automatically assess the distal radius and ulna (DRU) maturity from left hand radiographs. At the same time, we adapted the concept of densely connected mechanism in the proposed network architecture to reuse features and prevent gradient disappearance. Therefore, the model acquires two convincing advantages: first, our model preserves the maximum information flow and has a much faster convergence rate. Second, our model avoids overfitting even if training with limited data. The experimental dataset contains 1189 left-hand X-ray scans of children and teenagers. The proposed method achieves 85.27% and 91.68% for radius and ulna classification respectively. Extensive experiments prove that our model performs better than using other network structures.**

***Index Terms*—dense connection, ensemble learning, convolutional neural network, skeletal maturity**

I. INTRODUCTION

Assessment of skeletal maturity is an important instrumentality in managing children and adolescent's growth problems and helps the physician to find the optimal time to treat the disease. Two clinical methods are widely employed for skeletal maturity assessment: (1) Greulich and Pyle (G & P) method [1]. (2) Tanner-Whitehouse (TW1, TW2, TW3) method [2–4]. The G & P method is based on the comparison between the X-ray scan and the image set included in the atlas, but the subjectivity of different observers makes the diagnosis result unstable. The TW method gives a numerical score to every bone in the bone set, then add the score of all bones to evaluate the bone maturity. However, for the treatment of some specific diseases, such as adolescent idiopathic scoliosis (AIS), these methods are difficult to implement in the outpatient clinical setting and they are time-consuming [5].

Additionally, recognition of skeletal radiological images highly depends on the physician's analysis and it often takes too much time and cost. In recent years, image processing technology has made great progress in many fields [6–8]. Using image processing technology to analyze bone images can solve this problem. Computer-aided-diagnosis (CAD) method [9, 10] has been applied to many medical image processing tasks [11, 12]. In the early years, majority related works about CAD for bone age assessment are based on machine learning method. Zhang *et al.* [13] evaluated the bone age by using fuzzy classification based on the features extracted from radiographic images. Some related works [14, 15] adopted fuzzy logic to assess skeletal maturity. Seok *et al.* [16] utilized decision rules to evaluate the bone age. Mahmoodi *et al.* [17, 18] used Bayesian estimator for skeletal growth estimation. Kashif *et al.* [19, 20], Harmsen *et al.* [21] and Güraksın *et al.* [22] used support vector machines (SVM) for bone age assessment. Some related works [23, 24] utilized shallow neural networks to implement bone age assessment.

In order to avoid the hand-craft feature extraction process in machine learning algorithms, deep-learning solutions have been used in medical imaging applications in recent years. For instance, Chen *et al.* [25] and Lee *et al.* [26] employed convolutional neural network (CNN) to assess bone age by using palm and wrist radiographic images. Zhou *et al.* [27] used deep convolutional neural networks (DCNNs) and transfer learning for bone age classification. Spampinato *et al.* [28] proposed an automated bone age assessment (BAA) system by using CNN and Lee *et al.* [29] added a pretreatment module to the BAA system. Bian *et al.* [30] employed GoogleNet to assess bone age. Mutasa *et al.* [31] designed a 14 hidden layer-customized neural network to assess the bone age of adolescents. The model employs several some techniques including residual-style connections, inception layers, and spatial transformer layers. The mean absolute errors (MAE) for young females and older females were 0.561 years and 0.497 years respectively. For young males and older males, The MAEs were 0.585 years and 0.501 years respectively. Iglovikov *et al.* [32] studied bone age regression and classification using two VGG-style CNNs.The

The first two authors contributed equally to this work.
Corresponding author: Yong Hu (yhud@hku.hk)
This work was supported by National Natural Science Foundations of China under Grant No.61872351 and 61771465, International Science and Technology Cooperation Projects of Guangdong under Grant No. 2019A050510030, Strategic Priority CAS Project under Grant No. XDB38000000Major Projects from General Logistics Department of Peoples Liberation Army under Grant No. AWS13C008 and Shenzhen Key Basic Research Projects under Grant No. JCYJ20200507182506416 and JCYJ20180507182506416.
S. Wang, Y. Shen and B. He are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong Province 518000, P. R. China, Email: sq.wang@siat.ac.cn.

X. Wang and X. Zhao are with School of Data Science, University of Science and Technology of China, Hefei 230026, China.

P. Cheung, J. Cheung, K. Luk and Y. Hu are with Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong Email: yhud@hku.hk.

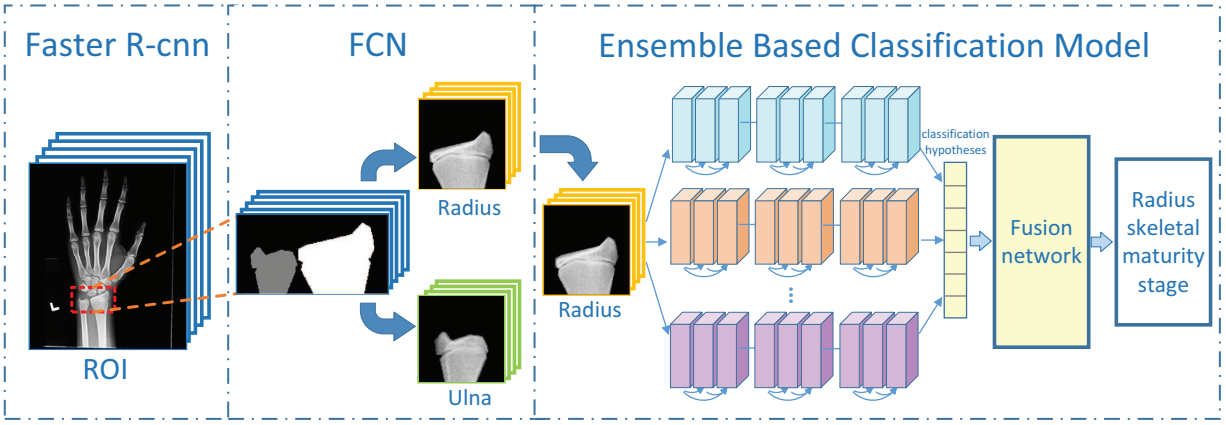


Fig. 1: Framework of our skeletal maturity assessment system.

MAEs of classification task are 6.16 months and 6.39 months for males and females respectively. Keatmanee *et al.* [33] proposed an automatic bone age assessment system by using CNNs. They evaluated various pre-trained models including VGG-16, ResNet-50 and Inception-V3. The obtained MAE is 6.53 months. The results indicate that VGG-16 performs better than others. Gou *et al.* [34] employed transfer learning and regression learning to evaluate bone age. The root mean squared errors are 0.70 year and 0.75 year for male and female cohorts respectively. Besides, the incremental learning based framwork *et al.* [35] and regression based CNN [36–38] were also used to estimate bone age using radiograph.

These bone age assessment systems mentioned above are mainly based on CNN structure. However, constructing very deep CNNs by simply stacking convolutional layers is not feasible, because too many layers can significantly impede the propagation of gradients, which is known as the vanishing-gradient problem. To solve this problem, related publications proposed some solutions, such as residual networks (ResNet) [39], highway networks [40] and densely connected convolutional networks (DenseNet) [41]. These methods shared a common idea: propagate the information by creating a short path from early layers to later layers. In these methods, DenseNet achieved better performance. DenseNet raises the efficiency of information propagation by connecting one layer's output to all the following layers. To take advantage of this mechanism, in this work, we present a model that automatically identifies the skeletal maturity stage of ulna and radius by using dense connection mechanism. Our model consists of three parts: (1) We utilize faster region-based convolutional network (Faster R-CNN) [42] to detect the region of interest (ROI) from radiological images. (2) For each ROI image, we segment the radius and ulna by using fully convolutional network (FCN) [43]. (3) We assess the radius and ulna skeletal maturity stage by using an ensemble DenseNet structure. Our test accuracy achieves 85.27% for radius maturity stage classification and 91.68% for ulna maturity stage classification.

In this study, we combine the dense connection mechanism with the ensemble model to improve the stability and accuracy of skeletal maturity assessment system. Considering accuracy, our model achieves the best performance compared with other mainstream neural network models. The rest of this paper is organized as follows. In Section II, we are going to explain the details of our framework. Section III will present the experimental results and analysis. We will discuss our experimental results in Section IV.

II. METHODS

The original images vary considerably in image size, palm position, intensity and contrast. Noise caused by the background exerts a strong interference on the ulna/radius classification. In order to obtain accurate assessment results, we use ROI detection and pixel-level segmentation to eliminate unnecessary noise and getting a purer image of ulna/radius. The framework of our skeletal maturity assessment system is shown in Fig.1.

A. Region of interest detection

Location of ulna/radius varies from a few pixels to a few hundred pixels in each image. In order to increase the proportion of valid information of the image which inputs to the classification module, the region of ulna and radius needs to be extracted. Faster-R-CNN is an efficient method to detect the region of interest (ROI). We cropped the original image to the same size to normalize the input image size, then we utilized trained Faster R-CNN to detect ulna and radius area.

B. Pixel-level segmentation

Bones except the ulna/radius and the background's noise in the ROI images still interference the accuracy of classification, we need to further remove external noise to make the ulna/radius's information more prominent. Pixel-level segmentation can greatly remove the noise in ROI images. In this task, we employed Faster R-CNN model to separate the ulna/radius from other bones and background noise.

In this work, the Faster R-CNN is first pre-trained on PASCAL VOC-2012 Dataset. It is trained by back-propagation and stochastic gradient descent (SGD) with mini-batch scheme. The initial learning rate is set as 0.01 and all layer weights are randomly initialized from a Gaussian distribution. The first step is to train the region proposal network (RPN) with

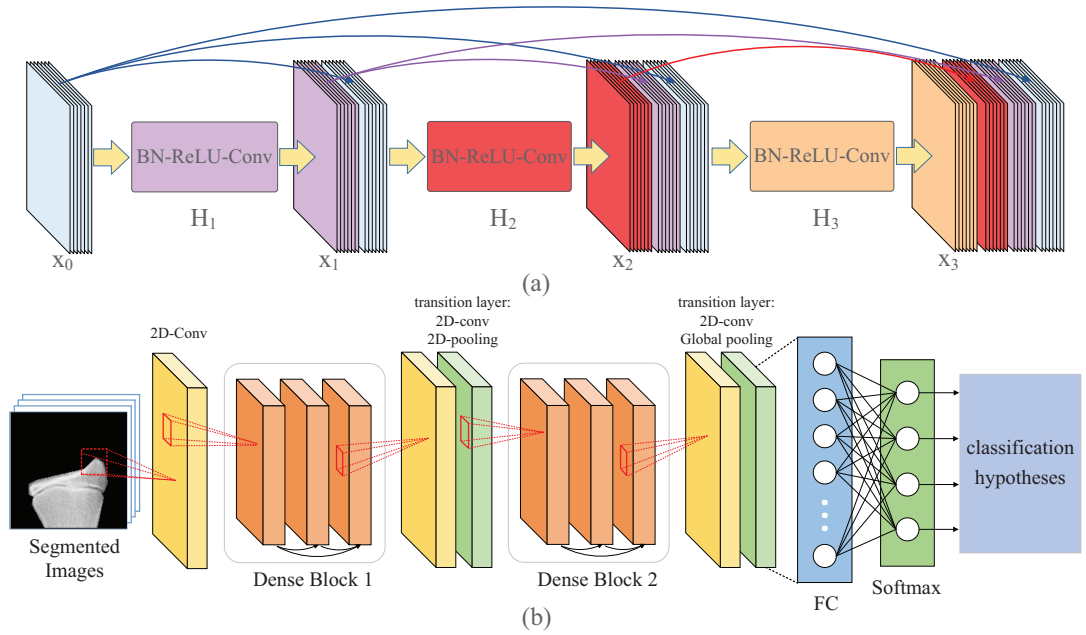


Fig. 2: (a) Constitution of dense connectivity with a 3-layer dense block. (b) The architecture of an individual DenseNet model which consists of two dense blocks.

8000 iterations, which is followed by training Fast R-CNN with 4000 iterations in the second step. Then, the third step is to train RPN with another 8000 iterations and finally the fourth step to train Fast R-CNN with another 4000 iterations. After the pre-train, the model is fine-tuned using our dataset including 400 left-hand X-ray scans. In this work, we employ LabelMe [44], a free annotation tool, to label the 400 left-hand X-ray scans. After that, we used the trained model to segment ROI images, we input all the ROI image into the trained model, the model output the category of each pixel. Then we masked out the background and other bones to get an image that only contains ulna/radius, as given in Fig. 1.

C. Data augmentation

Deep learning method requires a large number of training samples, the limited training set will cause overfitting and fail to generalize for application. Due to the high cost of medical image annotation and data collection, obtaining large-scale training set is a widespread challenge in medical image processing field. Data augmentation is a common and effective way to expand the training set and decrease the risk of overfitting. We increased the size of the training set with several geometric transformations, including horizontal flips, rotation (range from -20° to $+20^{\circ}$), horizontal scaling and vertical scaling by multiplying random numbers between 0.8 to 0.12.

D. Ensemble based densely-connected maturity stage classification model

The main idea of dense connection mechanism is to connect one layer's output to the following layers in a block. an advantage of this mechanism is that the gradient can flow directly from later layers to the earlier layers, which can prevent gradient disappearance. At the same time, the introduction of bottleneck layers can effectively reduce the number of redundant features. Define x_l as the output of l^{th} layer, x_l can be expressed by the following formula:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{1}$$

Where $[x_0, x_1, \ldots, x_{l-1}]$ represent the connectivity of the output produced in $0^{th} \ldots (l-1)^{th}$ and $H_l(\cdot)$ represent the non-linear transformation used in the l^{th} layer. The $H_l(\cdot)$ is composited by three consecutive operations: batch normalization followed by a rectified linear unit (ReLU) [45], and a convolution layer. Fig. 2(a) illustrates the basic framework of a dense block, each convolution layer includes k 3×3 convolution kernels. If each function H_l produces k feature-maps as output, the number of feature-maps that input to the l^{th} layer can be calculated as follows:

$$a_{l^{th}} = k \times (l-1) + k_0 \tag{2}$$

where k_0 is the number of channels of the input image. The whole network is divided into several densely-connected blocks, each block refers to as the dense block. Layers between dense blocks are composed of a batch normalization layer and a 1×1 convolutional layer followed by an average pooling layer, refer to as the transition layer in this model. To reduce the number of feature-maps at transition layer, 1×1 convolution kernel is used to reduce the input of $H_l(\cdot)$, if a dense block contains m feature-maps, the following transition layer generates $\lfloor \theta m \rfloor$ output feature maps, θ is referred to as the compression rate in this work. Fig. 2(b) illustrates the structure of an individual DenseNet model which consists of two dense blocks. Before entering the first dense block, a convolution is performed on the input images, at the end

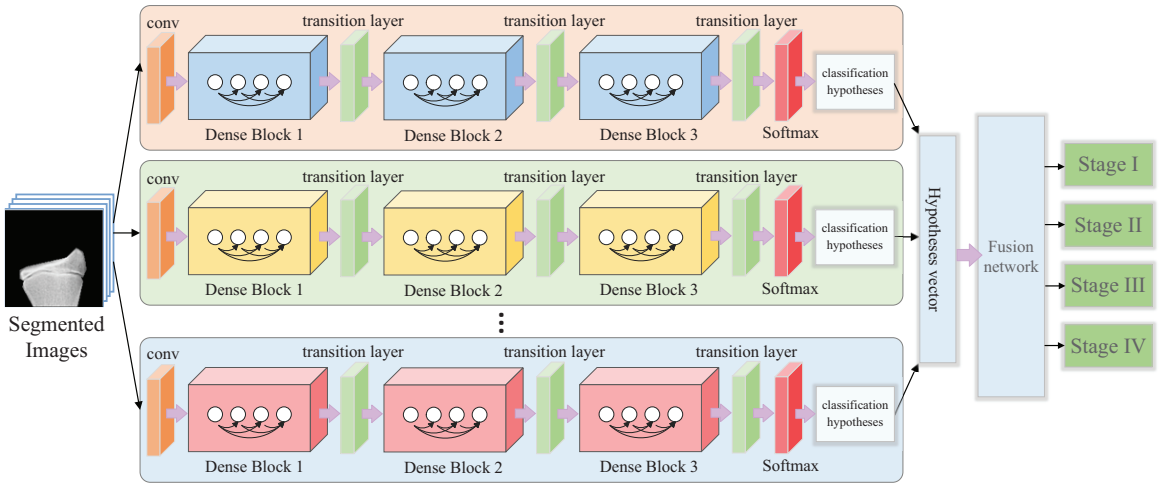


Fig. 3: The architecture of our proposed ensemble classification model.

of the last dense block, a global average pooling and a FC layer is performed, then a softmax classifier is implemented for classification.

In order to improve the accuracy of the classification model, we proposed an ensemble based densely-connected classification model to identify the maturity stage of ulna/radius. The model consists of several independent DenseNet models, each DenseNet model has different structure, and each DenseNet model classifies the input x_k to a particular maturity stage respectively. Classifier $DenseNet_i$ can be seen as a nonlinear function $f_i(\cdot)$, define $f_i(x_k)$ is the output of global average pooling layer in $DenseNet_i$. A conditional tag probability $p_k^i(y=c|f_i(x_k))$ by applying a softmax operation is defined as follows:

$$p(y=c\,|f_i(x_k)) = \frac{e^{f_i(x_k)^T w_{c_i}+b_{c_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}} \tag{3}$$

where $c \in 1, ..., M$, M indicates the total number of classes, w and b are the parameters of the FC layer. The probabilities assigned to categories by $DenseNet_i$ can be denoted as:

$$P_k^i = (\frac{e^{f_i(x_k)^T w_{1_i}+b_{1_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}, \frac{e^{f_i(x_k)^T w_{2_i}+b_{2_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}, ..., \frac{e^{f_i(x_k)^T w_{M_i}+b_{M_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}) \tag{4}$$

each individual model $DenseNet_i$ proposes corresponding classification hypothesis for input x_k, which can be denoted as follows:

$$S_k^i = \arg\max(\frac{e^{f_i(x_k)^T w_{1_i}+b_{1_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}, \frac{e^{f_i(x_k)^T w_{2_i}+b_{2_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}, ..., \frac{e^{f_i(x_k)^T w_{M_i}+b_{M_i}}}{\sum_{m=1}^{M} e^{f_i(x_k)^T w_{m_i}+b_{m_i}}}) \tag{5}$$

The ensemble model combines the outputs of these networks by training a fusion network. The fusion network automatically learns the weight between each classifier, and it is constructed by a fully connected neural network. When input the k^{th} sample x_k to the ensemble classification model, each classifier $DenseNet_i$ outputs the corresponding classification hypothesis of x_k. After combining the output of each classifier, we can get a classification hypotheses vector X_k which can be expressed as:

$$\mathrm{X}_k = \left(\mathrm{S}_\mathrm{k}^1, \mathrm{S}_\mathrm{k}^2, \ldots, \mathrm{S}_\mathrm{k}^\mathrm{i}, \ldots, \mathrm{S}_\mathrm{k}^\mathrm{I}\right) \tag{6}$$

where I indicates the total number of Densenet classifiers. The fusion network can be seen as a function $h_{weight}(\cdot)$ which combines the result of each classifier and assign corresponding weights to each classifier. The last FC layer's output can be expressed as $h_{weight}(\mathrm{X}_k)$. The final output of our ensemble classification model can be expressed as follows:

$$Y_{predict} = \arg\max(\frac{e^{h_{weight}(X_k)^T w_1+b_1}}{\sum_{m=1}^{M} e^{h_{weight}(X_k)^T w_m+b_m}}, \frac{e^{h_{weight}(X_k)^T w_2+b_2}}{\sum_{m=1}^{M} e^{h_{weight}(X_k)^T w_m+b_m}}, ..., \frac{e^{h_{weight}(X_k)^T w_M+b_M}}{\sum_{m=1}^{M} e^{h_{weight}(X_k)^T w_m+b_m}}) \tag{7}$$

the fusion network is trained to minimize the negative log-likelihood of the prediction, the loss function is computed as follows:

$$L = -\sum_{c=1}^{M} Y_c \log \frac{e^{h_{weight}(X_k)^T w_c+b_c}}{\sum_{m=1}^{M} e^{h_{weight}(X_k)^T w_m+b_m}} \tag{8}$$

where Y_c indicates the true probability that the sample belongs to category c.

Traditional ensemble algorithms are mostly based on weighted averaging and majority voting method which only consider the linear relationship between classifiers and most of them provide weights by artificial experience. Generally, the relationship between the labels of test samples and the multiple classifiers is unknown, only considering the linear relationships can't guarantee the reliability of the prediction. Our approach automatically learns the intricate weight relationships between classifiers and the labels of test samples, especially nonlinear relationships. Fig. 3 shows the structure of our ensemble classification model.

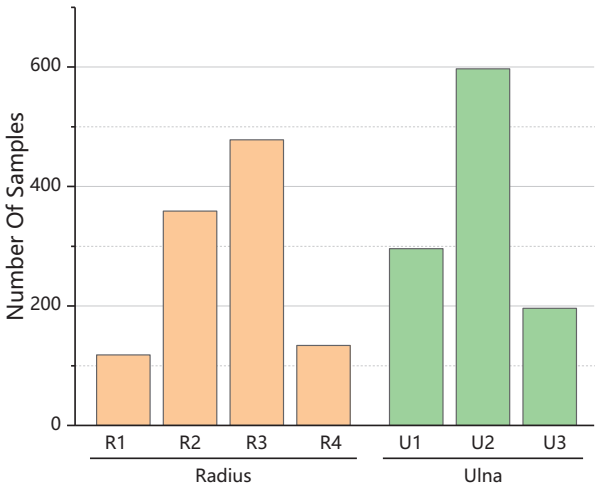


Fig. 4: Samples distribution of radius and ulna.

III. EXPERIMENTS AND RESULTS

A. Data and Implementation

Our dataset contains 1189 left-hand X-ray scans of children and teenagers (0 to 17 years old), from Duchess of Kent Children's Hospital, Pokfulam, Hong Kong, as shown in Fig.5. To validate the generalization capability of the proposed model, we used 5-fold cross validation scheme. For each fold, we used 70% of data for training, 10% of data for validation and 20% of data for test. Besides, 4 times data augmentation is employed. There are four growth stages for radius and three growth stages for ulna. The distribution for each class is shown in Fig. 4. Examples of each stage of radius and ulna images is shown in Fig. 6. In the following parts, we will introduce the morphological characteristics of each growth stage.

(1) Skeletal maturity stage of radius.

Stage1. The width of epiphysis is not as wide as metaphysis (Fig. 6(a)). This stage mainly appears in prepubertal phase (before 9.3 ± 1.5 years old).

Stage2. Physeal plate clearly visible, slight separation appears between medial part of epiphysis and the metaphysis, lateral border wider than metaphysis, standing height, long bone length and arm span at the growth peak during this stage (11.4 ± 2.1 years old), as shown in Fig. 6(b).

Stage3. The metaphysis is fully capped by epiphysis, the growth plate disappears and forming a sclerotic line. At this stage, the standing height stops growing. This stage always happens after the peak height velocity (PHV), most girls present with menarche at this stage. This stage is the peak of sexual development (mean age about 14.75 ± 2.0 years old), as shown in Fig. 6(c).

Stage4. The physeal line completely fuses with the lateral edge and the medial edge, the growth plate scar barely visible. At this stage, the arm span and the radial length stops growing, most adolescents present sexual maturation, this stage indicates the cessation of skeleton growth (17.3 ± 1.1 years old), as shown in Fig. 6(d).

(2) Skeletal maturity stage of ulna

Stage1. The width of epiphysis is not as wide as metaphysis (Fig. 6(e)). A styloid can be found on the medial end of the epiphysis. Standing height, sitting height and long bone length are at the growth peak during this stage (before 11.0 ± 1.4 years old).

Stage2. The epiphysis has the same width with the metaphysis, the medial physeal plate is narrow but unfused. The standing height, sitting height stops growing at this stage. This stage is the peak of sexual development, most girls present with menarche at this stage (13.4 ± 2.5 years old), as shown in Fig. 6(f).

Stage3. Growth plate scar is barely visible. The arm span and radial length stop growing at this stage. Most adolescents

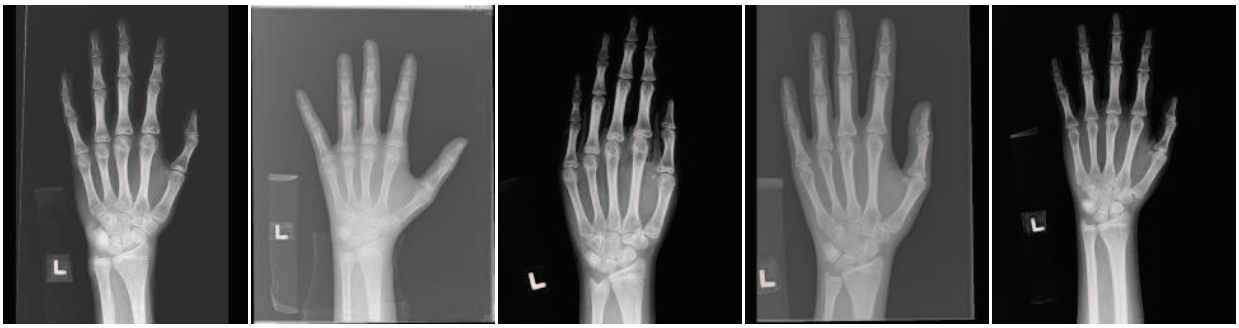
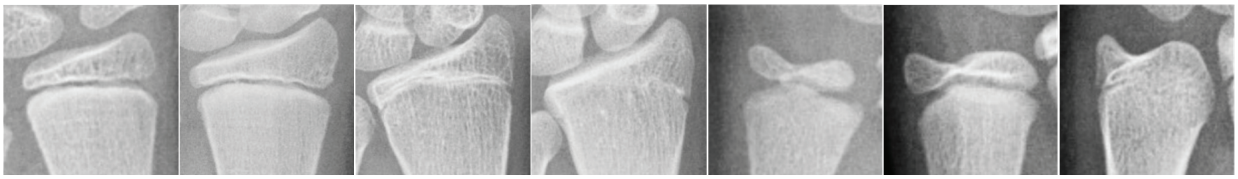


Fig. 5: Images of the original dataset.



(a) (b) (c) (d) (e) (f) (g)

Fig. 6: Examples of each skeletal maturity stage of radius and ulna.

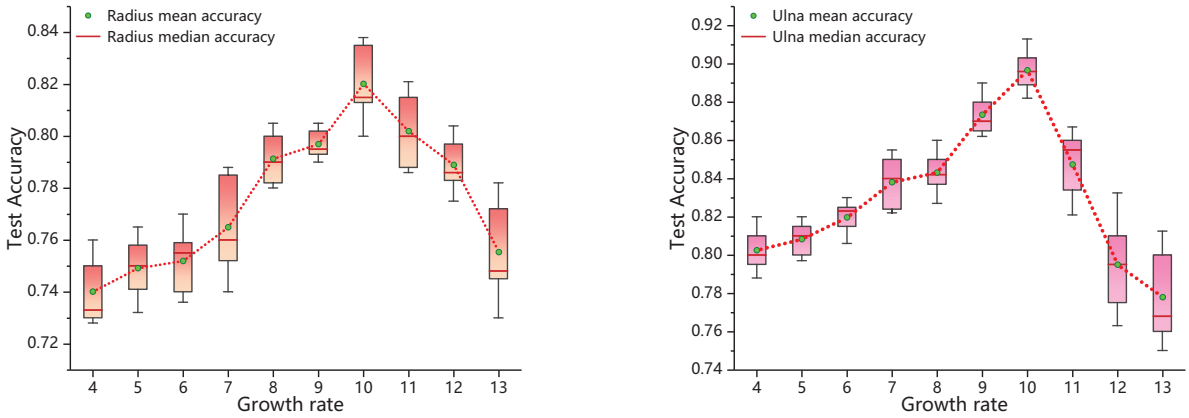


Fig. 7: Radius and ulna classification results with different growth rates.

present full sexual maturation. This stage indicates the cessation of skeleton growth (16.3 ± 1.3 years old), as shown in Fig. 6(g).

Our model was implemented in TensorFlow [46], we conducted experiments on a workstation with 12GB NVIDIA TESLA K40m GPU and the cuDNN library (v5) has been used for GPU acceleration.

B. Parameter analysis

In order to find well-performing individual DenseNet structures to build our ensemble classification model, we analyze the effects of different hyper-parameters on the classification result of individual model. The individual models were optimized with Nesterov momentum algorithm. In the following parts, we will present and analyze our experimental results.

Growth rate.

Growth rate controls the number of feature maps that each layer contributes to the global state. We fixed other hyper-parameters and tested the influence of growth rate on the accuracy; small changes of the growth rate can cause apparent fluctuation on the test accuracy. For different DenseNet structures, the accuracy of the model shows a convex function trend when growth rate increases, as shown in Fig. 7. Too large or too small growth rate will decrease the ulna/radius test accuracy, the best growth rate for radius/ulna classification is around 10.

Depth.

The influence of depth on the model performance is shown in Fig. 8. Under the same conditions ($growth\ rate = 10$, $blocks = 3$), when the number of blocks is fixed, change the total depth of the network means change the number of layers in each block. For radius classification, the accuracy of the model shows a convex function trend with depth increases. For ulna classification, test accuracy shows a negative linear relationship with depth and the model performs better with lower depth compared with higher depth. For the classification of radius/ulna, using deeper networks is more likely to occur overfitting phenomenon. The best depth for radius classification is about 16, and for ulna classification the best depth is about 10.

Number of blocks.

The transition layers between blocks implement the down-sampling operation. When the number of blocks increases means more down-sampling operation to feature-maps. An appropriate sampling degree can make the network perform better. We fixed other hyper-parameters to observe the effect of the number of blocks. It is evident that when the number of blocks smaller than 3, the accuracy increases at a high rate of speed. When the number of blocks around 3, we get the best accuracy for radius/ulna classification, as given in Fig. 9.

Compression rate.

We tested different compression rate in our model, as shown in Fig. 10. There presents the highest test accuracy when compression rate around 0.7. When model without compression operations ($compression\ rate = 1.0$), the accuracy is lower than the model which with appropriate compression operation. Compression operation can reduce the number of features, too many features may lead to overfitting. Discarding an appropriate number of features can improve the accuracy and generalization ability.

C. Model performance

According to sufficient experiments, we set $blocks = 3$, $depth = 16$, $compression\ rate = 0.7$ and $growth\ rate = 10$ for radius classification; for ulna classification, we set $blocks = 3$, $depth = 10$, $compression\ rate = 0.7$ and $growth\ rate = 10$, these structures are the optimal network structure we obtained for individual Densenet models, when other variable factors are under optimal conditions, the accuracy achieves 83.33% and 90.31% for radius and ulna classification respectively by using individual DenseNet model. Fig. 11 shows the trend of training accuracy and test accuracy when the number of iterations changes.

To acquire the best performance ensemble classification model, we selected five outstanding individual DenseNet structures from our sufficient experiments to build the ensemble model, the structure of each sub-classifier is shown in Table I. At the same time, we utilized a three-layer fully connected network as our fusion network.

We evaluated the performance of our individual model and the ensemble model for radius/ulna classification. The model

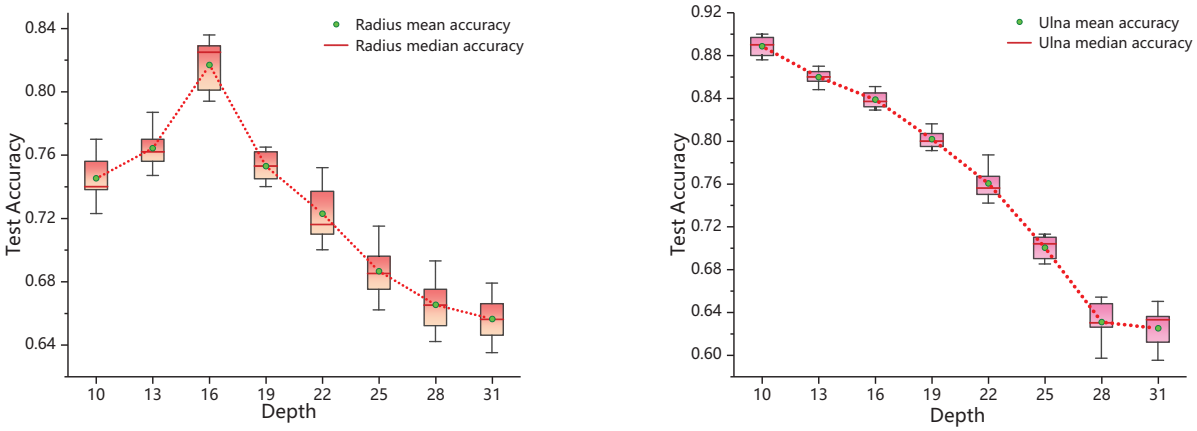


Fig. 8: Radius and ulna classification results with different depths.

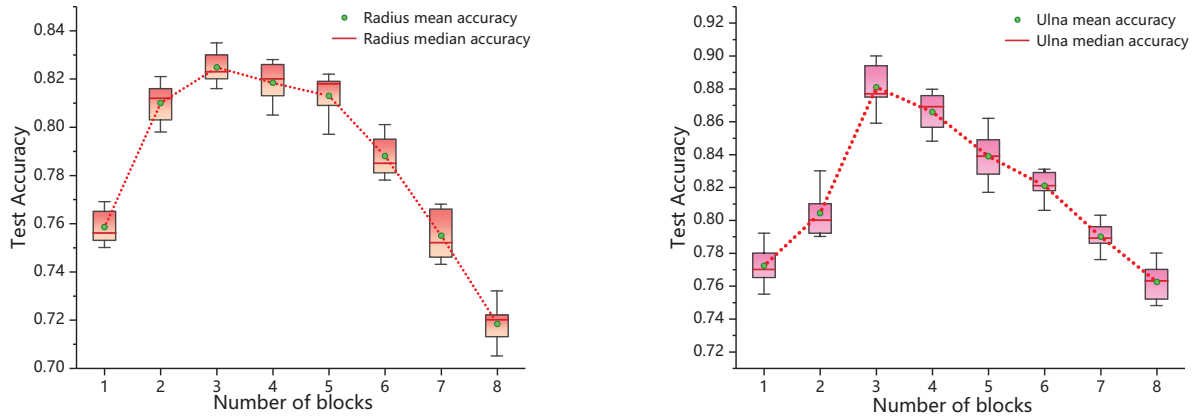


Fig. 9: Radius and ulna classification results with different number of blocks.

TABLE I: Structure of each sub-classifier.

Structure	Block	Depth	Growth rate	Structure	Block	Depth	Growth rate
DensNet-R-A	3	16	10	DensNet-U-A	3	10	10
DensNet-R-B	4	13	9	DensNet-U-B	3	13	9
DensNet-R-C	2	13	11	DensNet-U-C	4	13	11
DensNet-R-D	3	13	10	DensNet-U-D	3	17	9
DensNet-R-E	4	17	9	DensNet-U-E	2	11	11

is evaluated by three evaluation indicators: recall, precision and F1-score. The precision calculate formula as follows:

$$\text{Precision} = \frac{\text{Samples correctly classified as c}}{\text{Samples classified as c}} \tag{9}$$

The recall calculate formula:

$$\text{Recall} = \frac{\text{Samples correctly classified as c}}{\text{Samples of class c}} \tag{10}$$

The F1-score is given by:

$$\text{F1} - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

Table II and Table III show the performance of our best individual Densenet model and the performance of our ensemble model. The ensemble model shows relatively balanced and stable performance for the classification of each class, the ensemble model achieves encouraging performance and outperforms the individual model.

In addition, we compared the accuracy of the ensemble model and the accuracies of these five sub-classifiers which constitute the ensemble model, as shown in Table IV. The ensemble model shows outstanding performance, when other variable factors are under optimal conditions, the ensemble model shows the best accuracy with 85.27% for radius classification and 91.68% for ulna classification.

To explore the model performance according to different number of training samples, we compared the accuracy of the ensemble model and the accuracies of the five sub-classifiers when the number of samples changes, as shown in Fig. 12.

We utilized different multiples of data augmentation to change the number of samples, when the number of samples smaller than 4 times of the original dataset, the accuracy

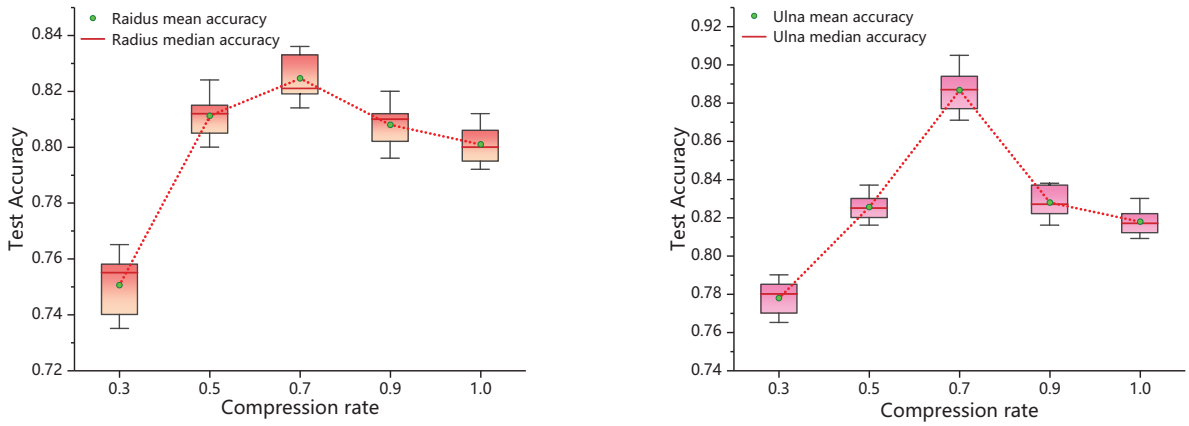


Fig. 10: Radius and ulna classification accuracy with diverse compression rates.

TABLE II: Performance of radius classification results at each stage.

Model	Skeletal Maturity of Radius	Recall	Precision	F1-score
Individual	Stage1	75.31%	75.85%	75.58%
	Stage2	85.12%	83.33%	84.21%
	Stage3	86.01%	81.25%	83.56%
	Stage4	79.65%	74.44%	76.96%
Ensemble	Stage1	81.81%	78.26%	80.00%
	Stage2	87.50%	87.50%	87.50%
	Stage3	87.17%	89.47%	88.27%
	Stage4	83.33%	81.08%	82.19%

increases sharply with the multiples of data augmentation increases. Sufficient experiments prove that 4 times data augmentation is a good choice for our skeletal maturity stage classification task.

In this work, we used confusion matrix to evaluate the classification performance for each category, Fig. 13 shows the confusion matrices for ulna and radius classification. The classification errors are mainly caused by adjacent maturity stages which have high similarity in the skeleton morphology. Fig. 14 (a) shows the image of radius just entering stage4, which is very easily confused with stage3. Similarly, Fig. 14 (b) shows the ulna image of the early stage2, the visual features of which are very similar to stage1. It is also a challenge for the clinician to recognize such samples in practice.

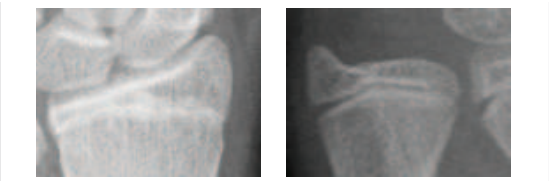


Fig. 14: Examples of radius and ulna images that are misclassified.

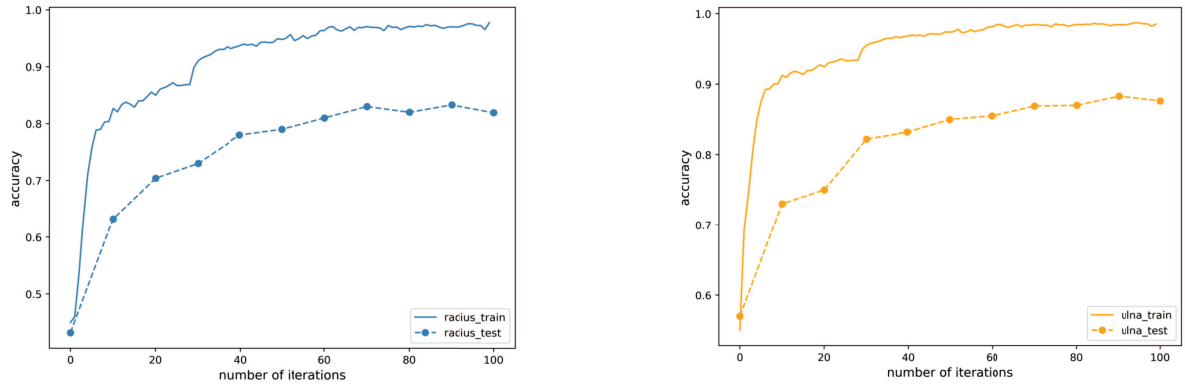


Fig. 11: Accuracy of the optimal individual Densenet model for radius (a) and ulna (b) classification.

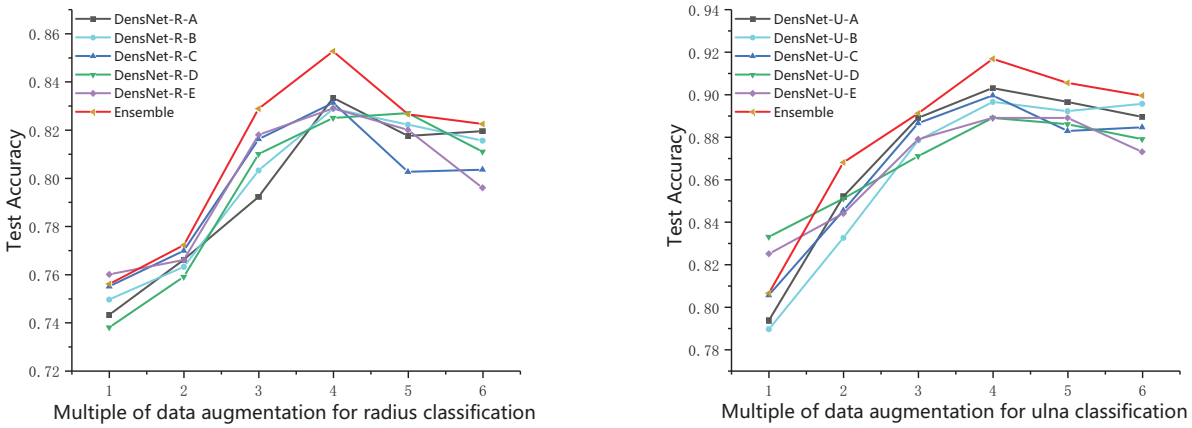


Fig. 12: Classification accuracy of the ensemble model with different size training datasets.

TABLE III: Performance of ulna classification results at each stage.

Model	Skeletal Maturity of Ulna	Recall	Precision	F1-score
Individual	Stage1	79.54%	92.10%	85.36%
	Stage2	96.33%	88.98%	92.51%
	Stage3	85.11%	90.91%	87.91%
Ensemble	Stage1	84.09%	97.36%	90.23%
	Stage2	98.17%	91.45%	94.70%
	Stage3	93.33%	89.36%	91.30%

TABLE IV: Comparison with each individual DensNet models.

Model	Train Acc	Validation Acc	Test Acc	Model	Train Acc	Validation Acc	Test Acc
DensNet-R-A	96.68%	85.45%	83.33%	DensNet-U-A	98.42%	91.55%	90.31%
DensNet-R-B	97.12%	86.02%	82.88%	DensNet-U-B	99.17%	92.32%	89.65%
DensNet-R-C	96.79%	86.26%	83.13%	DensNet-U-C	98.08%	93.33%	89.96%
DensNet-R-D	98.25%	84.13%	82.50%	DensNet-U-D	97.96%	91.71%	88.88%
DensNet-R-E	97.46%	85.33%	82.90%	DensNet-U-E	98.55%	90.97%	88.90%
Ensemble	-	-	**85.27%**	**Ensemble**	-	-	**91.68%**

IV. DISCUSSIONS

A. Comparison with other neural network models

We compare our ensemble model with other neural network models. Moreover, the test accuracy of different methods is summarized in Table V. The optimal CNN structure we obtained from the experiment is 32c7-64c5-64c3-128c3-128c3 ("32c7" means 32 convolution kernels with size of 7). Three observations can be made from our experiments:

1) Our model has faster convergence speed. Note that, only 100 iterations were trained until DenseNet classification model reaches convergence (Fig. 11), but 10000 iterations for LeNet5 (Fig. 15 (a)) and AlexNet (Fig. 15 (b)) reach convergence.

2) Compared with other neural network models, DenseNet has smaller generalization error and it is less prone to overfitting.

3) In the application of skeletal maturity stage classification, our method has higher accuracy compared with other neural network models.

Therefore, two advantages are enjoyed from the mechanism of dense connection. First, due to the gradient flows from later layers to earlier layers, it preserves the maximum information flow between layers, each layer receives additional supervision. Therefore, our model is easier to converge and reaches higher accuracy. Second, compared with other models for ulna/radius classification, DenseNet has less parameters, fewer parameters improve the generalization ability of the model and prevent overfitting even if we have a small train set which is an unavoidable problem in medical image processing.

TABLE V: Comparison with other neural network models.

Method	Accuracy of radius	Accuracy of ulna
LeNet5	68.59%	74.72%
AlexNet	67.88%	75.89%
ResNet	75.39%	84.82%
CNN	78.50%	85.33%
Our Method	**85.27%**	**91.68%**

B. How to improve the system

In the process of experiment, we found the parameter initialization has a significant influence on the model performance.

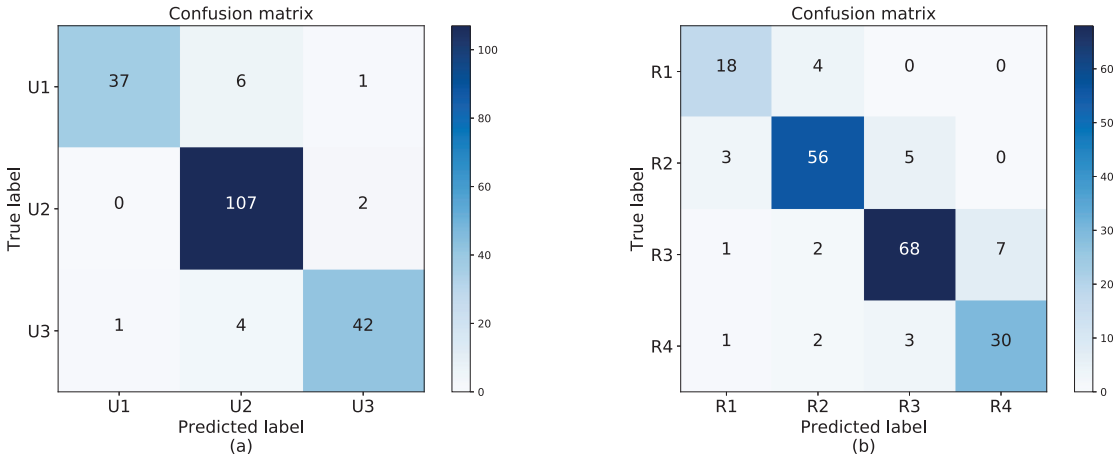


Fig. 13: (a) Confusion matrix for ulna classification. (b) Confusion matrix for radius classification.

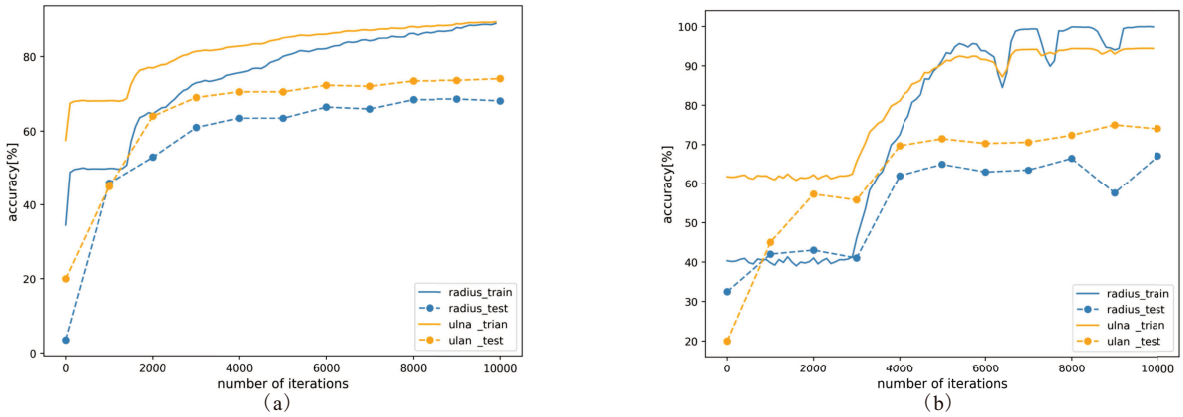


Fig. 15: (a) Accuracy of LeNet5 [47]. (b) Accuracy of AlexNet [48].

Although we have achieved satisfactory results, our parameter was initialized with random weights, this is detrimental to the learning speed of the neural networks. We still have room for improvement to provide more accurate classification results. In the future work, we expect to use the advantage of transfer learning mechanism to overcome the shortcoming of the limit dataset which is a widespread problem in medical image processing field. We plan to pre-train our model by using other datasets, such as CIFAR-10 [49] or ImageNet [50], then use the pre-trained network to build the skeletal maturity classification model.

C. Clinical application

In the treatment of the adolescent disease, it is essential to decide the treatment plan according to the patient's development situation. Study how to auto classify the skeletal maturity is meaningful for guide clinical management [5], the implementation of automatic skeletal maturity identification can reduce the workload of doctors. For example, it has significant clinical application value to quickly identify the skeletal maturity of patient with adolescent idiopathic scoliosis (AIS) and help physicians estimate the growth peak and growth cessation, determine clinical observational intervals as early as possible. In addition, using distal radius and ulna radiographs to assess skeletal maturity can provide the reference for physicians to find the time to initiate or end bracing therapy. At the same time, artificial assessment requires two or more physicians to assess skeletal maturity, this method can give diagnosis recommendations and provide a unified standard for a given examination and reduces the influence of interobserver variability.

V. CONCLUSIONS

In this work, we proposed an ensemble based automatic skeletal maturity assessment system by using dense connection mechanism. The dense connection mechanism was utilized with the ensemble framework to improve the stability and accuracy of skeletal maturity assessment system. A significant quantity of experiments have been done to find the influence of different hyper-parameters on the final result. We also demonstrated the superiority of the proposed model against using other models for skeletal maturity stage classification. Applying dense connectivity, our model may achieve better feature transfer efficiency with fewer parameters [41]. Our model has less chance to encounter the over-fitting problem even if training with limited data and it is much easier to converge and has greater optimization efficiency. The proposed method can be deployed in the clinical environment to help

orthopedist to identify the skeletal maturity via radius and ulna automatically and objectively.

REFERENCES

[1] W. W. Greulich, S. I. Pyle, and T. W. Todd, "Radiographic atlas of skeletal development of the hand and wrist," *Stanford: Stanford university press*, vol. 2, 1959.

[2] J. M. Tanner and R. H. Whitehouse, "Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty." *Archives of disease in childhood*, vol. 51, no. 3, pp. 170–179, 1976.

[3] R. Bull, P. Edwards, P. Kemp, S. Fry, and I. Hughes, "Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods," *Archives of disease in childhood*, vol. 81, no. 2, pp. 172–173, 1999.

[4] R. M. Malina and G. P. Beunen, "Assessment of skeletal maturity and prediction of adult height (tw3 method)," *American Journal of Human Biology*, vol. 14, no. 6, pp. 788–789, 2002.

[5] K. D. 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," *The Spine Journal*, vol. 14, no. 2, pp. 315–325, 2014.

[6] F. Shuang and C. L. P. Chen, "Fuzzy restricted boltzmann machine and deep belief network: A comparison on image reconstruction," *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1828–1833, 2017.

[7] J. Duan, L. Chen, and C. L. P. Chen, "Region-based multi-focus image fusion using guided filtering and greedy analysis," *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2932–2937, 2015.

[8] H. Sima, P. Guo, Y. Zou, Z. Wang, and M. Xu, "Bottom-up merging segmentation for color images with complex areas," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 3, pp. 354–365, 2018.

[9] K. Uemura, Y. Miyoshi, T. Kawahara, S. Yoneyama, Y. Hattori, J.-i. Teranishi, K. Kondo, M. Moriyama, S. Takebayashi, Y. Yokomizo *et al.*, "Prognostic value of a computer-aided diagnosis system involving bone scans among men treated with docetaxel for metastatic castration-resistant prostate cancer," *BMC cancer*, vol. 16, no. 1, p. 109, 2016.

[10] Y. Miyoshi, S. Yoneyama, T. Kawahara, Y. Hattori, J.-i. Teranishi, K. Kondo, M. Moriyama, S. Takebayashi, Y. Yokomizo, M. Yao *et al.*, "Prognostic value of the bone scan index using a computer-aided diagnosis system for bone scans in hormone-naive prostate cancer patients with bone metastases," *BMC cancer*, vol. 16, no. 1, p. 128, 2016.

[11] Y. Chen, J. Hsu, C. Hung, J. Wu, F. Lai, and S. Kuo, "Surgical wounds assessment system for self-care," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–16, 2018.

[12] X. Zhou, G. Bian, X. Xie, and Z. Hou, "An interventionalist-behavior-based data fusion framework for guidewire tracking in percutaneous coronary intervention," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14, 2018.

[13] A. Zhang, A. Gertych, and B. J. Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 299–310, 2007.

[14] M. Mansourvar, A. Asemi, R. G. Raj, S. A. Kareem, C. D. Antony, N. Idris, and M. S. Baba, "A fuzzy inference system for skeletal age assessment in living individual," *International Journal of Fuzzy Systems*, vol. 19, no. 3, pp. 838–848, 2017.

[15] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. Huang, "Bone age assessment of children using a digital hand atlas," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 322–331, 2007.

[16] J. Seok, J. Kasa-Vubu, M. DiPietro, and A. Girard, "Expert system for automated bone age determination," *Expert Systems with Applications*, vol. 50, pp. 75–88, 2016.

[17] S. Mahmoodi, B. S. Sharif, E. G. Chester, J. P. Owen, and R. Lee, "Skeletal growth estimation using radiographic image processing and analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 4, pp. 292–297, 2000.

[18] S. Mahmoodi, B. Sharif, E. Chester, J. Owen, and R. Lee, "Automated vision system for skeletal age assessment using knowledge based techniques," *IET Conference Proceedings*, pp. 809–813(4), 1997.

[19] M. Kashif, S. Jonas, D. Haak, and T. M. Deserno, "Bone age assessment meets sift," *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414, p. 941439, 2015.

[20] M. Kashif, T. M. Deserno, D. Haak, and S. Jonas, "Feature description with sift, surf, brief, brisk, or freak? a general question answered for bone age assessment," *Computers in Biology and Medicine*, vol. 68, pp. 67–75, 2016.

[21] 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 Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 190–197, 2013.

[22] G. E. Graksn, H. Hakl, and H. Uuz, "Support vector machines classification based on particle swarm optimization for bone age determination," *Applied Soft Computing*, vol. 24, pp. 597–602, 2014.

[23] A. Tristan-Vega and J. I. Arribas, "A radius and ulna tw3 bone age assessment system," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 5, pp. 1463–1476, 2008.

[24] M. Mansourvar, S. Shamshirband, R. G. Raj, R. Gunalan, and I. Mazinani, "An automated system for skeletal maturity assessment by extreme learning machines," *PloS one*, vol. 10, no. 9, p. e0138493, 2015.

[25] M. Chen, "Automated bone age classification with deep neural networks," *Stanford University Technical Report*, 2016.

[26] J. H. Lee and K. G. Kim, "Applying deep learning in medical images: The case of bone age estimation," *Healthcare Informatics Research*, vol. 24, no. 1, pp. 86–92, 2018.

[27] J. Zhou, Z. Li, W. Zhi, B. Liang, D. Moses, and L. Dawes, "Using convolutional neural networks and transfer learning for bone age classification," *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on*, pp. 1–6, 2017.

[28] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in x-ray images," pp. 41–51, 2017.

[29] H. Lee, S. Tajmir, J. Lee, M. Zissen, B. A. Yeshiwas, T. K. Alkasab, G. Choy, and S. Do, "Fully automated deep learning system for bone age assessment," *Journal of digital imaging*, vol. 30, no. 4, pp. 427–441, 2017.

[30] Z. Bian and R. Zhang, "Bone age assessment method based on deep convolutional neural network," *2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 194–197, 2018.

[31] S. Mutasa, P. D. Chang, C. Ruzal-Shapiro, and R. Ayyala, "Mabal: a novel deep-learning architecture for machine-assisted bone age labeling," *Journal of digital imaging*, vol. 31, no. 4, pp. 513–519, 2018.

[32] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets, "Paediatric bone age assessment using deep convolutional neural networks," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 300–308, 2018.

[33] C. Keatmanee, S. Klabwong, K. Osatavanichvong, and C. Suchato, "Performance of convolutional neural networks and transfer learning for skeletal bone age assessment," *THE BANGKOK MEDICAL JOURNAL*, vol. 15, no. 1, 2019.

[34] J. Han, Y. Jia, C. Zhao, and F. Gou, "Automatic bone age assessment combined with transfer learning and support vector regression," *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 61–66, 2018.

[35] S. E. Ayala-Raggi, F. M. Manzano, A. Barreto-Flores, S. Sánchez-Urrieta, J. F. Portillo-Robledo, V. E. Bautista-López, and P. Ayala-Raggi, "A supervised incremental learning technique for automatic recognition of the skeletal maturity, or can a machine learn to assess bone age without radiological training from experts?" *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 01, p. 1860002, 2018.

[36] P. Hao, S. Chokuwa, X. Xie, F. Wu, J. Wu, and C. Bai, "Skeletal bone age assessments for young children based on regression convolutional neural networks," *Mathematical Biosciences and Engineering*, vol. 16, no. 6, pp. 6454–6466, 2019.

[37] X. Ren, T. Li, X. Yang, S. Wang, S. Ahmad, L. Xiang, S. R. Stone, L. Li, Y. Zhan, D. Shen *et al.*, "Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph," *IEEE journal of biomedical and health informatics*, 2018.

[38] M. Zhang, D. Wu, Q. Liu, Q. Li, Y. Zhan, and X. S. Zhou, "Multi-task convolutional neural network for joint bone age assessment and ossification center detection from hand radiograph," *International Workshop on Machine Learning in Medical Imaging*, pp. 681–689, 2019.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *Advances in neural information processing systems*, pp. 2377–2385, 2015.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *CVPR*, vol. 1, no. 2, p. 3, 2017.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1137–1149, 2017.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[44] R. B. C. Torralba, A. and J. Yuen, "Labelme: Online image annotation and applications," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1467–1484, 2010.

[45] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, 2011.

[46] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and Isard, "Tensorflow: a system for large-scale machine learning," *OSDI*, vol. 16, pp. 265–283, 2016.

[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems 25*, pp. 1097–1105, 2012.

[49] Y. Yang, Q. J. Wu, and Y. Wang, "Autoencoder with invertible functions for dimension reduction and image reconstruction," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 7, pp. 1065–1079, 2018.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition*, pp. 248–255, 2009.