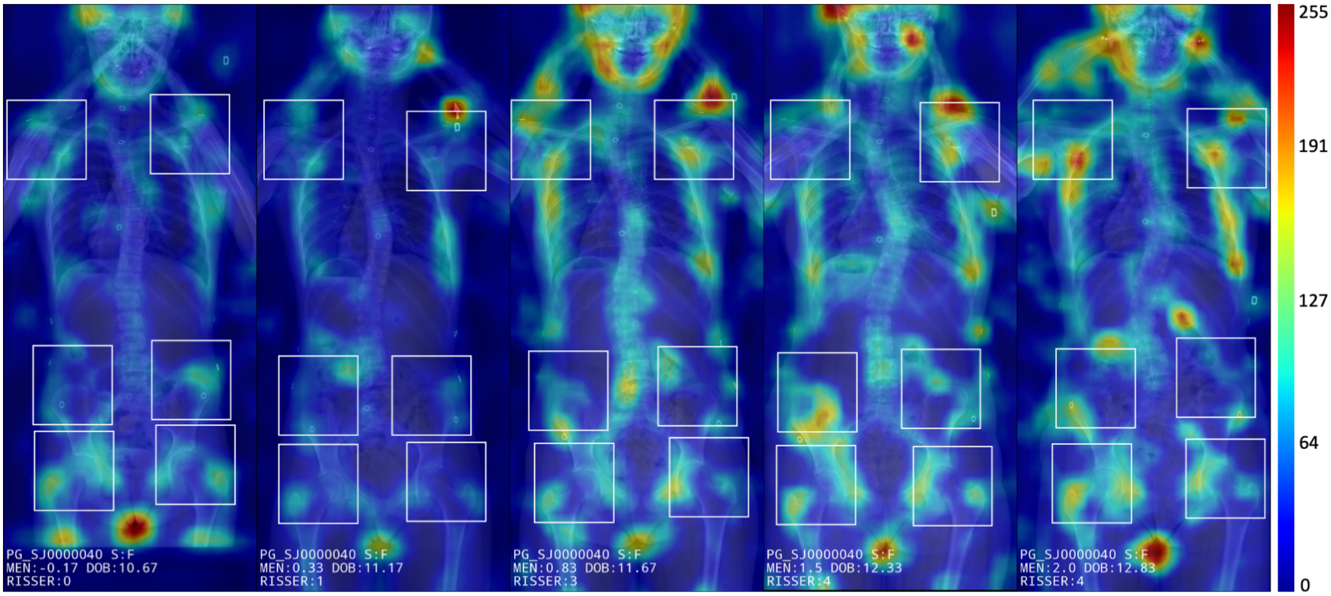
Authors' accepted manuscript
Article published in Computers in *Biology and Medicine*, Volume 136, September 2021, 104681
Version of record available at https://doi.org/10.1016/j.compbiomed.2021.104681

© 2021. This manuscript version is made available
under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Graphical Abstract

Automatic bone maturity grading from EOS radiographs in Adolescent Idiopathic Scoliosis

Eddie Magnide,Georges Wona Tchaha,Julie Joncas,Christian Bellefleur,Soraya Barchi,Marjolaine Roy-Beaudry,Stefan Parent,Guy Grimard,Hubert Labelle,Luc Duong



Authors' accepted manuscript
Article published in Computers in *Biology and Medicine*, Volume 136, September 2021, 104681
Version of record available at https://doi.org/10.1016/j.compbiomed.2021.104681

© 2021. This manuscript version is made available
under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Highlights

Automatic bone maturity grading from EOS radiographs in Adolescent Idiopathic Scoliosis

Eddie Magnide,Georges Wona Tchaha,Julie Joncas,Christian Bellefleur,Soraya Barchi,Marjolaine Roy-Beaudry,Stefan Parent,Guy Grimard,Hubert Labelle,Luc Duong

- Novel method for grading of skeletal maturity using machine learning in Adolescent Idiopathic Scoliosis (AIS)
- Automatic feature extraction using transfer learning from full standing radiographs
- Investigation of different ossification centers and validation at different key visits for follow up
- Visualization of class activation maps

Authors' accepted manuscript
Article published in Computers in *Biology and Medicine*, Volume 136, September 2021, 104681
Version of record available at https://doi.org/10.1016/j.compbiomed.2021.104681

© 2021. This manuscript version is made available
under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Automatic bone maturity grading from EOS radiographs in Adolescent Idiopathic Scoliosis

Eddie Magnide[a], Georges Wona Tchaha[a], Julie Joncas[b], Christian Bellefleur[b], Soraya Barchi[b], Marjolaine Roy-Beaudry[b], Stefan Parent[b,c], Guy Grimard[b,c], Hubert Labelle[b,c] and Luc Duong[a]

[a] *Department of Software and IT Engineering, École de technologie supérieure, Montreal, Canada*
[b] *Department of orthopedics, Sainte-Justine Hospital, Montreal, Canada*
[c] *Université de Montréal, Montreal, Canada*

ARTICLE INFO

Keywords:
Risser sign
Skeletal maturity
Bone aging
Convolutional neural networks
Scoliosis

ABSTRACT

Adolescent Idiopathic Scoliosis (AIS) is a deformation of the spine and it is routinely diagnosed using posteroanterior and lateral radiographs. The Risser sign used in skeletal maturity assessment is commonly accepted in AIS patient's management. However, the Risser sign is subject to inter-observer variability and it relies mainly on the observation of ossification on the iliac crests. This study proposes a new machine-learning-based approach for Risser sign skeletal maturity assessment using EOS radiographs. Regions of interest including right and left humeral heads; left and right femoral heads; and pelvis are extracted from the radiographs. First, a total of 24 image features is extracted from EOS radiographs using a ResNet101-type convolutional neural network (CNN), pre-trained from the ImageNet database. Then, a support vector machine (SVM) algorithm is used for the final Risser sign classification. The experimental results demonstrate an overall accuracy of 84%, 78%, and 80% respectively for iliac crests, humeral heads, and femoral heads. Class activation maps using Grad-CAM were also investigated to understand the features of our model. In conclusion, our machine learning approach is promising to incorporate a large number of image features for different regions of interest to improve Risser grading for skeletal maturity. Automatic classification could contribute to the management of AIS patients.

1. Introduction

Adolescent idiopathic scoliosis (AIS) is a 3D deformation of the spine that affects 1 to 3% of the at-risk population aged 10–16 years [45]. It is caused by various factors including genetic, biochemical, and morphological [12]. The deformation can worsen over time. AIS is treated by prescribing a brace, and in severe cases, the patient must undergo corrective surgery. Spine surgeons rely mostly on the monitoring of the scoliotic curve using posteroanterior and lateral radiographs. The severity of deformation is measured using the Cobb angle. The Cobb angle is determined on the posteroanterior radiograph by selecting the most tilted vertebra at the top and bottom of the curve [3]. The summation of the angles of these two vertebrae relative to the horizontal is the Cobb angle [3].

1.1. Bone maturity assessment in AIS

Bone maturity assessment is another important component of the management of AIS. Several authors have attempted to estimate bone age in different ways by comparing patients to reference X-rays. In general, the most used method is that of Greulich and Pyle performed on the left hand and wrist [16]. Another method, the Sanders maturity scale, based on a left hand radiograph, appears to be strongly prognostic of future scoliosis curve behavior [36]. Particularly, the thumb ossification composite index (TOCI) based on ossification of the thumb has demonstrated simplicity and high accuracy for predicting skeletal maturity, comparable with the Sanders maturity scale [22]. For its part, the method of Pyle and Hoerr targets the knee [2]. The one which is the most common in AIS treatment is the Risser sign [11]. The Risser sign is determined by observation of the growing ossification plates/cartilages of iliac crests. Two classifications American and French have been defined [17]. The American version is divided into 6 stages as shown in figure 1, namely:

- **Risser 0** where no illiac apophysis visible (ossification is absent),
- **Risser 1** where initial appearance of ossification of the illiac apophysis (ossification is 25%),
- **Risser 2** where migration halfway across the top of the illiac wing (ossification is at 50%),
- **Risser 3** where ossification is 75% of the distance,
- **Risser 4** where the ossification is crossing the iliac wing, but not fused to the illum and
- **Risser 5** where complete ossification of the iliac apophysis with fusion with the illum.

Given the subjective nature of the Risser sign, several studies have reported inter and intra-variability. Lack of consensus among clinicians affects the accuracy of the measurement. Some studies on measurement variability have shown acceptable results [15, 14]. However, the latest studies are less optimistic and consider agreement moderate [37, 23]. While the formal definition of the Risser sign relies on the observation of the iliac crests, other centers of ossification can be explored for proper grading. Thus,

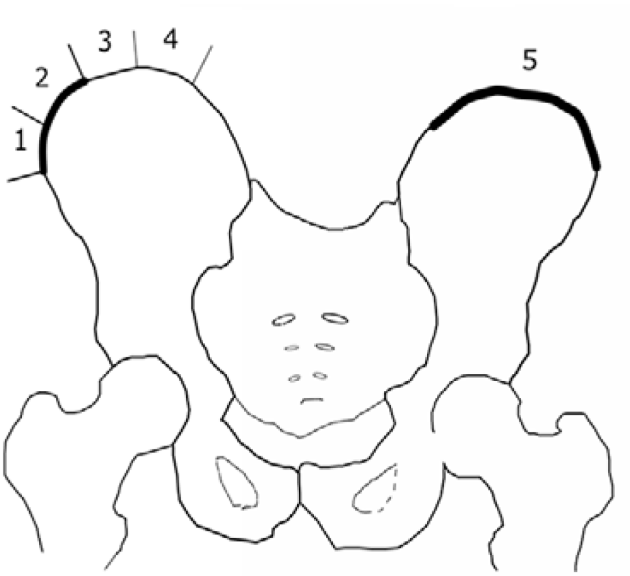


Figure 1: Visual illustration of iliac crest progression and corresponding Risser stages for American version [23] (At stage 0, there is no ossification until stage 6 when ossification becomes complete.)

Kaddioui et al. [23] demonstrated that the entire pelvic region could be used to better determine the Risser sign. Besides, several metrics make it possible to estimate bone age by observing other regions of the body [2].

1.2. Deep learning for classification of bone maturity

Recently, machine learning, especially deep learning methods were found to be successful at classifying automatically bone structures from radiographs and 3D models. Features extraction is performed using transfer learning based on a convolutional neuronal network (CNN). The wide availability of pre-trained models on natural images complexifies the choice of a proper model for bone aging classification. So far, the following models have been popularized: AlexNet, VGG, Inception, or U-Net.

Inception-ResNet model made it possible to detect and locate radius and ulna fractures on X-ray wrist with high sensitivity and specificity [42]. Pan et al. have shown that an automatic CNN-based approach using hand X-ray could facilitate the assessment of bone age [32]. Also, Rayan proposed to efficiently classify pediatric elbow abnormalities using CNN for diagnosis [33]. CNNs are now very useful for automating bone aging classification [43, 34, 25, 31, 18]. Tong et al. propose a model which consists of two modules: one is for feature extraction by CNNs (VGG architecture) and the other is for bone age estimation by support vector regression (SVR) using multiple kernel learning (MKL) [43]. In the same way, Ren et al. propose a fully automated deep learning solution to process X-ray images of the hand for bone age assessment using a CNN based on Inception-V3 architecture [34]. Lee et al. exploit transfer learning with a pre-trained CNN based on GoogLeNet to automatically extract important features from all bones on an ROI that was automatically segmented by a detection CNN [25]. Another study demonstrated the benefits of a customized neural network algorithm carefully calibrated to the evaluation of bone age utilizing a relatively large dataset [31]. Finally, an approach based on Inception-V3 architecture using images augmented and concatenated with the sex information won first place in a machine learning challenge [18]. These are all methods that have demonstrated the utility of CNNs in determining bone age.

CNN models have gained wide popularity in digital radiology. This approach often targets a clinical question related to anomaly detection or clinical classification [47]. To achieve this, Soffer et al [39] propose a methodology in 7 steps. After having formulated the problem, it is then necessary to carry out computer vision tasks which consist of classifying, detecting, segmenting, or optimizing input images [39]. Then data must be separated into training, validation, and test data. Preprocessing step is sometimes necessary to augment data and thus avoid overfitting [39]. At the implementation step, we must choose the most suitable hardware (GPU), software platform, and CNN architecture. Finally, the validation step is important to demonstrate the effectiveness of the model.

CNNs require a lot of images as input. CNNs demands an extensively large amount of data to achieve a well-behaved performance model [4]. However, it is sometimes very complicated to obtain clinical data given cost and confidentiality. To remedy this, two solutions are available to us: data augmentation or transfer learning [47]. Data augmentation is used to modify training data by transforming it randomly so that model does not see the same inputs during training iterations [49]. However, data augmentation includes artificially manipulated data, which can introduce bias in the training process. Transfer learning is another strategy to train a model on a small dataset. The idea is to harness the capabilities of the pre-trained model to extract the most important information from input images. Many models pre-trained are open to the public such as AlexNet [24], VGG [38], ResNet [19], Inception [41], and DenseNet [20].

The texture feature extraction methods are divided into seven classes: statistical approaches, structural approaches, transform-based approaches, model-based approaches, graph-based approaches, learning-based approaches, and entropy-based approaches [21]. Among extraction methods, the statistical approach GLCM (grey level co-occurrence matrix) is largely used [21]. It has been shown that, the GLCM-based approach has good performance in terms of processing time and complexity but for images with a large amount of noise, the GLCM features are not appropriate [30]. In the last few years, learning-based approaches in particular CNN-based have significantly been used [21]. If the textures have very large within-class appearance variations, CNN-based methods clearly perform the best, however at a cost of high computational complexity [28].

This study presents a new approach for bone maturity grading using different ossification centers. Our approach is validated on a large cohort of patients acquired at a different stage of their bone maturity. This work is organized as follows: Section 2 presents the methodology and the database used to validate our work, Sections 3 presents the results and Section 4 the discussion. Finally, in Section 5, a conclusion is provided.

2. Materials and Methods

2.1. Database description

A research protocol was submitted to the ethics review board and approval was obtained before the experimentation. A cohort of 53 AIS patients (88% of female, age = 13,65 ± 1,23 years), spanning over 5 visits at different Risser sign was recruited for this study at Sainte-Justine Hospital. The EOS radiographs were collected for each visit. EOS radiographs are acquired using two semi-simultaneous low-dose detectors to acquire a full or mid-body X-ray image of the patient while limiting the amount of patient irradiation [46]. EOS radiography is generated using the principle of the multi-wire proportional chamber developed by Charpak [9]. From the EOS radiographs, 6 regions of interest (the 2 left and right iliac crests, the 2 left and right femoral heads then the 2 left and right humeral heads) were extracted manually. Each sample is associated with:

- 6 regions of interest of size 600x600 pixels,
- the date of the visit,
- and the Risser sign labeled by the clinician.

Figure 2 illustrates images available for each sample. The modeling is done region of interest by region of interest. Left and right images will be considered as separate entries. Images are saved in a file structure organized by patient and by date while the annotations are recorded in a spreadsheet. Once the data is well structured and labeled, the next step will be to clean and balance it.

2.2. Data pre-processing

Before starting the modeling, a review of input data available allowed us to fill in the missing values with a regrading of the patient and, in some cases, to exclude missing data. After examination, 16 inconsistent examples were deleted while 4 missing annotations were re-estimated using the definition of the Risser sign (Figure 1). The size of the corpus went from 530 to 514 segmented images. Risser 0 dominates while Risser 2 and 5 present fewer samples (see figure 3A). To reduce this imbalance, there are 2 basic methods: sub-sampling by eliminating examples of majority classes or over-sampling by duplicating examples of minority classes. In the two methods, either one loses information which could have been useful, or one creates an over-fitting. On the other hand, the SMOTE-Tomek technique, for its part, combines over-sampling and under-sampling to find an acceptable compromise and reduce the possible errors linked to these two basic methods [5]. This is the main argument that motivated the choice of SMOTE-Tomek technique. After having applied this strategy, the data is better distributed (see figure 3B). At the end of this step, the pre-processed images are used to extract the features in 2 ways: first using transfer learning and then using the GLCM technique.

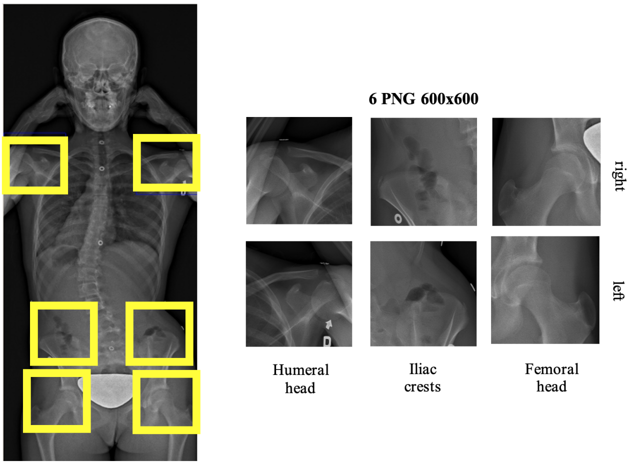


Figure 2: Raw data per visit per patient. (Each frontal EOS radiography is segmented into 6 sub-regions: right and left iliac crests, right and left humeral heads, and right and left femoral heads. Each image has a size of 600x600 pixels with gray-scale values.)

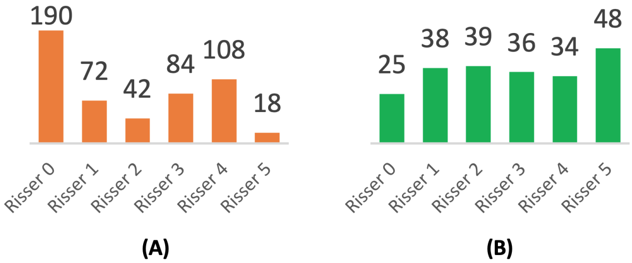


Figure 3: Distribution of Risser classes before and after the SMOTE-Tomek technique ((In figure 3A, we observe a large imbalance within the Risser classes using the 514 segmented and preprocessed images. SMOTE-Tomek technique reduced imbalance as can be seen in figure 3B. We present here the example of iliac crests with features extracted using transfer learning.)

2.3. Feature extraction: transfer learning

Transfer learning is convenient to adapt an existing model, learned from a large amount of data to a specific learning task. The parameters of the already trained CNN model are used to extract image features. The pre-trained CNN chosen is based on the ResNet101 architecture [27](Figure 4A). Pre-training is performed using the ImageNet database which contains millions of images [13]. As illustrated in figure 4B, the implemented CNN is composed of 4 groupings of Conv layers: res2,

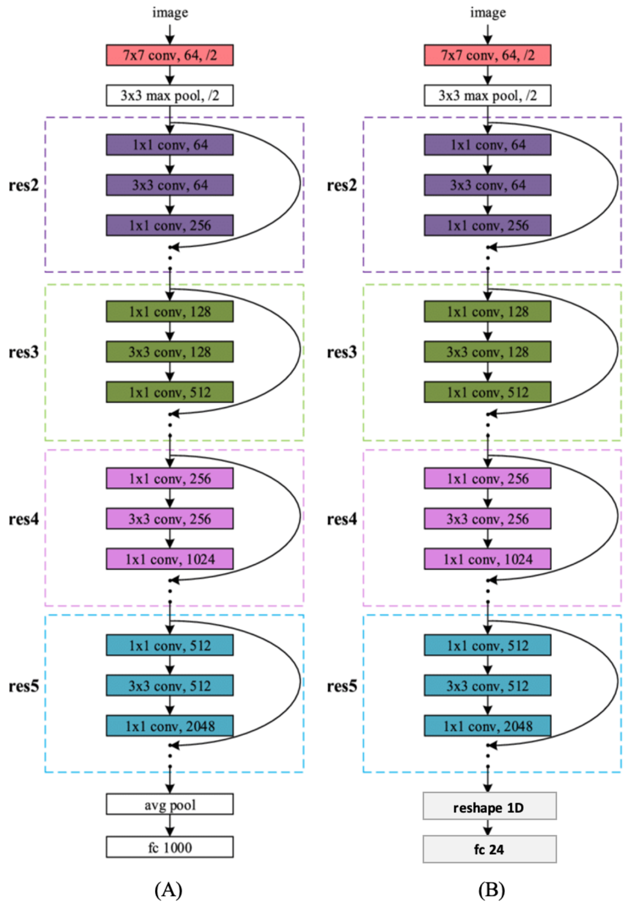


Figure 4: Architectures of ResNet101 (A) and pre-trained CNN are used to extract features (B) [27]. (In figure 4A, the original architecture of ResNet101. In figure 4B, pre-trained CNN is used for extraction after modification of the last 2 layers.)

res3, res4, and res5. Then, the last 2 layers of ResNet101 (Pool and FC) are replaced to generate 24 features as output. The software was developed in Python 3.7.4 using the Keras library with the TensorFlow library for deep learning [1] on an NVIDIA Quadro P2200 GPU.

2.4. Feature extraction: GLCM

To demonstrate the usefulness of our approach based on transfer learning, we will extract features using the GLCM or the grayscale co-occurrence matrix method technique. The goal is to then compare the results from the two approaches. The GLCM technique is a simple algorithm generally used to extract features from an image and will be used for baseline comparison with the deep features. GLCM is texture characteristics indicating the number of times that different combinations of gray levels are repeated in an image at a defined distance and angle. Four values are extracted for each of the 6 following properties: contrast, energy, homogeneity, correlation, dissimilarity, and second angular momentum (ASM). In total, we obtained a 1D vector of 24 features that represent the image. The implementation is done in Python 3.7.4 using Scikit-Image library [44]. To predict the Risser sign, the 24 features extracted for each region of interest are then entered into a classification algorithm.

2.5. Feature classification: SVM

Support Vector Machines (SVM) are supervised learning techniques that transform original space into higher dimensional space based on a kernel function. Then, the algorithm finds support vectors to maximize the separation between the classes. To train an SVM, 2 parameters must be adjusted:

- the kernel function which can be a linear function, a polynomial function with a given degree n, a Radial Basis Function (RBF) with a given gamma or a sigmoid function
- and the margin C which defines the separation distance between the classes.

After sampling with the SMOTE-Tomek technique, data are then separated at random with 80% for training-validation and 20% for testing. The Radial Basis Function (RBF) kernel was used given its ability to handle nonlinear boundaries between classes. This function is characterized by its gamma parameter which defines the shape of the curvature of the decision boundary. Thus, for our training, the value of gamma parameter varies between 0.0325 and 100 while the margin parameter C rather varies between 0.1 and 100. These two parameters were optimized using grid-search method that searches exhaustively through a manually specified subset of the parameter space [6].

The optimization is based on 10-fold cross validation [35]. The training-validation data set is divided into 10 folds : 9-folds for training and 1-fold for validation. The final error is estimated by averaging the errors committed in each fold [35]. After finding the optimal parameters for the SVM classifier, the testing data was used to evaluate both models. The evaluation is performed for each of the regions of interest using confusion matrix, precision, and recall. To go further in the analysis, the visualization of the data in the light of the classification will be carried out.

2.6. Data visualization: Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced to further understand the classification results. Grad-CAM offers a mathematical derivation that uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps [10]. The purpose of Grad-CAM is to find regions in an image that contributes the most to classification. Visualization of these could reveal new regions of interest. To generate Grad-CAM, we use a pre-trained CNN based on ResNet101 architecture [27] as shown in figure 4A. The target is the last layer convolution (res5).

3. Results

After data pre-processing, the ResNet101-SVM model is trained using features extracted for each region of interest.

Our model completes the training after 15 minutes per region of interest. At the end of the training, when the optimal C and gamma parameters are found, we evaluated and compared the ResNet101-SVM model to the GLCM-SVM model. Tables 1 and 2 illustrate the results of evaluation and comparison. The accuracy (Acc.) refers to overall accuracy for all Risser stages. The precision (Prec.) refers to precision for each Risser stage.

Table 1
Comparison between GLCM-SVM and ResNet101-SVM models for each region of interest

Iliac crests (IC)	GLCM-SVM						
	Risser	0	1	2	3	4	5
	Support	25	34	36	37	32	41
	Recall	0.40	0.74	0.86	0.81	0.88	0.95
	F1 score	0.44	0.75	0.86	0.82	0.82	0.92
	Prec.	**0.50**	0.76	0.86	0.83	0.78	0.89
	Acc.	0.80					
	ResNet101-SVM						
	Risser	0	1	2	3	4	5
	Support	25	38	39	36	34	48
	Recall	0.76	0.95	0.97	0.78	0.59	0.90
	F1 score	0.58	0.92	0.93	0.82	0.73	0.93
	Prec.	0.47	**0.90**	**0.88**	**0.88**	**0.95**	**0.98**
	Acc.	**0.84**					
Humeral heads (HH)	GLCM-SVM						
	Risser	0	1	2	3	4	5
	Support	26	38	25	38	29	45
	Recall	0.38	0.76	0.88	0.61	0.66	0.80
	F1 score	0.43	0.72	0.71	0.64	0.69	0.85
	Prec.	**0.48**	**0.76**	0.86	0.83	0.78	0.89
	Acc.	0.69					
	ResNet101-SVM						
	Risser	0	1	2	3	4	5
	Support	32	31	45	40	28	42
	Recall	0.69	0.81	0.87	0.75	0.82	0.74
	F1 score	0.52	0.77	0.91	0.85	0.81	0.85
	Prec.	0.42	0.74	**0.95**	**0.97**	**0.79**	**1.00**
	Acc.	**0.78**					
Femoral heads (FH)	GLCM-SVM						
	Risser	0	1	2	3	4	5
	Support	26	36	34	30	35	43
	Recall	0.62	0.69	0.88	0.80	0.71	1.00
	F1 score	0.56	0.77	0.87	0.76	0.77	0.97
	Prec.	**0.52**	0.86	**0.86**	0.73	**0.83**	0.93
	Acc.	0.80					
	ResNet101-SVM						
	Risser	0	1	2	3	4	5
	Support	25	38	43	35	38	34
	Recall	0.60	0.74	0.93	0.80	0.66	1.00
	F1 score	0.48	0.81	0.89	0.85	0.70	1.00
	Prec.	0.41	**0.90**	0.85	**0.90**	0.76	**1.00**
	Acc.	0.80					

For two regions of interest, overall accuracy is higher each time when using the ResNet101-SVM. Precision goes from 80% to 84% for iliac crests and from 69% to 78% for humeral heads. About femoral heads, precision is equivalent and evaluate at 80%. When we consider all regions of interest together, we obtain a precision of 79% and 76% respectively for ResNet101-SVM and GLCM-SVM. For most classes, precision and recall values are higher for ResNet101-SVM.

Graphical representations of confusion matrices are useful for checking the agreement between predicted values of a model and values annotated by the expert (ground truth). As illustrated in figure 5, we noticed a higher agreement for the ResNet101-SVM model from the confusion matrices.

Assessment of activation maps was performed by arranging side-by-side each radiograph in a single image per patient, to outline the bone maturity progression between each visit. Information of anonymized ID, sex, age, menarche, and Risser sign is marked along with regions of interest to be studied. Figures 6 and 7 shows examples of activation maps for two patients. Regions of interest previously identified (iliac crests, humeral heads, and femoral heads) present stronger activation. Apart from these three regions, we also notice the cranial bones and rib cage that have stronger activation.

4. Discussion

First, the results obtained show a significant improvement of our ResNet101-SVM compared against the GLCM-SVM and the VGG16 model [23]. The previous study using VGG16 model is built with 1830 pelvic radiographs and trained in 8 hours [23]. Comparatively, ResNet101-SVM is built with 257 EOS radiographs and trained in one quarter of hour. Moreover, we have demonstrated that different ossification centers, not only the pelvic region contributes to the classification results, and may provide additional information valuable for bone maturity assessment. Particularly, another study has demonstrated that the humeral heads can be used to gain additional insight into skeletal maturity of patients with scoliosis [26]. Regions of interest such as the cranial bones and rib cage that have demonstrated stronger activation (figures 6 and 7) can be explored and segmented for future work. In addition, the segmentation of regions of interest reduces the impact of artifacts due to implants.

Second, the ResNet101-SVM model for extracting features from the image demonstrated better performance in precision compared to the GLCM algorithm. The same is true in 2 other studies where the authors compared these approaches. Lopes and Valiati demonstrated the relevance of pre-trained CNNs in automatically determining tuberculosis from chest X-ray [29]. For their part, Byra et al. showed that the use of a pre-trained CNN for feature extraction improved results compared to the use of GLCM for the diagnosis of the amount of fat in the liver [8]. Therefore, these two studies support our conclusions even if the CNN architecture differs.

Third, we have demonstrated the ResNet101-SVM model is computationally efficient and it allow intuitive interpretation of the decision using activation maps. This visualization technique has already been used in several studies on bone maturity [7, 48, 40]. Interpretability

Table 2
Comparison between GLCM-SVM and ResNet101-SVM models using all regions of interest

		GLCM-SVM					
IC + HH + FH	Risser	0	1	2	3	4	5
	Support	98	93	107	115	99	114
	Recall	0.54	0.85	0.87	0.75	0.68	0.85
	F1 score	0.56	0.83	0.83	0.74	0.70	0.87
	Prec.	**0.58**	0.81	0.80	0.72	0.72	0.89
	Acc.	0.76					
		ResNet101-SVM					
	Risser	0	1	2	3	4	5
	Support	106	119	94	100	126	111
	Recall	0.68	0.76	0.91	0.78	0.70	0.93
	F1 score	0.57	0.79	0.90	0.78	0.80	0.94
	Prec.	0.49	**0.83**	**0.89**	**0.79**	**0.93**	**0.94**
	Acc.	**0.79**					

of the final prediction is very important in any clinical classification task. Our approach is appealling for a clinical use since the final decision is carried out in the final stage using SVMs, a well-established classification method offering good interpretability of the final results.

Despite the advantages of the proposed model, compared to the other two approaches, the limited amount of data available for training is a major issue. The use of the SMOTE-Tomek technique to compensate for the small amount of available examples is promising, but not perfect. Indeed, unlike the VGG16 approach [23], our modeling is based on resampled data. Also, the number of samples used is 7 times less than the number of samples used to implement the VGG16 model. VGG16 model achieved an overall accuracy of 78% [23] less or equal to ResNet101-SVM accuracies (84%, 78%, 80% and 79%). A comparable number of samples could have allowed a better comparison of the ResNet101-SVM and the VGG16. However, these two models were computed on two different datasets which may not be fairly compared.

Finally, for some Risser signs, GLCM-SVM performs better in precision than ResNet101-SVM. This is a weakness of our model, especially at low Risser signs. In fact, the ResNet101 CNN used is pre-trained with images from the ImageNet database. These images are classified into 1000 categories of real-life objects. The use of transfer learning allows us to look for interesting features in EOS radiographs. However, feature extraction is dependent on visible structures. The less useful information there is, the less precise the extraction. This is probably the reason why for low Risser signs our approach is less precise due to the lack of cartilage. On the contrary, the higher Risser signs, the more the precision increases because we have many more structures of ossification.

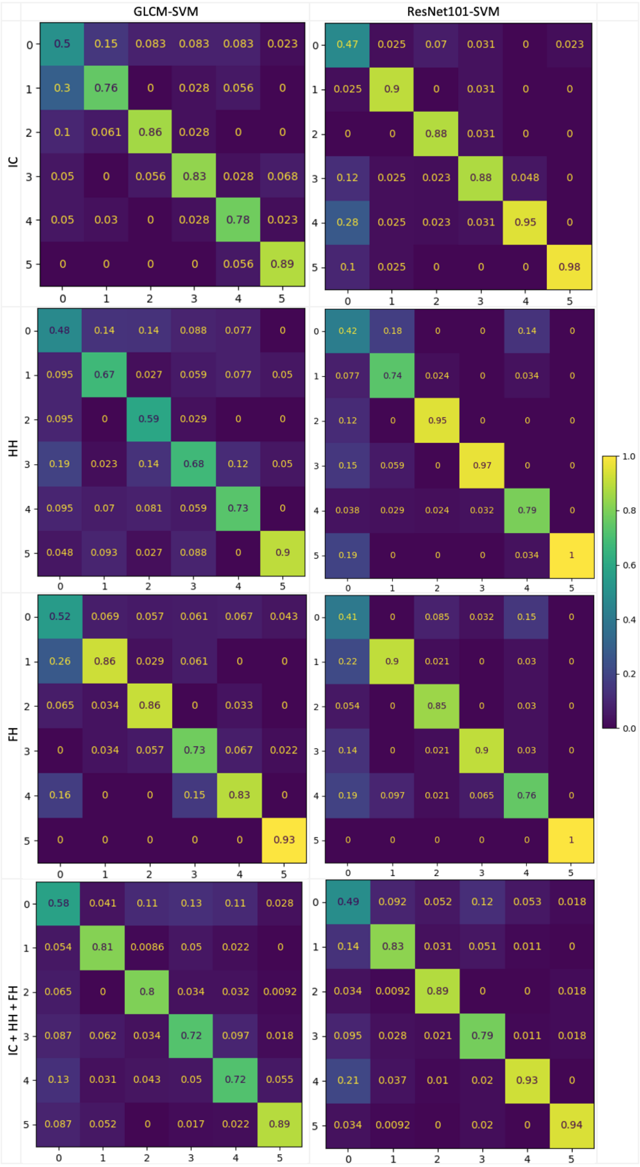


Figure 5: Comparison between the GLCM-SVM and ResNet101-SVM models with the standardized confusion matrices. (The agreement between the predicted values and those annotated by the expert is much more important in the case of ResNet101-SVM.)

5. Conclusion

We proposed a new classification approach based on the ResNet101-SVM architecture. The results demonstrated an improvement in precision compared to the GLCM-SVM approach, but also a reduction in computation time compared to the VGG16 approach [23]. In addition, this study demonstrated that other ossification centers such as the humeral heads and the femoral heads could contribute to the final prediction in our model. Our approach has a major advantage over previous studies, which is the ability to learn from the features extracted at different ossification centers and provide a decision based on these features, which is highly appealing for evaluating the final decision. Other bone maturity techniques, such Sanders maturity scale and TOCI will be investigated in future work. Also, given the availability of retrospective, longitudinal data, training of machine learning models adapted for time series, such

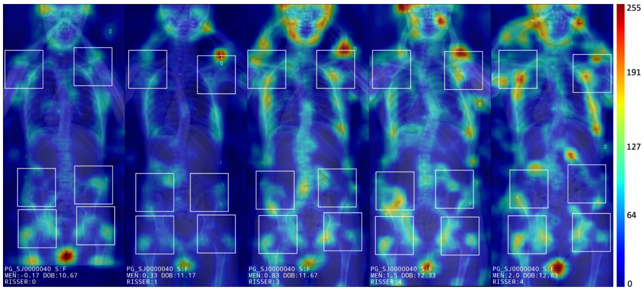


Figure 6: Grad-CAM maps for patient ID 40 over the 5 visits (The Risser sign variation : 0-1-3-4. Several regions in the images become redder, which translates to higher activation.)

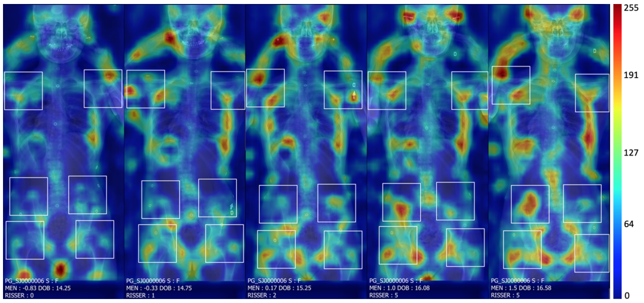


Figure 7: Grad-CAM maps for patient ID 06 over the 5 visits (The Risser sign variation : 0-1-2-5. Several regions in the images become redder, which translates to higher activation.)

as recurrent neural network (RNN) will be investigated in future work. Such models might be highly relevant to study bone progression directly from X-ray images, and it could provide a better estimate of bone maturity of AIS patients.

6. Acknowledgements

This research was funded by the *Chaire de recherche en sciences du mouvement du CHU Sainte-Justine*.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265–283.

[2] Adamsbaum, C., André, C., Merzoug, V., Kalifa, G., 2005. Âge osseux, intérêt diagnostique et limites. EMC-Pédiatrie 2, 1–11.

[3] Allen, S., Parent, E., Khorasani, M., Hill, D.L., Lou, E., Raso, J.V., 2008. Validity and reliability of active shape models for the estimation of cobb angle in patients with adolescent idiopathic scoliosis. Journal of digital imaging 21, 208–218.

[4] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. Journal of big Data 8, 1–74.

[5] Batista, G.E., Bazzan, A.L., Monard, M.C., 2003. Balancing training data for automated annotation of keywords: a case study., in: WOB, pp. 10–18.

[6] Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of machine learning research 13.

[7] Bui, T.D., Lee, J.J., Shin, J., 2019. Incorporated region detection and classification using deep convolutional networks for bone age assessment. Artificial intelligence in medicine 97, 1–8.

[8] Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michałowski, Ł., Paluszkiewicz, R., Ziarkiewicz-Wróblewska, B., Zieniewicz, K., Sobieraj, P., Nowicki, A., 2018. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. International journal of computer assisted radiology and surgery 13, 1895–1903.

[9] Charpak, G., 1996. Prospects for the use in medicine of new detectors of ionizing radiation. Bulletin de l'Academie nationale de medecine 180, 161.

[10] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 839–847.

[11] Coussement, A., 2001. Radiopédiatrie en pratique courante. Elsevier Masson.

[12] De Bodman, C., Zambelli, P.Y., Dayer, R.O.P., 2017. Scoliose idiopathique de l'adolescent: critères diagnostiques et prise en charge. Revue médicale suisse 13, 422–426.

[13] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

[14] Dhar, S., Dangerfield, P., Dorgan, J., Klenerman, L., 1993. Correlation between bone age and risser's sign in adolescent idiopathic scoliosis. Spine 18, 14–19.

[15] Goldberg, M.S., Poitras, B., Mayo, N.E., Labelle, H., Bourassa, R., Cloutier, R., 1988. Observer variation in assessing spinal curvature and skeletal development in adolescent idiopathic scoliosis. Spine 13, 1371–1377.

[16] Greulich, W.W., Pyle, S.I., 1959. Radiographic atlas of skeletal development of the hand and wrist. Stanford university press.

[17] Hacquebord, J.H., Leopold, S.S., 2012. In brief: The risser classification: a classic tool for the clinician treating adolescent idiopathic scoliosis. Clinical orthopaedics and related research 470, 2335–2338.

[18] 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., et al., 2019. The rsna pediatric bone age machine learning challenge. Radiology 290, 498–503.

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

[20] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

[21] Humeau-Heurtier, A., 2019. Texture feature extraction methods: A survey. IEEE Access 7, 8975–9000.

[22] Hung, A.L.H., Shi, B., Chow, S.K.H., Chau, W.W., Hung, V.W.Y., Wong, R.M.Y., Liu, K.L., Lam, T.P., Ng, B.K.W., Cheng, J.C.Y., 2018. Validation study of the thumb ossification composite index (toci) in idiopathic scoliosis: a stage-to-stage correlation with classic tanner-whitehouse and sanders simplified skeletal maturity systems. The Journal of bone and joint surgery. American volume 100, e88–1.

[23] Kaddioui, H., Duong, L., Joncas, J., Bellefleur, C., Nahle, I., Chémaly, O., Nault, M.L., Parent, S., Grimard, G., Labelle, H., 2020. Convolutional neural networks for automatic risser stage assessment. Radiology: Artificial Intelligence 2, e180063.

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

[25] Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B.A., Alkasab, T.K., Choy, G., Do, S., 2017. Fully automated deep learning system for bone age assessment. Journal of digital imaging 30, 427–441.

[26] Li, D.T., Cui, J.J., DeVries, S., Nicholson, A.D., Li, E., Petit, L.,

Kahan, J.B., Sanders, J.O., Liu, R.W., Cooperman, D.R., et al., 2018. Humeral head ossification predicts peak height velocity timing and percentage of growth remaining in children. Journal of pediatric orthopedics 38, e546.

[27] Liu, B., Liu, Q., Zhu, Z., Zhang, T., Yang, Y., 2019. Msst-resnet: Deep multi-scale spatiotemporal features for robust visual object tracking. Knowledge-Based Systems 164, 235–252.

[28] Liu, L., Fieguth, P., Wang, X., Pietikäinen, M., Hu, D., 2016. Evaluation of lbp and deep texture descriptors with a new robustness benchmark, in: European Conference on Computer Vision, Springer. pp. 69–86.

[29] Lopes, U., Valiati, J.F., 2017. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in biology and medicine 89, 135–143.

[30] Mehri, M., Héroux, P., Gomez-Krämer, P., Mullot, R., 2017. Texture feature benchmarking and evaluation for historical document image analysis. International Journal on Document Analysis and Recognition (IJDAR) 20, 1–35.

[31] Mutasa, S., Chang, P.D., Ruzal-Shapiro, C., Ayyala, R., 2018. Mabal: a novel deep-learning architecture for machine-assisted bone age labeling. Journal of digital imaging 31, 513–519.

[32] Pan, I., Baird, G.L., Mutasa, S., Merck, D., Ruzal-Shapiro, C., Swenson, D.W., Ayyala, R.S., 2020. Rethinking greulich and pyle: A deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. Radiology: Artificial Intelligence 2, e190198.

[33] Rayan, J.C., Reddy, N., Kan, J.H., Zhang, W., Annapragada, A., 2019. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiology: Artificial Intelligence 1, e180015.

[34] Ren, X., Li, T., Yang, X., Wang, S., Ahmad, S., Xiang, L., Stone, S.R., Li, L., Zhan, Y., Shen, D., et al., 2018. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE journal of biomedical and health informatics 23, 2030–2038.

[35] Rodriguez, J.D., Perez, A., Lozano, J.A., 2009. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence 32, 569–575.

[36] Sanders, J.O., Khoury, J.G., Kishan, S., Browne, R.H., Mooney III, J.F., Arnold, K.D., McConnell, S.J., Bauman, J.A., Finegold, D.N., 2008. Predicting scoliosis progression from skeletal maturity: a simplified classification during adolescence. JBJS 90, 540–553.

[37] Shuren, N., Kasser, J.R., Emans, J.B., Rand, F., 1992. Reevaluation of the use of the risser sign in idiopathic scoliosis. Spine 17, 359–361.

[38] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[39] Soffer, S., Ben-Cohen, A., Shimon, O., Amitai, M.M., Greenspan, H., Klang, E., 2019. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 290, 590–606.

[40] Souza, D., Oliveira, M.M., 2018. End-to-end bone age assessment with residual learning, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. pp. 197–203.

[41] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

[42] Thian, Y.L., Li, Y., Jagmohan, P., Sia, D., Chan, V.E.Y., Tan, R.T., 2019. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiology: Artificial Intelligence 1, e180001.

[43] 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, 1–8.

[44] van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., the scikit-image contributors, 2014. scikit-image: image processing in Python. PeerJ 2, e453. URL: `https://doi.org/10.7717/peerj.453`, doi:`10.7717/peerj.453`.

[45] Weinstein, S.L., Dolan, L.A., Cheng, J.C., Danielsson, A., Morcuende, J.A., 2008. Adolescent idiopathic scoliosis. The lancet 371, 1527–1537.

[46] Wybier, M., Bossard, P., 2013. Musculoskeletal imaging in progress: the eos imaging system. Joint Bone Spine 80, 238–243.

[47] Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. Insights into imaging 9, 611–629.

[48] Zhao, C., Han, J., Jia, Y., Fan, L., Gou, F., 2018. Versatile framework for medical image processing and analysis with application to automatic bone age assessment. Journal of Electrical and Computer Engineering 2018.

[49] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13001–13008.