Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Convolutional Neural Networks for Automatic Risser Stage Assessment

Houda Kaddioui, MD, MSc • Luc Duong, PhD • Julie Joncas, RN, BSc • Christian Bellefleur, MSc • Imad Nahle, MD • Olivier Chémaly, MD • Marie-Lyne Nault, MD, PhD • Stefan Parent, MD, PhD • Guy Grimard, MD • Hubert Labelle, MD

From the Department of Software and IT Engineering, Ecole de Technologie Supérieure, 1100 rue Notre-Dame Ouest, Montréal, QC, Canada H3C 1K3 (H.K., L.D.); Division of Orthopedics, Sainte-Justine Hospital, Montréal, Canada (J.J., C.B., I.N., O.C., S.P., G.G., H.L.); and Department of Surgery, Université de Montréal, Montréal, Canada (M.L.N., S.P., G.G., H.L.). Received November 20, 2018; revision requested December 21; revision received January 20, 2020; accepted January 27. **Address correspondence to** H.K. (e-mail: *houda.kaddioui@gmail.com*).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(3):e180063 • https://doi.org/10.1148/ryai.2020180063 • Content codes: IN MK

Purpose: To develop an automatic method for the assessment of the Risser stage using deep learning that could be used in the management panel of adolescent idiopathic scoliosis (AIS).

Materials and Methods: In this institutional review board approved–study, a total of 1830 posteroanterior radiographs of patients with AIS (age range, 10–18 years, 70% female) were collected retrospectively and graded manually by six trained readers using the United States Risser staging system. Each radiograph was preprocessed and cropped to include the entire pelvic region. A convolutional neural network was trained to automatically grade conventional radiographs according to the Risser classification. The network was then validated by comparing its accuracy against the interobserver variability of six trained graders from the authors' institution using the Fleiss κ statistical measure.

Results: Overall agreement between the six observers was fair, with a κ coefficient of 0.65 for the experienced graders and agreement of 74.5%. The automatic grading method obtained a κ coefficient of 0.72, which is a substantial agreement with the ground truth, and an overall accuracy of 78.0%.

Conclusion: The high accuracy of the model presented here compared with human readers suggests that this work may provide a new method for standardization of Risser grading. The model could assist physicians with the task, as well as provide additional insights in the assessment of bone maturity based on radiographs.

The Risser grade is widely used to assess bone maturity and the progressive potential of adolescent idiopathic scoliosis (AIS) (1–3). Since Risser introduced the comprehensive method for observing the ossification of the iliac crest from conventional radiographs (4), two main classification systems emerged: the United States classification (used in this study) and the French classification. The United States classification divides the ossification progression into six stages, where stage 0 is a nonossified iliac crest and 5 is a total fusion of the bones (Fig 1b). The assessment of bone maturity in the context of AIS is significant because patients with less mature bone are at increased risk of curve progression.

Even with a clear clinical definition, interpretation of plain radiographs is challenging due to: *(a)* different image qualities between acquisitions, *(b)* variability in radiographic systems, *(c)* severe deformities where the strict frontal condition is no longer respected, and *(d)* the continual cycle of bone ossification. Interobserver variability in the assessment of the Risser stage exists due to the rotated nature of the pelvis in AIS and subjective visual grading. Previous studies have established a lack of consensus concerning this variability. Goldberg et al (6) demonstrated a κ of 0.80, and Dhar et al (7) showed an agreement of 89.2%. In contrast, more recent studies showed a 50% agreement for all stages combined, while Shuren et al (8) showed moderate agreement between orthopedic surgeons and radiologists that can be as high as three Risser stages between the raters. Risser grading using an automated tool may be helpful in uncertain cases. We propose such a computerized tool using convolutional neural networks (CNNs) (9) to classify Risser stages from radiographs.

CNNs are a subtype of deep learning. The architecture of CNNs is inspired by the human hierarchical learning process and visual recognition pathways where information is sequentially processed with increased complexity (9). A comprehensive introduction of CNN models is available in Soffer et al (10). Among popular models, AlexNet, VGG, and U-Net are the most commonly used networks for image detection. AlexNet consists of five convolutional layers, and it was designed from 1.2 million natural images. VGG16/VGG19 is a deeper network, with 16 and 19 layers, respectively. U-Net is characterized by a contracting path and an expansive path that substitutes the fully connected layers. For bone detection, Inception-ResNet was recently introduced for fracture identification on wrist radiographs (11).

To the best of our knowledge, deep learning has not yet been applied for assessing Risser stage on radiographs. Hence, the goal of this study was to propose a new deep

Abbreviations

AIS = adolescent idiopathic scoliosis, CI = confidence interval, CNN = convolutional neural network

Summary

A deep learning network was developed to determine Risser stage from pelvic radiographs in adolescent patients; the network had similar accuracy to expert readers and thus could be implemented to aid physicians by providing a second opinion on staging.

Key Points

- The developed deep learning method to automate Risser stage assessment reached 78.0% accuracy, which was comparable to 74.5% agreement between expert readers.
- Risser stage assessment using deep learning models is promising for the evaluation of skeletal maturity in patients with adolescent idiopathic scoliosis and could reduce the propagation of error biases within clinical files.

learning technique for the automatic assessment of Risser stage. We validated the performance of our method against observers by evaluating the interobserver variability and found that the model performed similarly to experts. Automatic Risser grading using deep learning models could be developed as a tool to assist physicians and serve as a second opinion in institutions that either lack specialists or that have too few specialists to provide a second opinion on every case that might warrant one.

Materials and Methods

Study Design

Institutional review board approval and informed consent information were obtained for this retrospective study. A total of 1830 posteroanterior EOS and standard digital radiographs were collected between 1999 and 2017 from the scoliosis clinic from 1830 patients (age range, 10–18 years, 70% female) with confirmed AIS. The images included the cervical vertebrae and the femoral head (98.0%) or were full-body images (2.0%). The reference for Risser grading in this study was the United States Risser stage. The information was collected from the patients' scoliosis clinic records. The maximum Risser stage over the two iliac crests was set as the final label and was used as the ground truth by a trained technician and validated by an independent expert. In case of disagreement, a discussion about the case resulted in an agreed upon grade. There was no situation that necessitated the involvement of a third expert.

Radiograph Acquisition

The EOS images were acquired using EOS system II and III (EOS Imaging, Paris, France), and the conventional images were acquired using Fuji system FCR 7501 (Fujifilm, Tokyo, Japan).

Model Development

The Titan Xp graphics processing unit used for this research was donated by Nvidia (Santa Clara, Calif). The authors had full control over the data.

Interobserver and Evaluation of Agreement

To evaluate the interobserver variability, six graders were recruited. The group was composed of four orthopedic surgeons, one orthopedic fellow, and one research nurse. The graders were organized in two groups: senior experts (H.L. and S.P.) (more than 20 years of experience) and new experts (M.L.N., O.C., J.J., I.N.) (fewer than 10 years of experience). The overall agreement was computed first, followed by the agreement within groups. All graders assess the Risser stage on a regular basis. A balanced sample of 200 shuffled radiographs was provided to each grader (Fig 1). The readers were blinded to the sex, age, and demographic information about the patients; the recorded Risser stage; and the assessment of their peers. Each grader independently classified all 200 images, and the stages were based on the United States Risser classification.

Automatic Risser Grading

Training deep learning networks requires a large number of annotated images. Because the number of radiographs was limited in our dataset, we applied transfer learning using the VGG16 network (12). This approach consisted of reusing a CNN trained on a large dataset (eg, natural images) and adjusting its parameter to better fit our dataset. Transfer learning has been proven effective in practice for medical imaging (13,14).

Preprocessing of all radiographs was performed. The images were first cropped along the smallest edge and then resized to keep the aspect size ratio while including the entire pelvis, which resulted in 224 × 224-pixel images. A median filter was applied afterward to remove the salt-and-pepper noise. The dataset was then split into training and validation sets at an 80%:20% ratio. A third subset was left as a second testing set used for the validation of the accuracy against the experts as mentioned above. When the images were input to the network, convolution filters of a fixed size created a feature map by sliding over the entire image following a fixed stride. Convolution layers were followed by rectified linear unit layer to add nonlinearity and to improve the network's generalization (15). Afterward, a pooling layer was used to sample over the output of the previous layer, only keeping the most valuable information by retaining the maximum value in a given N × N window. The final layers of the network were specifically developed to train on the Risser grading task. This new set was randomly initialized and connected to the body of the original network. The fully connected layers resulted in a computed output of size 1 × 1 × C, where C is the number of different Risser stages (Fig 2).

The model parameters were initialized to pretrained weights optimized for the ImageNet dataset (16). To keep the parameters of the trained model, the first step was to freeze the superficial layers and only train the new layers over multiple iterations. This avoids propagation of the gradient over the entire network and prevents losing the discriminating parameters for the kernels, while allowing the filters to learn new parameters. After 30 iterations, the layers were "unfrozen," and training continued until sufficient accuracy was obtained, with a learning rate of 1×10^{-5}. The accuracy is defined as the number of correctly classified

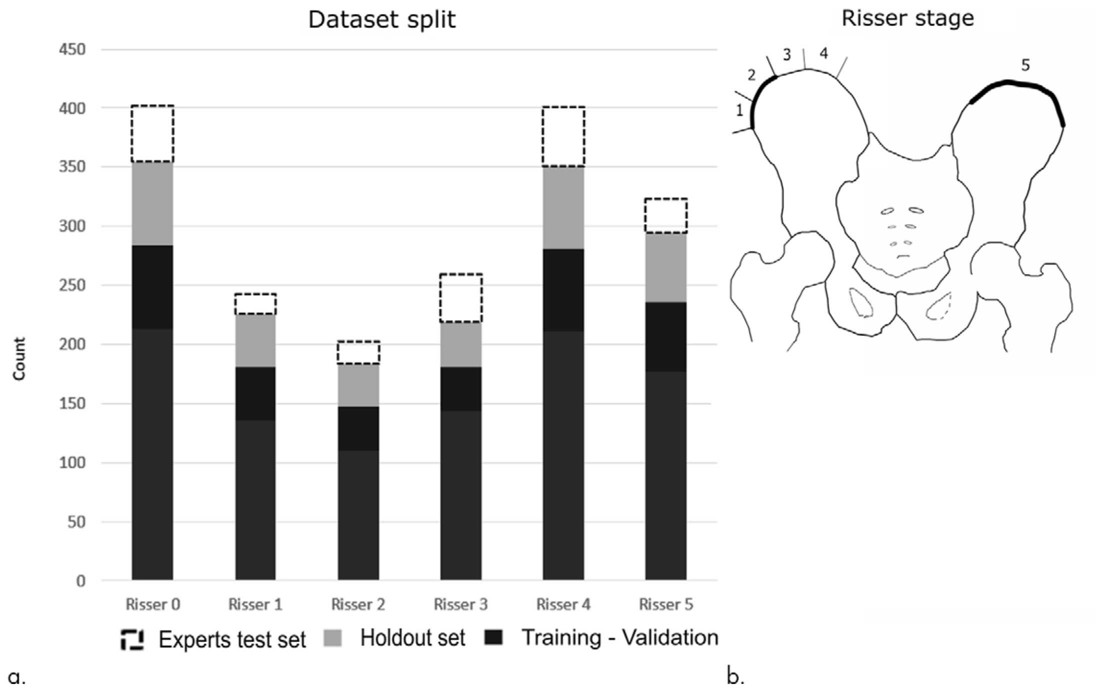


Figure 1: **(a)** Distribution of the Risser grade in the radiographic database. The expert test set consisted of 200 images to assess rater variability. The holdout set was used to test the model. The training-validation set was used to train and validate the model. **(b)** Visual illustration of iliac crest progression and corresponding Risser stages.

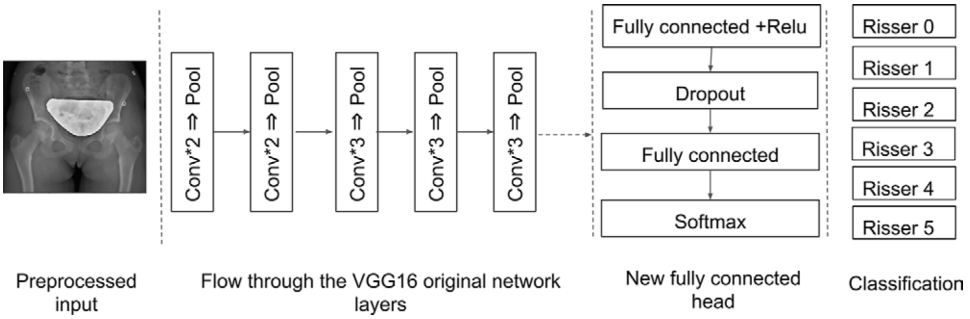


Figure 2: Feature extraction and classification workflow with convolutional neural networks. The output of the proposed method is the Risser grade (0–5). Conv = convolution, Relu = rectified linear unit.

images over the total number of images. Stochastic gradient descent was used for optimization to correct the predictions and guide the network toward accurate weights. After determining the final parameters, the training was performed for 10 folds to control for the effect of chance. To evaluate the network, we compared its accuracy with the agreement interval of the different grader groups. The software was developed in Python (version 2.7; Python Software Foundation, Wilmington, Del) using the Keras library *(https://keras.io/)* with the TensorFlow library *(https://www.tensorflow.org/)* for deep learning (17). The training phase took 8 hours on a professional workstation with a high-end graphics processing unit.

Statistical Analysis

To determine the interreader variability of the six graders, Fleiss κ was calculated. The κ coefficient measures the agreement between graders while accounting for the effect of chance. If the graders are in complete agreement, κ = 1, while if there is no agreement, κ = 0. When the analyzed group had more than two graders, the Fleiss variation was used (18). The results were compared with the criteria of Landis and Koch: Lower than zero corresponds to less than chance agreement, 0.01–0.20 to slight agreement, 0.21–0.40 to fair agreement, 0.41–0.60 to moderate agreement, 0.61–0.80 to substantial agreement, and 0.81–0.99 to almost perfect agreement (19). Groupwise and pairwise percentage of agreement were computed for a better interpretation of the observers' agreement. κ statistics and percentage of agreement were computed using R language (version 3.4.1; R Foundation for Statistical Computing, Vienna, Austria).

Results

Interobserver Agreement

To establish a baseline for the grading ability of our deep learning network, we first determined the interobserver agreement of Risser grading from trained experts. A total of six readers classified the images and determined the Risser grade. The overall agreement between the observers was fair with a value of κ = 0.62 (95% confidence interval [CI]: 0.46, 0.78). Senior experts (observers 5 and 6) had a κ coefficient of 0.65 (95% CI: 0.48,

Table 1: Pairwise κ Value of the Observers, the Ground Truth, and the Proposed Automatic Grading Method

Observer	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6	AGM	GT
1	1.00	0.62	0.53	0.55	0.71	0.68	0.64	0.63
2		1.00	0.50	0.50	0.62	0.55	0.54	0.57
3			1.00	0.59	0.57	0.65	0.58	0.52
4				1.00	0.53	0.58	0.57	0.49
5					1.00	0.65	0.69	0.60
6						1.00	0.60	0.52
AGM							1.00	0.72
GT								1.00

Note.—AGM = automatic grading method, GT = ground truth.

Table 2: Pairwise Percentage of Agreement for the Observers, the Ground Truth, and the Proposed Automatic Grading Method

Observer	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6	AGM	GT
1	100.0	71.0	68.5	64.5	81.0	75.5	72.0	71.0
2		100.0	62.5	61.0	71.0	65.5	65.0	66.0
3			100.0	68.0	68.5	74.5	68.5	62.5
4				100.0	63.5	67.5	66.4	59.0
5					100.0	74.5	76.0	69.0
6						100.0	67.5	62.0
AGM							100.0	78.0
GT								100.0

Note.—AGM = automatic grading method, GT = ground truth.

0.82) and had a total consensus on the Risser stage on 74.5% of the images. New experts (observers 1–4) had a κ coefficient of 0.58 (95% CI: 0.40, 0.76) and had a total consensus on 41.5% of the images. The pairwise κ coefficients and percentage of agreement for all observers are presented in Tables 1 and 2. The pairwise agreement ranged from fair (0.21–0.40) to moderate (0.41–0.60). The percentage of agreement of the experts with the ground truth (true Risser stage) was calculated and is reported in Figure 3 as the performances of each expert and the group performance over each class. The best performance of the group was obtained when the Risser stage was 0. There was no noticeable difference between the senior experts' and the new experts' performances and thus, no visible effect of time in the individual performance. Within stages, the senior experts were more consistent than the new experts.

Confusion matrices were used to map the classification results of the developed network and the experts' gradings. Analyzing the confusion matrix revealed high performances on Risser stage 0, 1, and 5; stages 3 and 4 had the most variability.

Automatic Risser Grading Method

Next, our model was tested on the same dataset given to the graders group. The automatic grading method showed a substantial agreement with the ground truth (κ = 0.72; 95% CI: 0.59, 0.85) and an accuracy of 78.0% (95% CI: 75.7%, 80.3%). Analysis of the network's output showed a misclassification limited to two stages (Fig 4a), while for graders the variability could be three or more stages (Fig 4b). Moreover, the misclassified images correspond to the most controversial images with the least agreement between the observers (Fig 5a). Finally, an analysis of the activated regions using the Keras-vis library (20) revealed the model's attention on the most important anatomic features (Fig 5b). The computing time at inference was less than 1 second per image. Together, the deep learning model performed in a comparable manner to the six expert readers.

Discussion

The Risser stage is a widely used indicator of skeletal maturity and progression potential of AIS. Although Risser staging is comprehensive and easy to implement, several authors have previously raised concerns regarding its efficacy and reliability. Studies suggest that the Risser system is subject to interobserver variability, does not reflect the velocity of the curve progression, and is not sensitive to rapid acceleration phases (2). Sanders et al introduced a new classification of bone maturity based on wrist radiographs (21). A study comparing the Risser and Sanders classifications showed a higher κ coefficient for the latter (22). Nault et al (23) also demonstrated the value of including more ossification in evaluating bone maturity, proposing a new Risser classification that includes the triradiate cartilage. Similarly, Troy et al proposed a revised classifica-

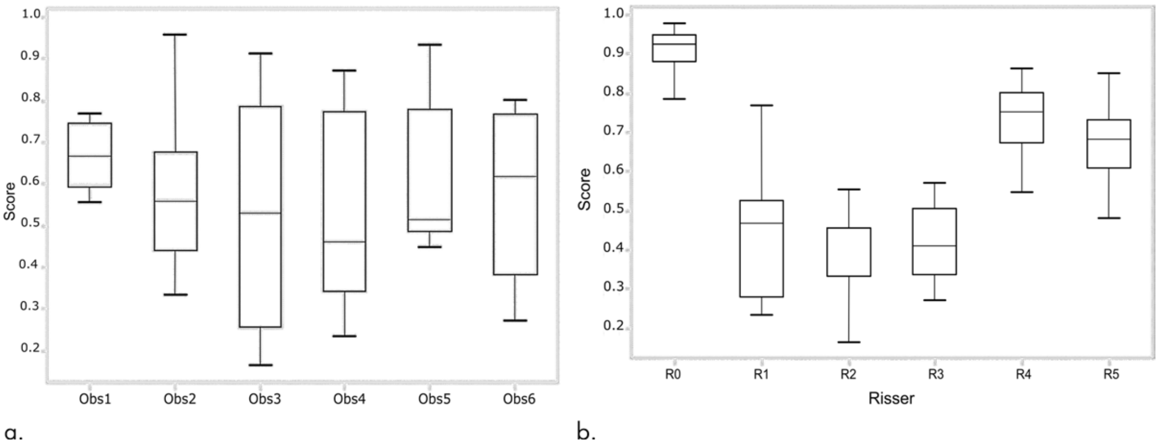
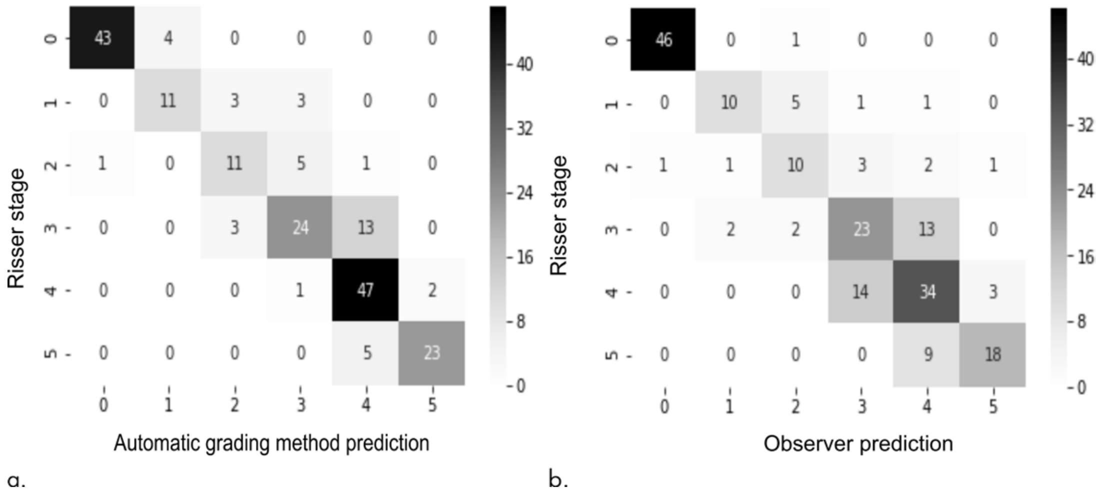


Figure 3: **(a)** Performance of each observer (Obs) in grading the test set. **(b)** Performance of all the observers for each Risser stage (R0–R5). The score represents the fraction of answers in agreement with the ground truth. The lower and upper quartiles are also shown.



	0	1	2	3	4	5
0	43	4	0	0	0	0
1	0	11	3	3	0	0
2	1	0	11	5	1	0
3	0	0	3	24	13	0
4	0	0	0	1	47	2
5	0	0	0	0	5	23

	0	1	2	3	4	5
0	46	0	1	0	0	0
1	0	10	5	1	1	0
2	1	1	10	3	2	1
3	0	2	2	23	13	0
4	0	0	0	14	34	3
5	0	0	0	0	9	18

Figure 4: **(a)** Confusion matrix for the automatic grading method. **(b)** Confusion matrix for one of the observers. The rows of the matrix show the values indicated by the observer, while the columns show the ground truth. The values on the diagonal of the matrix illustrate the number of samples correctly classified by Risser grade. The values above and below each value of the diagonal show misclassified samples.

tion with eight Risser stages, combining the United States and French classifications with the triradiate cartilage ossification. Their interobserver evaluation produced insufficient agreement (24). All these studies show a common concern regarding the grading variability among experts.

Previous published studies show a κ value of 0.31 to 0.80 (6,8). This broad range underlines the need for normalized databases, intraobserver and interobserver studies, and the development of automated grading systems. Our readers had fair to moderate agreement, matching the highest agreement values in the literature. However, the interpretation of κ values must consider two factors: first, the null hypothesis in a medical context should not be set as κ = 0; rather, a minimum acceptable agreement should be decided on. To our knowledge, no such value has been defined, hence the need to obtain the best possible agreement. The second factor is the effect of variability on the therapeutic decision; one study showed that the variability in assessing the Risser stage leads to several issues (3). In the clinical context, variability leads to missing classes and radiation exposures; when added to the impact of the treatment, this can be overwhelming for adolescents (5,25). Getting a second opinion might reduce this variability and thus reduce the propagation of an error bias within the patient's files. However, a second opinion is usually not easily available. Because our network had been trained on an agreement of two experts and validated on a group of six other graders, its classification would come as a second opinion. Moreover, some factors including time or the physical state or workload of a human expert can reduce the accuracy of the classification, whereas a network is invariant and independent of these factors.

Skeletal maturity evaluation is an integral part of pediatric radiology and orthopedics. However, manual grading of a large number of radiographs is time-consuming, and obtaining a second opinion to reduce variability is unfit for clinical settings. Deep learning has recently been introduced for radiographic assessment of skeletal maturity on carpograms using a five-layer CNN (26). When assessing the key regions, the network suggested that some carpal regions accounted for by clinicians might not be relevant, while some new regions should be considered. The recent deep learning bone age assessment models

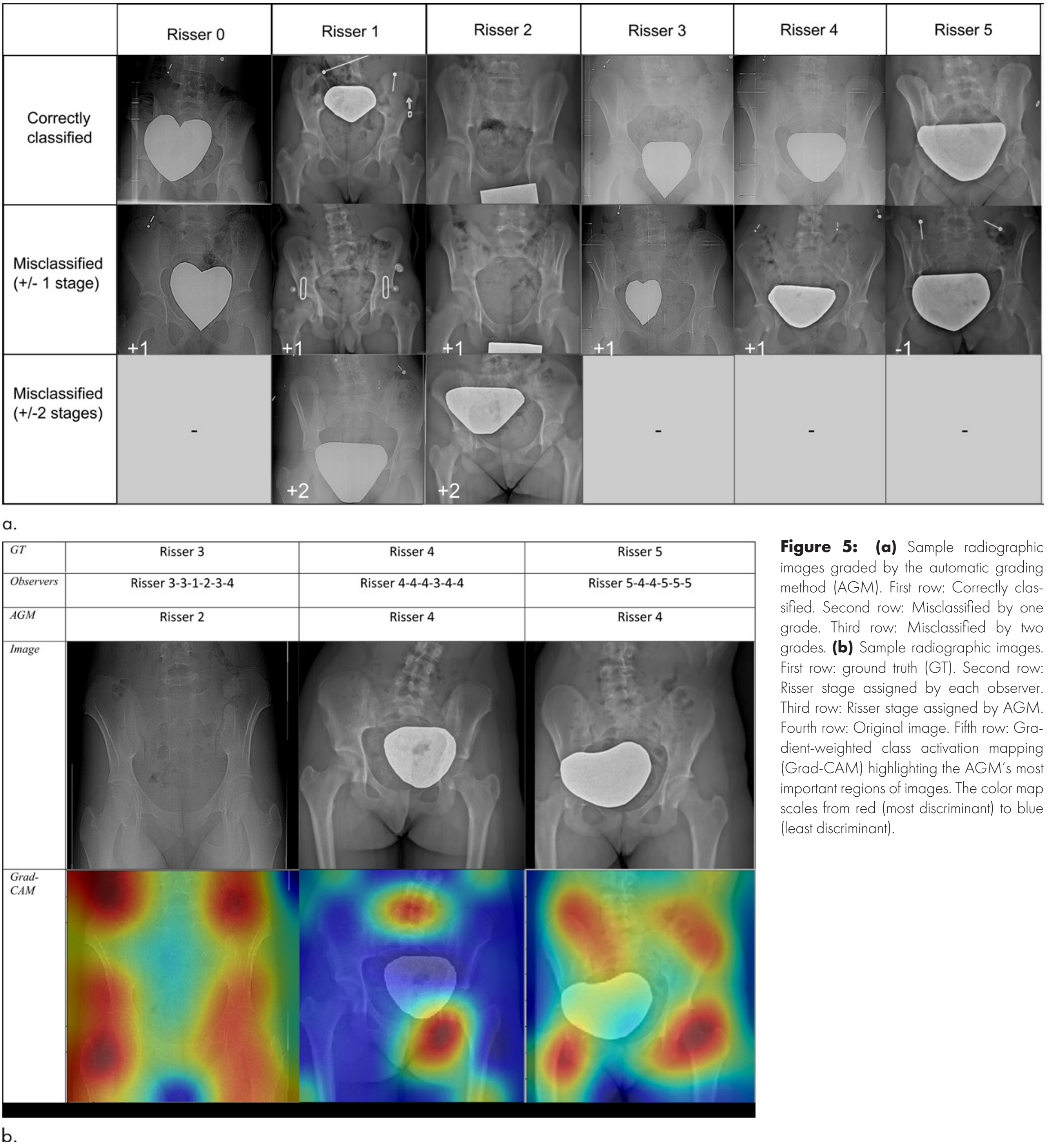


Figure 5: **(a)** Sample radiographic images graded by the automatic grading method (AGM). First row: Correctly classified. Second row: Misclassified by one grade. Third row: Misclassified by two grades. **(b)** Sample radiographic images. First row: ground truth (GT). Second row: Risser stage assigned by each observer. Third row: Risser stage assigned by AGM. Fourth row: Original image. Fifth row: Gradient-weighted class activation mapping (Grad-CAM) highlighting the AGM's most important regions of images. The color map scales from red (most discriminant) to blue (least discriminant).

not only yield satisfactory performance scores of 61% to 79%, but they also give interesting insights that could be further investigated (26–28). Similarly, our results illustrate that CNNs can be used to assign the Risser grade with satisfying accuracy. An automatic method is appealing because computerized approaches are highly predictive and give consistent output for the same input without internal variability. Furthermore, the result is given within seconds, and the classification errors are not aberrant, as shown in the confusion matrix. Finally, the network was trained to learn the most specific and invariant features, making it robust against different image variations, rotations, and contrasts, thereby overcoming the limitations of the Risser grading system. Thus, such a tool has the potential to be implemented to assist physicians in the assessment task.

Although different authors question the reliability of the Risser stage, the results of this study are promising and show the potential for a more accurate bone maturity assessment on radiographs. However, there were some limitations to this work. The ground truth was used based on the agreement of two observers, meaning that the network could be less accurate on a noisier dataset. Our work can be improved by collecting more radiographs and having additional graders agree on the final label. Finally, because the network was trained solely on radiographs of patients with AIS, an improvement to the methodology could be achieved by including more patients from different clinics. Additional reliability gain could be reached by diversifying the dataset.

We developed an automatic Risser grading method using a CNN, a deep learning approach. In addition, we evaluated interobserver variability at our institution. Our automatic method was able to perform within the known interobserver variability without internal variability. These results pave the way for more investigation on the feasibility of integrating automatic radiographic methods in clinical settings and its usefulness for the management of AIS.

Author contributions: Guarantors of integrity of entire study, H.K., L.D., J.J., I.N., S.P., H.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, H.K., L.D., S.P.; clinical studies, H.K., L.D., I.N., O.C., G.G.; experimental studies, H.K., L.D., M.L.N.; statistical analysis, H.K., L.D.; and manuscript editing, H.K., L.D., I.N., M.L.N., S.P., G.G., H.L.

Disclosures of Conflicts of Interest: **H.K.** disclosed no relevant relationships. **L.D.** disclosed no relevant relationships. **J.J.** disclosed no relevant relationships. **C.B.** disclosed no relevant relationships. **I.N.** disclosed no relevant relationships. **O.C.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author received payment from Ergoresearch for presentations on knee osteoarthritis to family physicians; author received travel accommodations from Johnson & Johnson for meeting on early onset scoliosis. Other relationships: disclosed no relevant relationships. **M.L.N.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives grant for HSC from Stryker, Smith & Nephew, Johnson & Johnson, and Wright Medical; patent for US9547897B2 (but has received no money). Other relationships: disclosed no relevant relationships. **S.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is paid consultant for K2M, EOS-Imaging, Medtronic, and DePuy Synthes Spine; institution receives grants from DePuy Synthes Spine, Canadian Institutes of Health, Pediatric Orthopaedic Society of North America, Scoliosis Research Society, Medtronic, EOS Imaging, Canadian Foundation for Innovation, Setting Scoliosis Straight Foundation, Natural Sciences and Engineering Council of Canada, Fonds de recherche Québec-Santé, Orthopaedic Research and Education Foundation; author receives royalties from EOS Imaging; stock in Spinologics; Academic Research Chair in spine deformities of the CHU Sainte-Justine (DePuy). Other relationships: disclosed no relevant relationships. **G.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: board member of Quebec Orthopedic Association; expert testimony (medical expertise) for Canadian Protective Medical Association; stock in Emovi. Other relationships: disclosed no relevant relationships. **H.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives grant from Canadian Institutes of Health Research; author has stock in Spinologics. Other relationships: disclosed no relevant relationships.

References

1. Izumi Y. The accuracy of Risser staging. Spine (Phila Pa 1976) 1995;20(17):1868–1871.
2. Reem J, Carney J, Stanley M, Cassidy J. Risser sign inter-rater and intra-rater agreement: is the Risser sign reliable? Skeletal Radiol 2009;38(4):371–375.
3. Hammond KE, Dierckman BD, Burnworth L, Meehan PL, Oswald TS. Inter-observer and intra-observer reliability of the Risser sign in a metropolitan scoliosis screening program. J Pediatr Orthop 2011;31(8):e80–e84.
4. Hacquebord JH, Leopold SS. In brief: The Risser classification: a classic tool for the clinician treating adolescent idiopathic scoliosis. Clin Orthop Relat Res 2012;470(8):2335–2338.
5. Weinstein SL, Dolan LA, Cheng JC, Danielsson A, Morcuende JA. Adolescent idiopathic scoliosis. Lancet 2008;371(9623):1527–1537.
6. Goldberg MS, Poitras B, Mayo NE, Labelle H, Bourassa R, Cloutier R. Observer variation in assessing spinal curvature and skeletal development in adolescent idiopathic scoliosis. Spine (Phila Pa 1976) 1988;13(12):1371–1377.
7. Dhar S, Dangerfield PH, Dorgan JC, Klenerman L. Correlation between bone age and Risser's sign in adolescent idiopathic scoliosis. Spine (Phila Pa 1976) 1993;18(1):14–19.
8. Shuren N, Kasser JR, Emans JB, Rand F. Reevaluation of the use of the Risser sign in idiopathic scoliosis. Spine (Phila Pa 1976) 1992;17(3):359–361.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
10. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019;290(3):590–606.
11. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019;1(1):e180001.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv:1409.1556 [preprint] https://arxiv.org/abs/1409.1556. Posted September 4, 2014. Accessed June 25, 2018.
13. Abdolmanafi A, Duong L, Dahdah N, Cheriet F. Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography. Biomed Opt Express 2017;8(2):1203–1220.
14. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
15. Novak R, Bahri Y, Abolafia DA, Pennington J, Sohl-Dickstein J. Sensitivity and generalization in neural networks: an empirical study. arXiv:1802.08760 [preprint] https://arxiv.org/abs/1802.08760. Posted February 23, 2018. Accessed March 25, 2019.
16. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009; 248–255.
17. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Berkeley, Calif: USENIX Association, 2016; 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.
18. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33(3):613–619.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
20. Kotikalapudi R, contributors. keras-vis. GitHub. https://github.com/raghakot/keras-vis. Published 2017. Accessed March 25, 2019.
21. Sanders JO, Khoury JG, Kishan S, et al. Predicting scoliosis progression from skeletal maturity: a simplified classification during adolescence. J Bone Joint Surg Am 2008;90(3): 540–553.
22. Minkara A, Bainton N, Tanaka M, et al. High risk of mismatch between Sanders and Risser staging in adolescent idiopathic scoliosis: are we guiding treatment using the wrong classification? J Pediatr Orthop 2018 Jan 22 [Epub ahead of print].
23. Nault ML, Parent S, Phan P, Roy-Beaudry M, Labelle H, Rivard M. A modified Risser grading system predicts the curve acceleration phase of female adolescent idiopathic scoliosis. J Bone Joint Surg Am 2010;92(5):1073–1081.
24. Troy MJ, Miller PE, Price N, et al. The "Risser+" grade: a new grading system to classify skeletal maturity in idiopathic scoliosis. Eur Spine J 2019;28(3):559–566.
25. Goldberg MS, Mayo NE, Poitras B, Scott S, Hanley J. The Ste-Justine Adolescent Idiopathic Scoliosis Cohort Study. part II: perception of health, self and body image, and participation in physical activities. Spine (Phila Pa 1976) 1994;19(14):1562–1572.
26. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in x-ray images. Med Image Anal 2017;36:41–51.
27. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30(4):427–441.
28. Torres F, Bravo MA, Salinas E, Triana G, Arbeláez P. Bone age detection via carpogram analysis using convolutional neural networks. In: Romero E, Lepore N, Brieva J, García JD, eds. Proceedings of SPIE: 13th International Conference on Medical Information Processing and Analysis. Vol 10572. Bellingham, Wash: International Society for Optics and Photonics, 2017; 1057217.