

Article

Data Imbalance Immunity Bone Age Assessment System Using Independent Autoencoders

Ching-Tung Peng ^{1,*} , Yung-Kuan Chan ² and Shyr-Shen Yu ¹¹ Department of Computer Science and Engineering, National Chung Hsing University, Taichung City 402, Taiwan² Department of Management Information Systems, National Chung Hsing University, Taichung City 402, Taiwan

* Correspondence: roger@teamail.ltu.edu.tw

Abstract: Bone age assessment (BAA) is an important indicator of child maturity. Generally, a person is evaluated for bone age mostly during puberty stage; compared to toddlers and post-puberty stages, the data of bone age at puberty stage are much easier to obtain. As a result, the amount of bone age data collected at the toddler and post-puberty stages are often much fewer than the amount of bone age data collected at the puberty stage. This so-called data imbalance problem affects the prediction accuracy. To deal with this problem, in this paper, a data imbalance immunity bone age assessment (DIIBAA) system is proposed. It consists of two branches, the first branch consists of a CNN-based autoencoder and a CNN-based scoring network. This branch builds three autoencoders for the bone age data of toddlers, puberty, and post-puberty stages, respectively. Since the three types of autoencoders do not interfere with each other, there is no data imbalance problem in the first branch. After that, the outputs of the three autoencoders are input into the scoring network, and the autoencoder which produces the image with the highest score is regarded as the final prediction result. In the experiments, imbalanced training data with a positive and negative sample ratio of 1:2 are used, which has been alleviated compared to the original highly imbalanced data. In addition, since the scoring network converts the classification problem into an image quality scoring problem, it does not use the classification features of the image. Therefore, in the second branch, we also add the classification features to the DIIBAA system. At this time, DIIBAA considers both image quality features and classification features. Finally, the DenseNet169-based autoencoders are employed in the experiments, and the obtained evaluation accuracies are improved compared to the baseline network.

Keywords: bone age assessment; data imbalance; CNNs; autoencoder; scoring network

Citation: Peng, C.-T.; Chan, Y.-K.; Yu, S.-S. Data Imbalance Immunity Bone Age Assessment System Using Independent Autoencoders. *Appl. Sci.* **2022**, *12*, 7974. <https://doi.org/10.3390/app12167974>

Academic Editors: Charles Tijus, Teen-Hang Meen and Chun-Yen Chang

Received: 30 June 2022

Accepted: 6 August 2022

Published: 9 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

For children with different growth conditions, the differences in the growth process caused by the environment, genes, diet, etc., can usually be reflected by bone maturity, and then used to evaluate the development of children in cases such as endocrine disorders and pediatric syndromes. Therefore, assessing bone age is an important task in clinical application. In general, estimating the bone age of the non-dominant hand is the most commonly used method. Using the non-dominant hand to estimate bone age has several advantages. First, there are a lot of bones in palms. Second, the non-dominant hand is used less frequently and is less affected by external forces. Because most people's non-dominant hand is the left hand, most estimates of bone age are based on the left hand. The assessment of bone age is mostly before and after puberty. For adults, due to the fusion of the growth plate, it is difficult to use the condition of the palm bones for assessment. Traditionally, there are two most commonly used methods for bone age assessment: the Greulich & Pyle Atlas (GP) method [1] and the Tanner Whitehouse (TW) method [2,3]. In the GP method, the X-ray to be assessed for bone age is compared with the GP atlas to give bone age. This

approach has several disadvantages. First of all, it takes a long time to assess bone age, and it also requires professional medical personnel to evaluate. Secondly, due to manual evaluation, there are human errors. In the TW method, 20 bones are scored separately, these scores are then combined to obtain a composite score, which is then mapped to a table to calculate bone age. Since bone age is determined by scoring, a possible problem is that an overly simplistic scoring method may lead to inaccurate estimates of bone age.

Since the above disadvantages lie in traditional bone age assessment methods, in the past, automatic bone age assessment systems have been proposed in many studies, while due to the development of deep learning technology, more and more systems for medical purposes are based on deep learning, including medical human–computer interfaces, medical diagnosis systems, medical data analysis, medical registration systems, etc. These various applications are increasingly inseparable from deep learning. For automatic bone age assessment systems, many studies also build their systems based on CNN. In these studies, the way of estimating bone age can be divided into two types. First, directly use CNNs to build a regression model for bone age [4,5], since this method is to fit all bone ages of palm bone X-ray by one CNN, in this case, the main problem is that if the estimation is wrong, it is easy to have great errors. Another way is to divide the palm bone X-ray into several stages according to the bone age [6,7]. At this time, the CNNs are dealing with the classification problem, however, the main problem of this method is that if the stages are divided too coarsely, the estimation of bone age will be not accurate enough. If the stages are divided too finely, the difference between the stages will be not obvious, which will cause it to be difficult to fit by CNNs and prone to overfitting. In order to avoid these problems, a compromise solution can be considered. First, the palm bone X-rays are divided into several rough stages, and for each stage an independent regression model is trained to estimate the bone age. Therefore, the accuracy of classification of the rough stages is very important. In this paper, we mainly deal with the problem of data imbalance in the classification at the coarse stage, and use the classification results to the CNN regression network to calculate bone age.

Although many existing bone age studies have trained a large number of palm bone X-ray databases [8], which has led to good results on classification, however, due to differences in the growth process of children, due to differences in races, diets, and different growth environments, the development process of children varies in different regions. Therefore, it is necessary for each region to build its own bone age estimation system. For this reason, this paper is based on the palm bone X-ray films taken by children in Taiwan to build a system. In this case, we consider a problem: it will take a long time for X-ray data to accumulate to a sufficient number. Especially in toddler and post-puberty stages, there are few outpatient cases, so the data at these two stages will be greatly lacking. In the initial research stage of the bone age system, there will be a problem of very unbalanced data. For example, in this paper, in the three stages we use, the proportion of the database we collected in females is approximately 5:90:5, which is a very biased data distribution. In our follow-up experiments, we will find that when training directly with CNNs all the classification results are in the puberty stage.

This kind of X-ray data imbalance problem often occurs. For example, disease cases are usually much smaller than those without disease, or, as in the case of this paper, outpatient cases during toddler and post-puberty stages are very scarce, but these are a small number of cases, usually the system still must be able to distinguish them, so many existing studies are devoted to alleviating the problem of serious imbalances of X-ray data. The most commonly used methods are several: re-sampling the data by undersampling [9,10] or oversampling [11–14], using transfer learning [15–17] to pre-train network parameters with other datasets with relatively sufficient data, and using the target task dataset to fine-tune the model. In addition, the loss function of its CNNs is improved [18–21]. The loss function is usually designed to compensate for a small number of classes during training.

The undersampling and oversampling methods each have their shortcomings. In oversampling, if the amount of data of certain classes is quite lacking, no matter how the

data is augmented, the data generated by the augmentation is always generated from the original small amount of classes. Therefore, the features that can be learned in CNNs are still very limited. In undersampling, if we have to discard those samples from a large number of classes, it is a waste of these discarded data, especially in medical imaging where image collection is more difficult. By using the loss function to compensate for the small number of classes, the network is made to learn more towards the small number of classes to alleviate the data imbalance problem, but the network always learns all classes of datasets on the same network, which is still susceptible to a large number of classes. Therefore, in this paper, we use three (three-stages target in this paper) autoencoders [22] and a CNN-based scoring network to alleviate the data imbalance problem, because in each stage each autoencoder is independent, they do not influence each other, so the small number stage will not be disturbed by the large number stage. In DIIBAA, the classification problem is turned into an image quality scoring problem. Its classification is mainly based on the features of image quality. However, the classification's features are still precious information. Therefore, we also add a classifier for the bone age stage and integrate these two different results for the final judgment result. After the classification is completed, we use a CNN-based regression network to determine the bone age value.

In the subsequent section of this article, we first describe the dataset used in this article and detail the DIIBAA method proposed in this article in Section 2, which includes two branches which use image quality features and classification features, respectively. And finally, we use the two to jointly determine the results of bone age stage prediction. After the stage prediction, we build the bone age regression models to evaluate the bone age. In Section 3, we will show the experimental results and their comparisons. In Section 4, we discuss the advantage and limitations of our method. In Section 5, we give the conclusion of the DIIBAA method proposed in this article.

2. Materials and Methods

The datasets used in this paper are shown in Table 1. These stages are based on [23] (in fact, they have a total of 6 stages, we consider that the two stages are too similar, so the adjacent stages are merged). S_1 is the toddler stage, S_2 is the puberty stage, and S_3 is the post-puberty stage. The data set is from the Taoyuan Cheng Hsin General Hospital, and the labels are judged by two professional physicians. These two professional physicians refer to the clinical database in the hospital to give the bone age stage and the bone age value. In order to reduce human error, our bone age value is the average judgment of the two physicians. These data were collected by the hospital in 2018 to 2020. It can be found that in these three stages, regardless of gender, they are very unbalanced. As the follow-up experiments show, in this extremely unbalanced situation, when directly using CNN training, they are very unbalanced and both ends (S_1 and S_2) of stages have weak discernment.

Table 1. Bone age dataset used in this paper.

Stage	Number of Female	Number of Male
S_1	66	63
S_2	1140	875
S_3	73	133

In order to clearly illustrate the input data of this paper, we sample three pictures from each stage of the data in Table 1, as shown in Figure 1, from left to right are toddlers (S_1), puberty (S_2) and post-puberty (S_3) stages.



Figure 1. Three-stage metacarpal X-ray images.

The DIIBAA consists of two branches, as shown in Figure 2. Branch 1 consists of three autoencoders and a scoring network. This branch converts the classification problem into an image quality assessment problem, and because the high-quality images and low-quality images have been converted into 1:2, the problem of data imbalance in this branch can be greatly alleviated. In the second branch, the classification information is actually very valuable since the classification features are not used in the first branch, but the network trained directly using all the data is extremely imbalanced and relationships are almost indistinguishable for each stage. Therefore, in order to utilize classified information, we use mixup data augmentation [24] on branch 2 to alleviate the problem of data imbalance and make the network slightly more distinguishable. For the samples that are ambiguous in the first branch, we use branch 2 to judge them, and finally, give a CNN-based regression network to the three bone age stages, respectively, and use it to give the bone age values. In short, we first use branch 1 to predict, then check its results. If the check is passed, the bone age prediction stage is the result of branch 1, otherwise, branch 2 is used as the result of the bone age prediction stage.

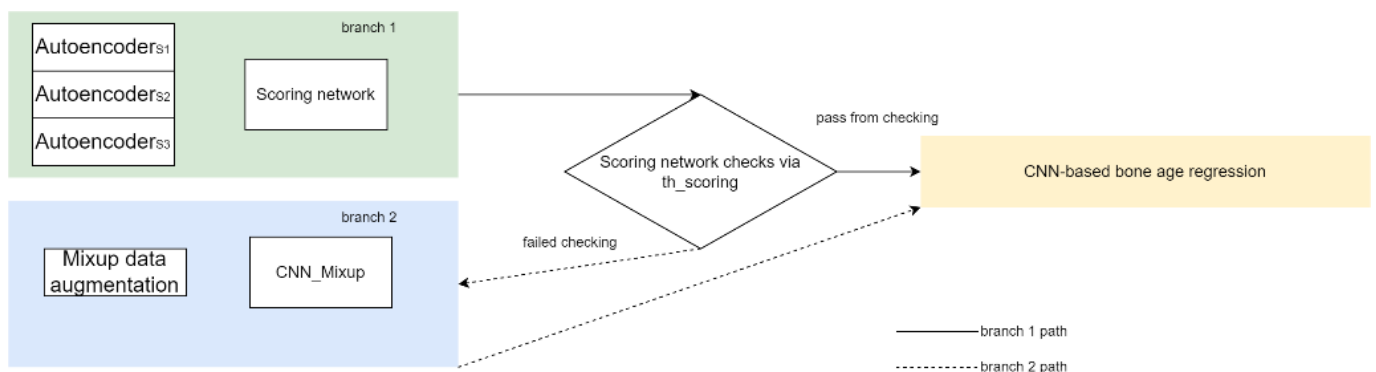


Figure 2. Framework of DAIIBAA.

2.1. Autoencoder-Based Scoring Classification System

In branch 1, we use three autoencoders to represent three stages (S_{1-3} in Table 1). In each autoencoder, the input and output image sizes are 224×224 . During the encoding process, we have a total of three convolutional layers, their kernel size are all 3×3 , and the number of channels from the first layer to the third layer are 512, 256, and 128, respectively. After each convolutional layer, a pooling layer is followed, and in each pooling is the reduced feature map which is $1/2$ times that of the previous layer. In the decoder part, the convolution kernels from shallow to deep layers are 128, 256, and 512, respectively. The rest of the convolution kernel parameters are the same as those of the encoder. After convolution, they are all upsampled to double the size of the previous layer, and finally, a single-channel 1×1 convolution layer will be connected. The output of this layer

represents the reconstructed image. The sigmoid function is used in this layer for nonlinear transformation, and the rest are all ReLU. The architecture of a CNN-based autoencoder is as shown Figure 3. In the scoring network, we use DesnseNet169 [25], we use it for the regression of score, so the output layer activation function is sigmoid. In the following, we will divide the training and testing phases to illustrate the branch 1 part in this paper.

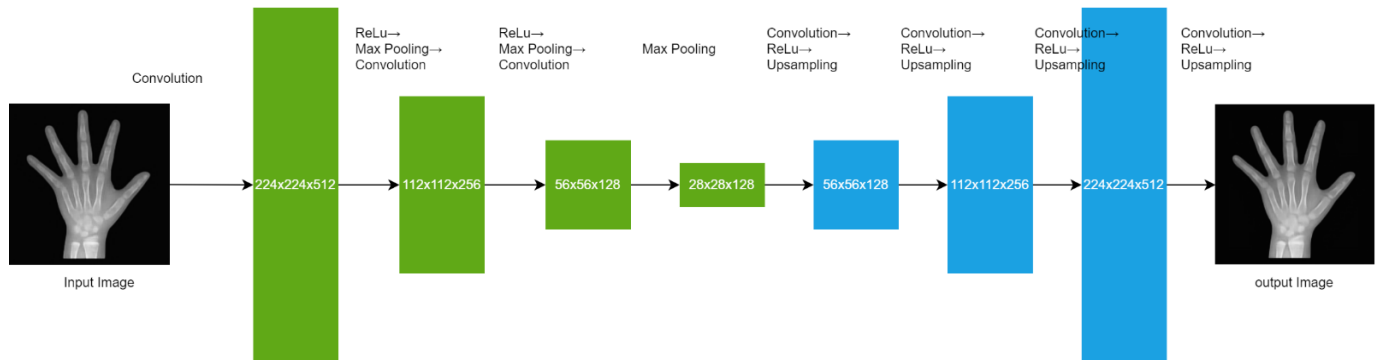


Figure 3. The architecture of CNN-based autoencoder.

In the phase of training, we first train three autoencoders. As shown in Figure 4, we use the corresponding autoencoders for training. For example, in the blue branch, we only use the data in the S_1 stage for training. In the orange branch, only the training of the data in the S_2 stage is used. During the training process, we add the perturbed data of the other stages to weaken the reconstruction ability of the images that are not of our own stage. The input of the perturbed data pair is the original image, and the output (that is labeled) is the result of perturbing the original image. In this paper, the perturbation in the branch is Gaussian noise [26]. Density in Gaussian noise simulation are added to the image in a random way. Its SNR value is about 6.989 to 10, and the SNR value is calculated as: $SNR = 10 \times \log_{10}(\text{signal}/\text{noise})$. The number of perturbed degraded images in this paper is 5% of the images of this stage.

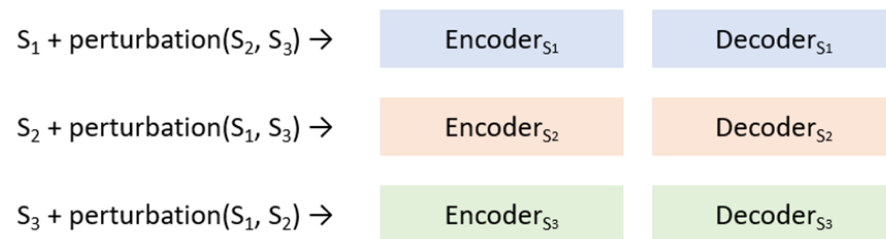


Figure 4. Three autoencoders in branch 1.

After the training of the three autoencoders is completed, we will then use the reconstruction results of these three autoencoders to generate data for the scoring network as training data. The training data we generate will eventually generate two classes, T and F. T is the result after the training image is reconstructed by the autoencoder of its own corresponding stage. For example, the image of the S_1 stage is reconstructed by autoencoders1, and F is the result of the training image reconstructed by the autoencoder of its own non-corresponding stage. For example, the image of the S_1 stage is reconstructed by autoencoders₂ or autoencoders₃, such as (1) and (2):

$$T = \{ \text{Decoder}_{S_1}(\text{Encoder}_{S_1}(S_1)), \text{Decoder}_{S_2}(\text{Encoder}_{S_2}(S_2)), \text{Decoder}_{S_3}(\text{Encoder}_{S_3}(S_3)) \}, \quad (1)$$

$$F = \{ \text{Decoder}_{S1}(\text{Encoder}_{S1}(S_2, S_3)), \\ \text{Decoder}_{S2}(\text{Encoder}_{S2}(S_1, S_3)), \\ \text{Decoder}_{S3}(\text{Encoder}_{S3}(S_1, S_2)) \}, \quad (2)$$

Among them, Decoder_{S1} and Encoder_{S1} are the encoding and decoding functions (CNNs) of stage S_1 , respectively, and S_1 , S_2 and S_3 are the training X-ray data. After taking the data of the two sets of T and F, we set T as 1.0 points, F as 0.0 points, using such a configuration to train the scoring network, here we can see that the ratio of T to F is 1:2 (each image will generate 1 T and 2 F), which is a considerable alleviate for the original extremely imbalanced data (Table 1).

In the test, when an image X is an unknown stage, we first input X to three different autoencoders, they will generate X^{S1_r} , X^{S2_r} and X^{S3_r} , three images, as shown in Figure 5.

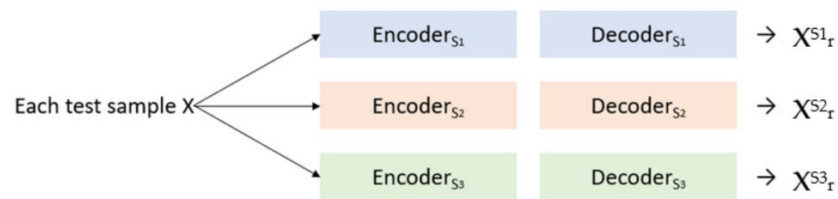


Figure 5. Reconstruct the test image X with three autoencoders separately.

Then, input X^{S1_r} , X^{S2_r} and X^{S3_r} into the scoring network, respectively, and each will receive a score of either X^{S1_r} , X^{S2_r} or X^{S3_r} , as shown in Figure 6:

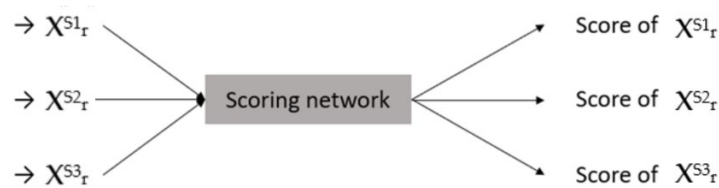


Figure 6. Scoring network.

Finally, X is predicted to be the stage corresponding to the autoencoder network corresponding to the highest score of X^{S1_r} , X^{S2_r} and X^{S3_r} , as shown in Equation (3):

$$\text{Predict result} = \max \{ \text{Score of } X_r^{s1}, X_r^{s2}, X_r^{s3} \}, \quad (3)$$

In branch 1, our principle is that if the image is input into the non-corresponding autoencoders, these autoencoders should not be able to reconstruct it well, conversely, if the X-ray image is input to the corresponding autoencoder there should be excellent reconstructed image quality. Therefore, in the scoring network, we can say that it predicts the image quality to give the input image score. Therefore, if the image has a better reconstruction effect in a certain autoencoder, the image should be the corresponding class for this autoencoder.

2.2. Use Mixup Augmentation to Train CNN Network

Since the above-mentioned way of independent autoencoders does not mainly take into account the fact that the images are actually distinguishable features, in order to keep these features from being wasted, we use a CNN that also builds classification purposes to make DIIBAA's system also take into account classification features. Since CNN is directly used to classify these extremely unbalanced data, they have almost no classification ability (experimental description in the next chapter). Therefore, we use the mixup augmentation [24] method to augment the data of S_1 and S_3 . The mixup augmentation is as follows (4): The formula:

$$\text{image}_{\text{mixup}} = \lambda \times \text{image}_x + (1 - \lambda) \times \text{image}_y, \quad (4)$$

In the above equation, $\text{image}_{\text{mixup}}$ represents the synthesized image, and λ is a value of 0–1. Here we generate λ in a random way, image_x and image_y are two randomly selected images, respectively, and here we use Formula (4) for the X-ray films of the S_1 and S_3 stages to generate synthetic images until the number of images in S_1 and S_3 is equal to the number of images in S_2 . Figure 7 is a mixup operation, using (4) to superimpose two images at the same bone age stage with the random λ .

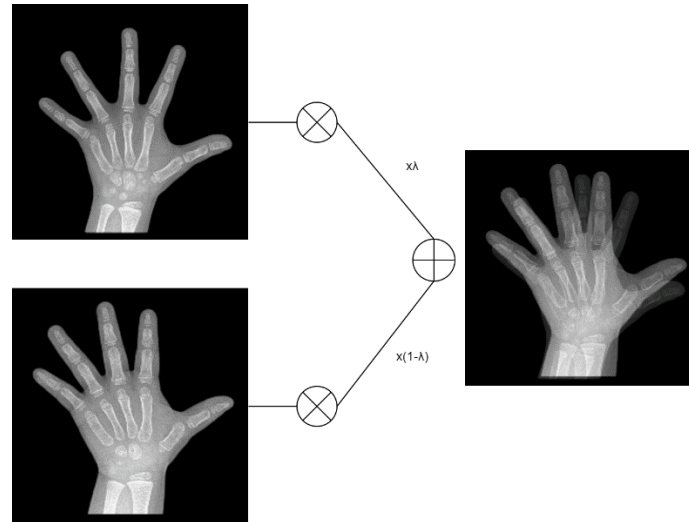


Figure 7. The operation of mixup.

With this data we train a CNN network, called, $\text{CNN}_{\text{mixup}}$, which is an improvement compared to the indistinguishable (without mixup) augmentation, as shown in Tables 2 and 3 (female and male, respectively):

Table 2. Compare the CNN trained with or without mix-up augmentation in the female.

Stage	w/o Mixup Augmentation	with Mixup Augmentation
S_1	0.00%	41.67%
S_2	100.00%	95.67%
S_3	0.00%	30.77%

Table 3. Compare the CNN trained with or without mix-up augmentation in the male.

Stage	w/o Mixup Augmentation	with Mixup Augmentation
S_1	0.00%	41.67%
S_2	93.14%	94.86%
S_3	33.33%	66.67%

When using mixup augmentation, compared to 0% at both ends before no augmentation, with mixup augmentation the two end stages in female are increased from 0%, 0% to 41.67%, 30.77%, and male from 0.00% and 33.33% to 41.67%, 66.67%, respectively. Such a result shows that $\text{CNN}_{\text{mixup}}$ has some discrimination ability.

2.3. Integration 2model

Next, we must integrate the results of the two branches. First, we use the results of the scoring network. If the difference between their top two scores $> \text{th_score}$, then since the network with the highest score is significantly better reconstructed than the other two, this indicates that it has a fairly high chance for this stage, so in this case, we take the final classification result as the result of the scoring network, otherwise the predicted result using $\text{CNN}_{\text{mixup}}$ is the final output result. The details are shown in Algorithm 1:

Algorithm 1. Integration 2 model

Input: Encoder_ S_1 , Encoder_ S_2 , Encoder_ S_3 , Scoring_CNN, CNN_mixup, image_x, th_score
 $Sc_1(\text{image_x}) = \text{Scoring_CNN}(\text{Encoder_}S_1(\text{image_x}))$
 $Sc_2(\text{image_x}) = \text{Scoring_CNN}(\text{Encoder_}S_2(\text{image_x}))$
 $Sc_3(\text{image_x}) = \text{Scoring_CNN}(\text{Encoder_}S_3(\text{image_x}))$
 $hs = \text{highest score}(Sc_1(\text{image_x}), Sc_2(\text{image_x}), Sc_3(\text{image_x}))$
 $shs = \text{second highest score}(Sc_1(\text{image_x}), Sc_2(\text{image_x}), Sc_3(\text{image_x}))$
If $hs - shs > th_score$
 Result = hs
Else
 Result = result of CNN_mixup
Return Result

Next, we find a suitable th_score, we use Algorithm 2 to find this threshold. On the validation set, at an interval of 0.1, find the th_score with the highest accuracy for the overall validation.

Algorithm 2. Finding th_score

Input: Algorithm 1, validation data
Highest accuracy = 0.0
For th from 0.05 to 0.5
 Accuracy = Algorithm 1(validation data, th)
If Accuracy > Highest accuracy
 th_score = th
Return th_score

In this step, we use a CNN regression model to assessment bone age values. In the dataset of Table 1, bone age values are in years and include decimal points (example: 6.5 years means 6 years and 6 months). We built three independent regression models for each of the three bone maturation stages ($S_1 \sim S_3$), and before training the CNN regression model, we normalized the value of bone age to range from 0 to 1. We use the Equation (5) for normalization.

$$BA'_{S_n} = \frac{BA_{S_n} - \text{minimum } BA_{S_n}}{\text{maximum } BA_{S_n} - \text{minimum } BA_{S_n}}, \quad (5)$$

In (4), BA'_{S_n} represents the normalized bone age value, and BA_{S_n} represents the original bone age value. After regressing the results given by the network, we use Equation (6) to calculate the true bone age value.

$$BA = \text{base} + BA' * \text{scale}, \quad (6)$$

In (5), BA represents the final predicted bone age, the base is the minimum bone age value in the connection, as shown in the Table 4, the scale is the bone age range at this stage, maximum BA–minimum BA.

Table 4. CNN-based bone age regression.

Stage	Maximum BA of Female	Minimum BA of Female	Maximum BA of Male	Minimum BA of Male
S_1	2.5	0.5	2.7	1.0
S_2	15.5	5.0	16.0	4.4
S_3	18.0	12.0	19.0	13.5

3. Results

In our experiments, we use DenseNet169 as the backbone of CNNs, and the baseline represents the original DenseNet169 architecture, and directly classify the three stages in

Table 1. The branch 1 column represents the results using three independent autoencoders and scoring networks. The branch 2 column uses the exact same architecture as the baseline. The difference is that the training data is augmented by mixup. DIIBAA is the final staged result of the system in this paper. The th_score calculated by Algorithm 2 is 0.45 and 0.6 for female and male, respectively, according to the validation set results. These results are shown in Tables 5 and 6, which represent boys and girls, respectively. The following accuracies are all test set accuracies, 80% are training and validation sets, and 20% are test sets. The amount of data is shown in Table 1 above.

Table 5. Results on female.

Stage	Baseline	Branch 1	Branch 2	DIIBAA ($th_Score = 0.45$)
S ₁	0.00%	50.00%	41.67%	66.67%
S ₂	100.00%	86.14%	95.67%	99.13%
S ₃	0.00%	53.84%	30.77%	61.54%

Table 6. Results on male.

Stage	Baseline	Branch 1	Branch 2	DIIBAA ($th_Score = 0.5$)
S ₁	33.33%	50.00%	41.67%	75.00%
S ₂	93.14%	89.14%	94.86%	97.71%
S ₃	0.00%	59.26%	66.67%	77.78%

As these two tables show, the baseline, S₁ and S₃ have almost no recognition ability, and even in females, all inputs are predicted as S₂. Although the performance of the autoencoder drops in S₂, it has at least some discriminative ability in the two minority stages S₁ and S₃, the drop in the S₂ stage does not imply that the baseline has a better ability to identify S₂, for example, for male, it just predicts all inputs as S₂. In branch 1, although its performance is worse than that of autoencoder at both ends, it has less degradation to S₂, which allows us to use the results of branch 2 to improve the insufficiency of branch 1 for S₂ accuracy.

Next, in Tables 7 and 8, we list the results for all th_scores . It can be found that the th_score needs to be high enough to have good performance. This is because when the th_score is too low (because branch 1 uses the image quality score as the classification basis) the image quality is not much different, and the result of branch 1 is easily susceptible to errors. Therefore, the th_score is used to adjust the output result, so that the classification result of branch 1 is used when th_score is appropriately high, otherwise the result of branch 2 is used as the finally result.

Table 7. th_score on female.

	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
S ₁	58.33%	58.33%	58.33%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%
S ₂	96.53%	96.70%	97.40%	98.27%	99.13%	99.13%	99.13%	99.13%	99.13%	99.13%
S ₃	53.85%	53.85%	61.54%	61.54%	61.54%	61.54%	61.54%	61.54%	61.54%	61.54%

Table 8. th_score on male.

	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
S ₁	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	75.00%
S ₂	92.00%	93.14%	93.71%	94.29%	94.86%	95.43%	95.43%	97.14%	97.71%	97.71%
S ₃	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	70.37%	74.07%	74.07%	77.78%

Here, we must compare with other methods for data imbalance problem alleviation. We adopt three methods commonly used to alleviate the data imbalance on CNN on our

data set. We employ these methods on the task of bone age stage identification task and use their staged identification results to input into the corresponding bone age estimation regression network. We compare the application research of these three methods in the medical images. The three methods are: oversampling method, ref. [14] using DCGAN [27] to perform data augmentation for the minority classes, ref. [21] using focal loss [28] to deal with data imbalance problem, ref. [17] is based on transfer learning (ImageNet as the source domain dataset) and data augmentation to deal with the imbalance problem. In the first two methods, we still use densenet169 as the backbone network. In [17], since they pointed out that Xception is the most superior in their method, we use Xception as the backbone network in the method of [17]. Tables 9 and 10 are the comparison of accuracy and MAE, respectively.

Table 9. Accuracy on three common methods for data imbalance alleviation.

Gender	Ref. [14] DCGAN-Based Data Augmentation	Ref. [21] Focal Loss	Ref. [17] Transfer Learning on Xception	Our Method (DIIBAA)
Female (S ₁)	30.77%	46.15%	61.54%	66.67%
Female (S ₂)	99.56%	92.67%	96.55%	99.13%
Female (S ₃)	0.00%	41.66%	25.00%	61.54%
Male (S ₁)	7.41%	70.37%	59.25%	75.00%
Male (S ₂)	99.42%	92.00%	97.14%	97.71%
Male (S ₃)	25.00%	25.00%	0.00%	77.78%

Table 10. MAE on three common methods for data imbalance alleviation.

Gender	Ref. [14] DCGAN-Based Data Augmentation	Ref. [21] Focal Loss	Ref. [17] Transfer Learning on Xception	Our Method (DIIBAA)	w/o Staging
Female	1.24 (year)	1.33 (year)	1.28 (year)	1.06 (year)	8.86 (year)
Male	1.94 (year)	1.83 (year)	1.67 (year)	1.34 (year)	8.88 (year)

As shown in Tables 9 and 10, DIIBAA outperforms the other three methods on the problem of this article. We explain these reasons for each method separately. There are two main problems with the DCGAN-based method. First, if the DCGAN cannot be well trained, their help for data augmentation is limited. In addition, as we mentioned earlier, this kind of oversampling method usually just augments the existing data, they are helpful in the expansion of the decision boundary, but since the amount of data in our two stages is too scarce, this method will have ineffective results. For focal loss, it strengthens training on hard samples (equivalent to the minority stages in this article), and minority stages are actually too scarce, so the focal loss is still not enough for training in hard samples. Therefore, its method has a considerable improvement over the baseline network, but still not accurate enough. In the transfer learning-based method, since this method does not directly design data imbalance, it may be more helpful for slightly imbalanced data, but it performs poorly for the more difficult cases of this article.

In the comparison of MAE, due to the stage prediction error, the error caused by it is even greater, and DIIBAA is still better than the above three methods because of its error rate. Of course, the MAE of DIIBAA has better results.

Finally, we use experiments to explain why we need to classify the mature stages of bone age first, instead of directly using a regression CNN for all stages. We calculate the MAE of the two to compare, and the results are shown in last column of Table 10.

It can be found that when the bone age is directly regressed without distinguishing stages, the error of their MAE is very large, compared with the first stage. This is because when all stages are mixed training, the range between them is larger and the variation is larger. This makes CNN regression harder to fit. When the stages are separated, each independent network fit is less problematic and therefore more likely to have accurate results.

4. Discussion

Although our results have been greatly improved compared to the case where the baseline has almost no discriminative ability, there is still the possibility of further improvement in the future. In the branch 1 part, there still can be improvement in the S1 and S3 stages. In future, we can try to enhance it in the following ways, such as adding richer perturbation of the autoencoder, and the improvement of the loss function of the scoring network, etc. I believe that under the framework of this paper, there is still a certain potential to further improve this extremely imbalanced data. In the part of mixup augmentation, this is a way of oversampling. Maybe we can think about how to make the data generated in the process of oversampling more diverse, so that the decision boundary of a small number of classes can be more practically expanded, and then make the branch 2 network more distinguishable. If the two branches can be further improved, I believe that the DIIBAA system can give more reliable classification results.

Additionally, this paper uses autoencoders and scoring networks to build a system to alleviate data imbalanced problem. We also suggest that this method is quite suitable to alleviate the data imbalance problem encountered by other medical images, but there is a problem to consider, we convert the data ratio to 1:2. However, if the dataset has many classes, such as a dataset with 100 classes, then the ratio of DIIBAA will become 1:99 in the training of the scoring network. At this time, we can consider the operation of undersampling the F category, which is different from directly undersampling a large number of categories, because in 99% of the data in F, each data actually has another 99 of the same origin. For the data generated by the graph, each original graph has a high probability to exist in F, and must exist in T, so data will not be wasted.

Finally, with the migration of the ages, the living conditions of human beings will inevitably change, and with the changes in eating habits, living habits, etc., the developmental process of children will inevitably change, and the method in this article is built using a changeless data set (2018 to 2020), in the face of future changes, the system must be able to update the model with the times, such as adding online learning methods.

5. Conclusions

Automatic BAA is an important task for pediatrics, and in order to accurately assess bone age, in the experiment, we show that assessment of bone age in divided stages is a better way. Therefore, the stage is independently given to a regression network before we must first perform stage prediction on the unknown X-ray films, and in order to train the stage discriminant classifier, we meet a serious data imbalance problem. If we directly use them for CNN training, the CNN has almost no discriminative ability. It is just that almost all samples are predicted as a large number of stages. Therefore, this paper mainly converts the classification problem into an image quality assessment problem, because the autoencoders in these stages are independent and do not interfere with each other, which naturally avoids data imbalance problems. In addition, considering that the classification features between these X-ray films are also valuable information, in order to effectively utilize them, we use mixup data augmentation to make branch 2 in DAIIBA system have some classification ability, and finally we use *th_score* to leverage the advantages of both of them. Finally, we also show that the correct prediction of bone age stage is helpful for the prediction of bone age value.

Author Contributions: Methodology, C.-T.P.; Resources, Y.-K.C. and S.-S.Y.; Supervision, Y.-K.C. and S.-S.Y.; Writing—original draft, C.-T.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Garn, S.M. Radiographic atlas of skeletal development of the hand and wrist. *Am. J. Hum. Genet.* **1959**, *11*, 282.
- Tanner, J.M. *Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method)*; Academic Press: Cambridge, MA, USA, 1983; pp. 50–106.
- Tanner, J.M.; Goldstein, H.; Cameron, N. *Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method)*; Wiley: Hoboken, NJ, USA, 2001.
- Ren, X.; Li, T.; Yang, X.; Wang, S.; Ahmad, S.; Xiang, L.; Stone, S.R.; Li, L.; Zhan, Y.; Shen, D.; et al. Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment from Hand Radiograph. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 2030–2038. [[CrossRef](#)] [[PubMed](#)]
- He, J.; Jiang, D. Fully automatic model based on se-resnet for bone age assessment. *IEEE Access* **2021**, *9*, 62460–62466. [[CrossRef](#)]
- Lee, H.; Tajmir, S.; Lee, J.; Zissen, M.; Yeshiwas, B.A.; Alkasab, T.K.; Choy, G.; Do, S. Fully automated deep learning system for bone age assessment. *J. Digit. Imaging* **2017**, *30*, 427–441. [[CrossRef](#)]
- Dehghani, F.; Karimian, A.; Siros, M. Assessing the bone age of children in an automatic manner newborn to 18 years range. *J. Digit. Imaging* **2019**, *33*, 399–407. [[CrossRef](#)]
- Halabi, S.S.; Prevedello, L.; Kalpathy-Cramer, J.; Mammonov, A.B.; Bilbily, A.; Cicero, M.; Pan, I.; Pereira, L.A.; Sousa, R.; Abdala, N.; et al. The RSNA pediatric bone age machine learning challenge. *Radiology* **2019**, *290*, 498–503. [[CrossRef](#)]
- Hertel, R.; Benlamri, R. COV-SNET: A deep learning model for X-ray-based COVID-19 classification. *Inform. Med. Unlocked* **2021**, *24*, 100620. [[CrossRef](#)]
- Marginean, A.N.; Muntean, D.D.; Muntean, G.A.; Priscu, A.; Groza, A.; Slavescu, R.R.; Timbus, C.L.; Munteanu, G.Z.; Morosanu, C.O.; Cosnarovici, M.M.; et al. Reliable learning with pde-based cnns and densenets for detecting covid-19, pneumonia, and tuberculosis from chest x-ray images. *Mathematics* **2021**, *9*, 434. [[CrossRef](#)]
- Khatami, A.; Babaie, M.; Khosravi, A.; Tizhoosh, H.R.; Nahavandi, S. Parallel deep solutions for image retrieval from imbalanced medical imaging archives. *Appl. Soft Comput.* **2018**, *63*, 197–205. [[CrossRef](#)]
- Vieira, P.; Sousa, O.; Magalhaes, D.; Rabelo, R.; Silva, R. Detecting pulmonary diseases using deep features in x-ray images. *Pattern Recognit.* **2021**, *119*, 108081. [[CrossRef](#)]
- Motamed, S.; Rogalla, P.; Khalvati, F. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Inform. Med. Unlocked* **2021**, *27*, 100779. [[CrossRef](#)] [[PubMed](#)]
- Saini, M.; Susan, S. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Appl. Soft Comput.* **2020**, *97*, 106759. [[CrossRef](#)]
- Rahman, M.; Cao, Y.; Sun, X.; Li, B.; Hao, Y. Deep pre-trained networks as a feature extractor with xgboost to detect tuberculosis from chest x-ray. *Comput. Electr. Eng.* **2021**, *93*, 107252. [[CrossRef](#)]
- Mamalakakis, M.; Swift, A.J.; Vorselaars, B.; Ray, S.; Weeks, S.; Ding, W.; Clayton, R.H.; Mackenzie, L.S.; Banerjee, A. Denrescov-19: A deep transfer learning network for robust automatic classification of covid-19, pneumonia, and tuberculosis from x-rays. *Comput. Med. Imaging Graph.* **2021**, *94*, 102008. [[CrossRef](#)] [[PubMed](#)]
- Lujan-García, J.E.; Yanez-Marquez, C.; Villuendas-Rey, Y.; Camacho-Nieto, O. A transfer learning method for pneumonia classification and visualization. *Appl. Sci.* **2020**, *10*, 2908. [[CrossRef](#)]
- Zhou, Y.-J.; Xie, X.-L.; Zhou, X.-H.; Liu, S.-Q.; Bian, G.-B.; Hou, Z.-G. A real-time multifunctional framework for guidewire morphological and positional analysis in interventional x-ray fluoroscopy. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *13*, 657–667. [[CrossRef](#)]
- Calderon-Ramirez, S.; Yang, S.; Moemeni, A.; Elizondo, D.; Colreavy-Donnelly, S.; Chavarría-Estrada, L.F.; Molina-Cabello, M.A. Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images. *Appl. Soft Comput.* **2021**, *111*, 107692. [[CrossRef](#)]
- Aviles-Rivero, A.I.; Sellars, P.; Schönlief, C.-B.; Papadakis, N. Graphxcovid: Explainable deep graph diffusion pseudo-labelling for identifying covid-19 on chest x-rays. *Pattern Recognit.* **2022**, *122*, 108274. [[CrossRef](#)]
- Tran, G.S.; Nghiem, T.P.; Nguyen, V.T.; Luong, C.M.; Burie, J.-C. Improving accuracy of lung nodule classification using deep learning with focal loss. *J. Healthc. Eng.* **2019**, *2019*, 5156416. [[CrossRef](#)]
- Tschannen, M.; Bachem, O.; Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv* **2018**, arXiv:1812.05069.
- Gilsanz, V.; Ratib, O. *Hand Bone Age: A Digital Atlas of Skeletal Maturity*; Springer: Berlin/Heidelberg, Germany, 2005.
- Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
- Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
- Lefkimmiatis, S. Universal denoising networks: A novel CNN architecture for image denoising. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA, 18–22 June 2018; pp. 3204–3213.
- Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.