Bone age assessment method based on fine-grained image classification using multiple regions of interest

Keji Mao\textsuperscript{a}, Wei Lu\textsuperscript{a}, Kunxiu Wu\textsuperscript{a}, Jiafa Mao\textsuperscript{a} and Guanglin Dai\textsuperscript{b}

\textsuperscript{a}College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, People’s Republic of China; \textsuperscript{b}College of Information Engineering, Zhejiang University of Technology, Hangzhou, People’s Republic of China

\textbf{ABSTRACT}

Bone age assessment is commonly used to determine the growth status and growth potential of children. In this paper, the bone age assessment is regarded as a fine-grained image classification problem as bone age assessment is usually performed on radiographs of the left hand. An end-to-end bone age assessment model was proposed. This model is composed of four parts: feature extractor, Region of Interest (ROI) selection subnet, guidance subnet, and assessment subnet. Feature extractor is implemented based on Convolutional Neural Networks (CNNs), ResNet50 was used to extract image features. ROI selection subnet is used to select multiple informative ROIs that contain representative images features in the radiograph. Guidance subnet can guide the ROI selection subnet to select ROI more appropriately. Assessment subnet is used for bone age assessment by utilizing the extracted image features. The proposed model can extract the most informative ROIs in the radiographs, and use these ROIs to improve the accuracy of bone age assessment. In this paper, the bone age assessment model is tested on a public data set. The experimental results show that the proposed bone age assessment model has the highest accuracy, and the Mean Absolute Error (MAE) reaches 6.65 months.

\textbf{ARTICLE HISTORY}

Received 20 September 2021
Accepted 12 December 2021

\textbf{KEYWORDS}

Bone age assessment; fine-grained image classification; region of interest; convolutional neural network

1. Introduction

Bone age can reflect the maturity of the individual’s skeleton. Compared to chronological age, bone age can better reflect the growth status and growth potential of an individual, especially for children in critical growth spurts. The final height of an individual can be predicted by combining one’s current height and bone age. Therefore, bone age is widely used in the selection of athletes. In addition, bone age is often used as one of the diagnostic criteria for some endocrine diseases and genetic diseases as these diseases can lead to abnormal bone age. Given the high application value of bone age, bone age assessment has been widely studied.

Bone age assessment is usually performed on the radiographs of the left hand. Traditional bone age assessment methods are performed manually, experts make bone age assessments by analyzing the shape, size, texture, and other characteristics of the bones on the radiograph. Traditional bone age assessment methods are usually divided into three categories: counting method, mapping method, and scoring method. The counting method simply makes bone age assessment based on the number of ossification centers present in the left hand. The mapping method takes the whole radiographs as the comparison object, compares it with the radiographs in the standard library, then selects the most similar radiograph, and uses its bone age as the result of the bone age assessment. The GP (Greulich-Pyle) atlas method (Bayer, 1959) is the most representative method among the mapping methods. The scoring method first sets several bones on the left hand as reference bones, then assesses the developmental stage of each reference bone according to the shape, size, and texture of each bone. Each developmental stage corresponds to a certain maturity score. When using the scoring method for bone age assessment, experts first assess the developmental stage of all reference bones, then integrate the maturity scores of all reference bones to convert the final bone age. TW (Tanner-Whitehouse) series of scoring methods (Morris, 1986) are the most representative scoring methods among the scoring methods for their high accuracy. Long assessment time and low assessment accuracy are two major
drawbacks of traditional bone age assessment methods. In addition, both expertise and practical experience are necessary to perform an accurate bone age assessment. Traditional bone age assessment methods are affected by intra- and inter-observer variability (Bert et al., 2001). This means that the results of the bone age assessment may be different from one expert to another. Even the same expert may have different assessment results for the same radiograph at different times.

Thanks to the development of deep learning techniques, deep learning-based bone age assessment methods have been extensively studied. Bone age assessment can be done in seconds using these methods. The assessment accuracy of these methods is comparable to that of experts. Convolutional neural networks are commonly used in these deep learning-based methods for their powerful feature extraction and abstraction capabilities. Some methods directly extract image features of whole radiographs and use these features to perform bone age assessment. An end-to-end model was usually used in these methods, for the input radiographs, the model can directly output the bone age assessment results. However, there exist many irrelevant areas in the radiographs, those areas can affect the accuracy of bone age assessment. So some other methods were brought up. These methods first segment the radiographs to extract representative ROIs, then use these ROIs for bone age assessment. These methods are highly interpretable, while the selection of regions of interest requires a certain level of expertise. In addition, a large amount of additional annotation is required for ROI extraction, such as the coordinates of key points and the size of the bound box.

In this paper, a novel end-to-end bone age assessment model is proposed, which can extract multiple informative ROIs in radiographs, and use these ROIs to improve the accuracy of bone age assessment. During model training, the only required annotation is the bone age corresponding to the radiographs. Fine-grained classification classifies different subcategories within the same category, resulting in more accurate classification results. The most prominent feature of fine-grained classification is the large intra-class variation and small inter-class variation. Bone age assessment is considered a fine-grained classification problem in this paper for the following reasons: (1) Large intra-class variation. Bone age assessment is performed by combining the shape, size, and texture information of several reference bones. For two radiographs with the same bone age, they may have only a small part of the reference bone that are similar, while the others are quite different. (2) Small inter-class variation. Two radiographs with different bone ages may differ significantly in only one reference bone, as some bones play a significant role in bone age assessment.

The main contributions of this paper are as follows:

1. In this paper, a novel end-to-end bone age assessment model was proposed. For the input radiographs, multiple informative ROIs can be selected, and these ROIs are used to make accurate bone age assessments.
2. Improved the image feature fusion method in the proposed model. After the feature fusion method is improved, the spatial and semantic information in different image feature maps can be better fused, so the model can extract ROI more accurately.
3. Improved classification loss function. With the improved classification loss function, the model also produces good evaluation results in the case of sample imbalance.
4. Used a public dataset to validate the proposed model. The experimental results demonstrate that the proposed can be well applied to the bone age assessment task, and the accuracy of the bone age assessment outperforms other methods, with the assessment results achieving a MAE of 6.55 months.

The remainder of this paper is organized as follows: Section 2 of this paper summarizes the related work in the field of bone age assessment in China and abroad; Section 3 provides a detailed description of the methods used in this paper; Section 4 describes the experimental setup and results of experiments; Section 5 presents the conclusions and future directions.

2. Related work

In the past few years, various bone age assessment methods have been proposed, among which deep learning-based bone age assessment methods have been widely studied. Spampinato et al. (2017) used the Overfeat network (Sermanet et al., 2013) to extract image features from radiographs using a regression network for bone age assessment. In their method, a deformation layer was added to the network to increase the robustness of the network thus making it adaptable to different hand postures. Tong et al. (2018) designed a bone age assessment model combining heterogeneous feature learning. In their method, a convolutional neural network based on VGG (Simonyan & Zisserman, 2014) was used for extracting image features from radiographs. Then a support vector regression machine with multiple kernel learning was used for assessing bone age, utilizing the extracted features in combination with information such as race and gender. Wu et al. (2019) designed a bone age assessment model based on the TW scoring method, using a Mask R-CNN subnet to extract the
ROIs of the reference bones, and then using a residual attention network to perform bone age assessment. Pan et al. (2020) also started from the perspective of image preprocessing to improve bone age assessment. They combined UNet (Ronneberger et al., 2015) with reinforcement learning methods to extract ROIs of reference bones from radiographs with only a small number of annotated images. Lee et al. (2020) manually labeled key points on the radiographs and used these points to select several ROIs from large to small. These ROIs were classified using convolutional neural networks to make bone age assessments. Ultimately, the smallest ROI yielded the best bone age assessment results. In order to train a bone age assessment model on small-scale radiograph datasets, a light convolutional neural network was designed and proposed by Miao et al. (2020). A SuperPoint network was used to extract robust feature maps. These extracted feature maps are then downsampled and fed into an optimized classification network to produce bone age assessment results. Marouf et al. (2020) considered gender information as a fundamental aspect of bone age assessment. The bone age assessment model they designed contains two parts, both using convolutional neural networks to reach the target task. The first part is used for gender classification to determine the gender of the input image. The second part is used to perform the bone age assessment, and the input to this part includes the gender information obtained in the previous part with the image information from the radiographs.

Most of the above studies have performed bone age assessments based on entire radiographs. These methods use features extracted by CNNs or manually designed features for bone age assessment, thus the accuracy of bone age assessment depends on the merit of the extracted features. Other methods first extract different ROIs in the radiographs and then use the extracted ROIs for bone age assessment. A large amount of extra annotation is needed in these methods. In addition to the bone age values corresponding to the radiographs, the localization of ROIs often requires information such as the size of the localization frame and the coordinates of the key points of the ROIs. Moreover, most of the models proposed in these approaches are implemented in multiple stages and cannot be trained end-to-end.

Our proposed model is an end-to-end trainable bone age assessment model. The model is able to extract multiple informative ROIs in radiograph and use these ROIs to improve the accuracy of bone age assessment. During model training, no extra annotation is required, only the image-level annotation of the real bone age value is needed.

3. Method

In this paper, a novel bone age assessment model is proposed. The proposed model was based on the fine-grained classification model as the baseline model, which is proposed by Yang et al. (2018). The proposed bone age assessment model consists of four parts, which are feature extractor, ROI selection subnet, guidance subnet, and assessment subnet. The overall structure of the bone age assessment model is shown in Figure 1.

The input of the model is a radiograph of the left hand. The feature extractor is used to extract image features from the radiograph. Then the ROI selection subnet selects K ROIs with the most representative image features. The guidance subnet evaluates the ROIs selected by the ROI selection subnet, calculates the classification confidence of each ROI. These classification confidences are feedback to the ROI selection subnet, update the ROI selection subnet to better select ROIs. Finally, the assessment subnet combines the image features of the whole radiograph as well as the K selected ROI regions to perform bone age assessment and output the corresponding results. In this paper, K is set as 4.

The entire bone age assessment network can be trained end-to-end. The core idea of the model is to select the most informative ROIs to help improve the accuracy of bone age assessment. During the model training, the ROI selection subnet is continuously optimized based on the feedback from the guidance subnet, so that the most informative ROIs can be selected. Ultimately, these extracted ROIs can effectively improve the accuracy of bone age assessment.

3.1. Feature extractor

In this paper, ResNet50 (He et al., 2016) is used as the base network of the feature extractor. The fully connected layer and Softmax layer of the ResNet50 network are removed and only the feature extraction part of the network is used. Image features in the radiographs are extracted by the feature extractor and feature maps are generated. Then feature maps are input to the ROI selection subnet. The use of residual connection structure is the most important feature of Resnet, which is shown in Figure 2.

For the input feature map $X$, the purpose of a convolution layer is to fit a potentially constant mapping function: $H(X) = X$. In the residual connection structure, the residual connection link is introduced (dashed line in Figure 3). With this structure, the input $X$ can be passed directly to the next layer of the network. The function to be fitted becomes $H'(X) = F(X) + X$. Thus the objective of this convolutional layer changed to learning a residual function $F(X) = H'(X) − X$. In this way, the output of
the mapping function is more sensitive to changes in the input after introducing residual connections. During backpropagation, small loss gradients are more likely to reach the shallow neuron residual connectivity structure solves the network degradation problem in deep networks, allowing deeper network structures to be trained.

3.2. ROI selection subnet

The role of the ROI selection subnet is to select multiple most informative ROIs in the radiographs based on the image features extracted by the feature extractor.

Image features extracted by CNNs often have two aspects of information, respectively semantic information and spatial information. Semantic information plays a key role in image classification, and spatial information plays a key role in the localization of key features. In the shallow part of the network, simple image features such as size, textures, and edges in the image are usually extracted. These shallow features have weak semantic information but strong spatial information. As the depth of the network increases, more complex and abstract features in the image are extracted, and these deep features have strong semantic information but lack spatial information. In order to make full use of the semantic and spatial information in the radiograph, the FPN (Feature Pyramid Networks) (Lin et al., 2017a) + PAN (Path Aggregation Network) (Liu et al., 2018) structure is used in this paper for feature fusion. The structure of FPN + PAN is shown in Figure 3.

\[ M_1, M_2, M_3 \] are the feature maps extracted by the feature extractor, including shallow features and deep
features. These feature maps are first fused using the FPN structure. The FPN is a top-down pyramid structure, and the different feature maps are fused by using the lateral connection, new feature maps \( \hat{M}_1, \hat{M}_2, \) and \( \hat{M}_3 \) are obtained after fusion. Figure 4 shows the structure of the lateral connection, we use the feature maps \( M_1, M_2 \) as examples to illustrate how lateral connection works. We use the bilinear interpolation method to upsample the feature map \( M_2 \), a new feature map \( M_2' \) is obtained, which has the same size as the feature map \( M_2 \). Then a \( 1 \times 1 \) convolution operation is performed on \( M_2' \). After that \( M_2' \) is fused with the feature map \( M_1 \) by element-wise operation. Since deep features tend to have more semantic information, the semantic information in the feature map is all enhanced after fusing by the FPN structure. The structure of PAN is used after the FPN structure, which is a bottom-up pyramid structure. The shallow feature maps with rich spatial information are fused with the deep feature maps to enhance the spatial information. In the PAN structure, averaging pooling is used to downsample the shallow features, after which the new features are fused with the deep features using the element wise operation. Finally, image features \( \tilde{M}_1, \tilde{M}_2, \tilde{M}_3 \) with rich spatial information and strong semantic information are obtained.

Using the feature maps \( \tilde{M}_1, \tilde{M}_2, \tilde{M}_3 \), the ROI selection subnet is able to calculate the amount of information of each ROI in the original map, which is defined as \( I(R) \) and is used to evaluate how informative the ROI is. The purpose of the ROI selection subnet is to select K ROIs in order according to the amount of information of the ROIs, denoted as \( R_1, R_2, \ldots, R_K \). Here, \( K \) is a configurable parameter. The \( K \) sorted ROIs should satisfy the following conditions: \( I(R_1) > I(R_2) > \cdots > I(R_K) \).

### 3.3. Teaching subnet

The role of the guiding subnet is to optimize the ROI selection subnet. For the \( K \) ROIs selected by the ROI selection subnet, the guidance subnet calculates their classification confidence, defined as \( C(R) \). The calculated classification confidence, \( C(R_1), C(R_2), \ldots, C(R_K) \) are feedback to the ROI selection subnet. The classification confidence refers to the probability of classifying an ROI as its corresponding true label. Greater classification confidence means that the corresponding ROI is more helpful for the classification of the whole radiograph, and the more informative the ROI should be. The ROI selection subnet is continuously optimized in the model training process based on the feedback from the guidance subnet so that the final selected \( K \) ROIs satisfy formula (1):

\[
\forall i, j \in [1, K], \text{ if } I(R_i) \geq I(R_j), C(R_i) \geq C(R_j) \tag{1}
\]

That is, for the ROI selected in the ROI selection subnet, if the amount of information is greater, the confidence level should be also higher.

### 3.4. Assessment subnet

The role of the assessment subnet is to perform bone age assessment. For the inputted radiograph, the feature extractor extracts the feature from the entire picture, denote as \( F_A \). The ROI selection subnet selects \( K \) most informative ROIs, \( R_1, R_2, \ldots, R_K \), resize them to the same size as the whole film. Then feature extractor is used to extract image features of these \( K \) ROIs, denote as: \( F_1, F_2, \ldots, F_K \). These \( K+1 \) image features including that from the whole film and that form \( K \) ROIs are connected to form a new feature \( F_{ALL} \). After the calculation of \( F_{ALL} \) by two fully connected layers, the results of bone age assessment are obtained. To increase the nonlinearity of the model, Relu is used as the activation function in the middle of the last two fully-connected layers.

### 3.5. Loss function

The overall loss function of the proposed model consists of three components, which are ROI selection loss,
guidance loss, and assessment loss, denoted as $L_R, L_G, L_A$, respectively.

The ROI selection loss uses a pairwise ranking loss to align the informativeness ranking with the confidence ranking of $K$ ROIs. The ROI selection loss $L_R$ is defined as follows:

$$L_R = \sum_{i,s \in [1,K], i < s} f(l_s - l_i)$$  \hfill (2)

Here, the hinge loss function $f(x) = \max(1 - x, 0)$ is used as the loss function $f$. When there is any inconsistency between the informativeness ranking and the confidence ranking, the inconsistent ranking pairs are penalized, leading to an increase in loss. And conversely, when the images are correctly classified, the loss will be small, and consequently, the loss function will be larger. In this paper, Focal Loss (Lin et al., 2017b) is used to calculate the classification loss, which is defined as follows:

$$L = -\sum_{i=1}^{M} (1 - C(R_i))^{\gamma} \log C(R_i) \; \text{for the selected} \; K \; \text{ROIs},$$  \hfill (3)

Here, the first term represents the sum of classification loss for the selected $K$ ROIs, the second term represents the classification loss for the whole film. $R_k$ represents the whole radiograph. In this paper, Focal Loss (Lin et al., 2017b) is used to calculate the classification loss, which is defined as follows:

$$FL = -(1 - p_t)^{\gamma} \log(p_t)$$  \hfill (4)

Here, $p_t$ is the classification confidence of classifying the images as true labels and $\gamma$ is a settable parameter. Focal Loss forces the network to focus more on the difficult and misclassified samples during training. When the images are correctly classified, the loss will be small, and conversely, when the images are incorrectly classified, the loss function will be larger. In this paper, $\gamma$ is set as 2 throughout.

The definition of assessment loss $L_A$ is shown in formula (5):

$$L_A = -(1 - C(R_{ALL}))^{\gamma} \log(R_{ALL})$$  \hfill (5)

The assessment loss represents the classification loss obtained when the bone age assessment is performed by the assessment subnet. Where $R_{ALL}$ represents the image features of the whole radiograph combined with the $K$ selected ROIs.

The loss of the whole proposed model $L_{All}$ is defined as follows:

$$L_{All} = L_R + L_G + L_A$$  \hfill (6)

### 4. Experimental and results

#### 4.1. Data overview

The dataset used in this paper is a publicly available dataset for bone age assessment, which is published by the Radiological Society of North America (RSNA) in 2017 (Halabi et al., 2018), namely the RSNA dataset.

The dataset contains 14,236 radiographs of the left hand from two children’s hospitals with corresponding reports. The whole dataset is divided into three parts: training set, validation set, and test set, in which there are 12,611 samples in the training set, 1,423 samples in the validation set, and 200 samples in the test set. For each image, the average of the bone age values from the corresponding report and the bone age assessment values from 3 additional experts were used as its reference bone age value, which was used as the training label in the model training. All images corresponded to a range of bone ages from 0 to 228 months, and their distribution is shown in Table 1.

#### 4.2. Evaluation metric

For bone age assessment tasks, the commonly used evaluation metric is the MAE, which is defined as follows:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |BA_{p_i} - BA_{gt_i}|$$  \hfill (7)

where $m$ is the total number of samples for which bone age assessment was performed, $BA_{p_i}$ is the bone age assessment result of the $i$-th sample, and $BA_{gt_i}$ is the label value of the $i$-th sample.
In the assessment of bone age, we usually use one year as a threshold to assess whether an adolescent is developing early or late. Using MAE as an assessment metric does not reflect the superiority of our model. Therefore, in this paper, we use additional metrics, which are the accuracy of absolute error within 1 year and the accuracy of absolute error within 0.5 years, respectively.

In this paper, both MAE and accuracy within 1 and 0.5 years are used as evaluation metrics.

4.3. Experimental environment configuration and experimental parameter

The configuration of the experimental environment is shown in Table 2.

In the model training phase, the batch size is set to 4 and the epoch is set to 120, and an epoch is completed when all images in the training set have been trained.

The Stochastic Gradient Descent (SGD) optimizer was used in model training also, with the learning rate set to 0.0001, the weight decay set to 0.01, and the momentum set to 0.9.

In addition, MultiStep LR was used in the experiments. It is a learning rate decay strategy that dynamically adjusts the learning rate during the model training. The learning rate of the model changes to 10% of the current learning rate after 60 and 100 epochs of training are completed.

4.4. Base network selection

In the proposed bone age assessment model, all three subnets share one feature extractor. Therefore, in order to obtain better accuracy of bone age assessment, it is especially important to choose a convolutional neural network with powerful feature extraction ability as the base network.

We conducted experiments to compare several popular convolutional neural networks, which include ResNet50, Inceptionv3 (Szegedy et al., 2016b), Inception-ResNetv2 (Szegedy et al., 2016a), and Densenet (Huang et al., 2017). In all experiments in this paper, the data set was divided into months, with different months of bone age considered as different classifications. The evaluation results obtained from training on the dataset using the four above-mentioned CNNs are shown in Table 3.

4.5. Feature fusion methods

The ROI selection subnet uses the feature maps extracted by the feature extractor to select the most representative ROIs, therefore improving the bone age assessment accuracy of the model. The key point here is the quality of the feature maps. It is essential to adopt the feature fusion method to fuse the shallow feature maps with the deep feature maps. After feature fusion, both semantic and spatial information in the feature map is enhanced, which will be beneficial with the ROI selection subnet.

In the baseline model, the structure of FPN is used for feature fusion. However, the FPN is a top-down pyramid structure that can fuse the semantic information from the deep feature map to the shallow feature map well, but not the spatial information from the shallow feature map to the deep feature map. PAN structure can solve this problem by fusing the spatial information from the shallow feature map to the deep feature map. Therefore, in this paper, a PAN structure is added behind the FPN structure to form an FPN + PAN structure for feature fusion. A set of experiments was designed to verify the usefulness of different feature fusion methods, and the results are shown in Table 4. The experimental results showed that when the FPN + PAN structure was used for feature fusion, the bone age assessment accuracy of the whole model was higher and the MAE decreased by 0.18 months.

4.6. Determination of parameters in focal loss

In this paper, Focal Loss was used as the classification loss function. $\gamma$ is a settable parameter in the Focal Loss function. When $\gamma$ is set to greater than 0, it makes the model focus more on hard-to-classify samples during training by reducing the loss of easy-to-classify samples. We set $\gamma$ to 0.25, 0.5, 2, and 5 to conduct the experiments separately,
Table 5. Performance of models with different values of $\gamma$.

| Value of $\gamma$ | MAE/months |
|-------------------|------------|
| 0.25              | 7.31       |
| 0.5               | 7.10       |
| 2                 | 6.65       |
| 5                 | 6.87       |

Table 6. Performance comparison of different bone age assessment methods.

| Methods                    | MAE of male/months | MAE of female/months | MAE of all/months | Accuracy within 1 year/\% | Accuracy within 0.5 years/\% |
|----------------------------|--------------------|-----------------------|-------------------|---------------------------|-----------------------------|
| Lee et al. (2020)          | -                  | -                     | 8.89              | -                         | -                           |
| Wu et al. (2019)           | -                  | -                     | 7.38              | -                         | -                           |
| Pan et al. (2020)          | -                  | -                     | 7.35              | -                         | -                           |
| Baseline model             | 7.36               | 6.84                  | 7.08              | 86.34                     | 62.03                       |
| Baseline model + FPN + PAN | 7.12               | 6.71                  | 6.90              | 87.42                     | 62.44                       |
| Baseline model + FL        | 6.91               | 6.81                  | 6.86              | 85.95                     | 61.05                       |
| Baseline model + FL + FPN + PAN (our method) | 6.65 | 6.64 | 6.65 | 88.53 | 62.87 |

and the experimental results are shown in Table 5. When $\gamma$ is set to 2, our model can obtain the best experimental results.

4.7. Performance comparison

Table 6 shows the experimental performance of several different methods of bone age assessment, including the model proposed in this paper as well as those proposed in other papers.

The first three rows of the table indicate the experimental performance of the methods proposed in recent years, which have been experimented on the RSNA dataset. The last four rows of the table indicate the experimental performance of the experiments designed in this paper. Since the radiographs of different genders with the same bone age have large differences, the data set was separated by gender and the experiments were conducted separately. These three sets of experiments are used to verify the improvement of the baseline model by different feature fusion structures and different classification loss functions. The experimental results showed that the experimental performance improved with the use of Focal Loss as the loss function, and the MAE decreased by 0.22 months. After using the FPN + PAN structure as the feature fusion structure, the MAE decreased by 0.18 months. These results demonstrate that changes in the feature fusion method and improvements in the classification loss function both contribute to the accuracy improvement of the baseline model.

The proposed bone age assessment model finally reached the highest accuracy of bone age assessment with an MAE of 6.65 months, which outperforms other bone age assessment methods.

In addition, we counted the accuracy of each method’s assessment error within 1 year of age and within 0.5 years of age, show them in the last two columns of Table 6. The results show that our proposed model obtains the highest accuracy rate. The accuracy within 1 year of age is 88.53\%, and the accuracy within 0.5 years of age is 62.87\%.

5. Conclusion and future work

In this paper, bone age assessment is considered a fine-grained classification problem. Based on the fine-grained classification model, a novel bone age assessment model that can be trained end-to-end is designed and implemented. This model is able to select multiple informative ROIs from radiographs and use these ROIs to improve the accuracy of bone age assessment. The experiment results demonstrate that the model can successfully extract the most informative ROIs and perform bone age assessment with an MAE of 6.65 months, which is better than other current studies. In future studies, two improvements are expected: first, the morphology of bones in the left hand at different age stages has large differences. Samples of radiographs at different stages are considered to be trained separately to obtain better classification accuracy. Second, when performing the analysis of experimental results we found that some very poor bone age assessment results were caused by abnormal hand posture. Therefore data enhancement methods are considered to improve the robustness of the model so that the model can adapt to radiographs with different hand postures.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the Basic Public Welfare Research Project of Zhejiang Province [Grant Number LGG22F020014]; National Natural Science Foundation of China [Grant Number 62072410].

Data availability statement

The data that support the findings of this study are available in kaggle at https://www.kaggle.com/kmader/rsna-bone-age
References

Berst, M. J., Dolan, L., Bogdanowicz, M. M., Stevens, M. A., Chow, S., & Brandser, E. A. (2001). Effect of knowledge of chronologic age on the variability of pediatric bone age estimated by the greulich and pyle standards. *American Journal of Roentgenology*, 176(2), 507–510. https://doi.org/10.2214/ajr.176.2.1760507

Greulich, W. W., & Pyle, S. I. (1959). Radiographic atlas of skeletal development of the hand and wrist. *The American Journal Of The Medical Sciences*, 238(3), 393. https://doi.org/10.1097/00000441-195909000-00030

Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., Pan, I., Pereira, L. A., Sousa, R. T., Abdala, N., Kitamura, F. C. (2018). The RSNA pediatric bone age machine learning challenge. *Radiology*, 290(2), 498–503. https://doi.org/10.1148/radiol.2018180736

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June 26–July 01). Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90

Huang, G., Liu, Z., Laurens, V., & Weinberger, K. Q. (2017, June 21–26). Densely connected convolutional networks. 2017 IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2017.243

Lee, J. H., Kim, Y. J., & Kim, K. G. (2020). Deep learning or automated skeletal bone age estimation using deep learning and hand X-ray images. *Biomedical Engineering Letters*, 10(3), 323–331. https://doi.org/10.1007/s13534-020-00151-y

Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a, July 21–26). Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.106

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017b). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 99, 2999–3007. https://doi.org/10.1109/tpami.2017.324

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018, June 18–23). Path aggregation network for instance segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2018.00913

Marouf, M., Siddiqi, R., Bashir, F., & Vohra, B. (2020, January 29–30). Automated hand X-Ray based gender classification and bone age assessment using convolutional neural network. 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). http://dx.doi.org/10.1109/iomet48670.2020.9073878

Mi, J., Yue, H., Wu, X., Xu, D., & Chen, W. (2020, November 06–08). Bone age assessment based on SuperPoint features.

Morris, L. L. (1986). Assessment of skeletal maturity and prediction of adult height (tw2 method). *American Journal of Human Biology*, 14(6), 788–789. https://doi.org/10.2307/2802657

Pan, X., Zhao, Y., Chen, H., Wei, D., Zhao, C., & Wei, Z. (2020). Fully automated bone age assessment on large-scale hand x-ray dataset. *International Journal of Biomedical Imaging*, 2020, 1–2. https://doi.org/10.1155/2020/8460493

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2013). OverFeat: Integrated recognition, localization and detection using convolutional networks. *Eprint Arxiv*. arXiv: 1312.6229, 2013.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., & Leonardi, R. (2017). Deep learning or automated skeletal bone age assessment in x-rays images. *Medical Image Analysis*, 36, 41–51. https://doi.org/10.1016/j.media.2016.10.010

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016a, February 04–09). Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.308

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016b, June 25–July 02). Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2818–2826). https://doi.org/10.1109/cvpr.2016.308

Tong, C., Liang, B., Li, J., & Zheng, Z. (2018). A deep automated skeletal bone age assessment model with heterogeneous features learning. *Journal of Medical Systems*, 42(12), 12. https://doi.org/10.1007/s10916-018-1091-6

Wu, E., Kong, B., Wang, X., Bai, J., Lu, Y., Gao, F., & others (2019, April 08–11). Residual attention based network for hand bone age assessment. 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019) (pp. 1158–1161). https://doi.org/10.1109/isbi.2019.8759332

Yang, Z., Luo, T., Dong, W., Hu, Z., & Wang, L. (2018). Learning to navigate for fine-grained classification. *Lecture Notes in Computer Science* 438–454. https://doi.org/10.1007/978-3-030-01264-9_26

2020 Chinese Automation Congress (CAC) (pp. 3921–3926). IEEE. http://dx.doi.org/10.1109/cac51589.2020.9327736