Convolutional Neural Networks for Automatic Risser Stage Assessment

Kaddioui Houda., M.D.Msc[1], Duong Luc., Ph.D[1], Joncas Julie[2]., Bellefleur Christian[2]., Nahle Imad, M.D[2]., Chemaly Olivier. M.D[2]., Nault Marie-Lyne. M.D,Ph.D[3], Parent Stefan. M.D, Ph.D. [2,3], Grimard Guy, M.D[2,3]., Labelle Hubert, M.D[2,3].

[1]Dept. of software and IT engineering, Ecole de Technologie Supérieure, Montréal, Quebec, Canada

[2]Division of Orthopedics, Sainte-Justine Hospital, 3175 Côte-Sainte-Catherine, Montréal, QC H3T 1C5, Canada

[3]Dept. of surgery, Université de Montréal

Corresponding author: Houda Kaddioui, Dept. of software and IT engineering, Ecole de Technologie Supérieure, Montréal, Quebec, Canada, houda.kaddioui@gmail.com , +1 4384997855

Article type: Original research

Abbreviations: AIS= adolescent idiopathic scoliosis; CI = confidence interval

Summary Statement: A deep learning network was developed to determine Risser stage on adolescent pelvic radiographs. The network had similar accuracy to expert readers, and thus could be implemented to aid physicians to provide 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 AIS and could reduce the propagation of error biases within clinical files.

Authors' accepted manuscript. Article published in *Radiology: Artificial Intelligence* on May 27 2020.
 The final published version is available at https://doi.org/10.1148/ryai.2020180063

Abstract

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

Methods: In this institutional review board approved study, a total of 1830 posteroanterior radiographs (ages 10-18, 70% female) of AIS patients were collected retrospectively and graded manually by six trained readers using the United States Risser definition. Each radiograph was pre-processed and cropped to include the entire pelvic region. A convolutional neural network was trained to automatically grade conventional radiographs according to the Risser classification method. The network was then validated by comparing its accuracy against the inter-observer variability of six trained graders from our institution using Fleiss Kappa.

Results: Overall agreement between the six observers was fair, with a kappa coefficient of 0.60 for the experienced graders and an agreement of 74.5%. The automatic grading method obtained a kappa 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 our model compared to 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 bone maturity from radiographs.

Authors' accepted manuscript. Article published in *Radiology: Artificial Intelligence* on May 27 2020

Introduction

The Risser grade is widely used to assess bone maturity and the progressive potential of Adolescent Idiopathic Scoliosis (AIS) (1,2,3). 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 (used in this study) and the French classifications. The United States classification divides the ossification progression into six stages, where stage 0 is a non-ossified iliac crest and 5 is a total fusion of the bones (Figure 1b). The assessment of bone maturity in the context of AIS is significant since patients with an less mature bones have a greater 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. Inter-observer 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 kappa of 0.80 and Dhar, et al (7) showed an agreement of 89.2%. In contrast, more recent studies showed a 50% agreement all stages combined, while Shuren, et al (8), showed moderate agreement between orthopedic surgeons and radiologists that can go up to three stages between the raters. Risser grading using an automated tool may provide assistance in uncertain cases. We propose such a computerized tool using convolutional neural networks (9) to classify Risser stages from radiographs.

Convolutional neural networks (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

Authors' accepted manuscript. Article published in *Radiology: Artificial Intelligence* on May 27 2020
 The final published version is available at https://doi.org/10.1148/ryai.2020180063

introduction of CNN models is available in (10). Among popular models, AlexNet, VGG and U-Net are the most commonly used network 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 network is characterized by a contracting path and an expansive path that substitutes the fully connected layers. For bone detection, Inception-ResNet has been recently introduced for fracture identification on wrist radiographs (11).

To the best of our knowledge, deep learning has not yet been applied for the assessment of the Risser stage on radiographs. Hence the goal of this study is to propose a new deep learning technique for the automatic assessment of the Risser stage. We validated the performance of our method against observers by evaluating the inter-observer 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 with lack of specialists.

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 1999-2017 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 patient's 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 needed a third expert's involvement.

Radiograph Acquisition

The EOS images were acquired using EOS system II and III (EOS Imaging, Paris, France) and the conventional image 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 Corporation. (Santa Clara, USA). The authors had full control over the data.

Inter-observer and Evaluation of Agreement

To evaluate the inter-observer 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 (more than twenty years of experience) and new experts (less than ten years of experience). The overall agreement was first computed, 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 (Figure 1). The readers were blinded about the sex, age, demographic information of 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. Since 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 (e.g. 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 afterwards to remove the salt-and-pepper noise. The dataset was then split into training and validation set 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 non-linearity and to improve the network's generalization (15). Afterwards, 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 Risser stages (Figure 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 a 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 1e $^{-5}$. The accuracy is defined as the number of correctly classified 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 2.7 using the Keras library with the Tensorflow library for deep learning(17).The training phase took eight hours on a professional workstation with high-end graphics processing unit.

Statistical Analysis

To determine the inter-reader variability of the six graders, Fleiss Kappa was calculated. Kappa coefficients (κ) 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 to Landis and Koch's agreement scale: lower than zero corresponds to less than chance agreement, 0.01–0.20 slight agreement, 0.21– 0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–0.99 almost perfect agreement (19). Groupwise and pairwise percentage of agreement were computed for a better interpretation of the observers' agreement. Kappa statistics and percentage of agreement were computed using R language version 3.4.1.

Results

Inter-observer agreement

In order to have a baseline for the grading ability of our deep learning network, we first determined the inter-observer agreement of Risser grading from trained experts. A total of six readers classified the images an determined the Risser grade. The overall agreement between the observers was fair with a value of k = 0.62 (CI: 0.46 – 0.78). Senior experts (Obs. 5 and Obs. 6) had a kappa coefficient of 0.65 (CI: 0.48 – 0.82) and had a total consensus on the Risser stage on 74.5% of the images. New experts (Obs. 1-4) had a kappa coefficient of 0.58 (CI: 0.40 – 0.76) and had a total consensus on 41.5% of the images. The pairwise kappa coefficients and percentage of agreement for all observers are presented in Table 1 and Table 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, thus no visible effect of time in the individual performance. The senior experts were more consistent within stages 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 while stages 3 and 4 had the most variability.

<u>Automatic Risser Grading Method</u>

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 (k=0.72; CI: 0.59 – 0.85), and an accuracy of 78.0% (CI: 75.7% – 80.3%). Analysis of the network's output showed a misclassification limited to 2 stages (Figure 4a), while the graders could have a variability of 3 or more stages (Figure 4b). Moreover, the misclassified images correspond to the most controversial images with the less agreement between the observers (Figure 5a). Finally, an analysis of the activated regions using the Keras-vis library (20) revealed the model's attention on the most important anatomical features (Figure 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 inter-observer 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 kappa coefficient for the latter (22). Following this theory, Nault et al proposed a new Risser classification that includes the triradiate cartilage(23). Similarly, Hresko et al proposed a revised classification with eight Risser stages, combining the United States and French classifications with the triradiate cartilage ossification. Their inter-observer evaluation produced insufficient agreement (24). All these studies show a common concern regarding the grading variability among experts.

Previous literature reports show a kappa value of 0.31 - 0.80 (6,8). This broad range underlines the need for normalized databases, intra and inter-observer studies, and for developing automated grading systems. Our readers had fair to moderate agreement, matching the literature's highest agreement values. However, the interpretation of kappa values must consider two factors: first, the null hypothesis in a medical context should not be set as k=0, but rather, a minimal acceptable agreement should be decided upon. 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: a 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, 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. Since 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, the physical state or work load 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 receiving a second opinion to reduce its variability is unfit for clinical settings. Deep learning has recently been introduced for radiographic assessment of skeletal maturity on carpograms using a five layered convolutional neural network (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 not only yield satisfactory performance scores of 61% - 79%, but they also give interesting insights that could be further investigated (26–28). Similarly, our results illustrate that a CNNs can be used to assign the Risser grade with satisfying accuracy. An automatic method is appealing since 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 thus overcoming the limitations of the Risser grading system. Hence, such a tool has the potential to be implemented in order to assist physicians in the assessment task.

Although different authors question the reliability of the Risser stage, this study is promising and shows the potential for a more accurate bone maturity assessment on radiographs in the future. However, there are 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, since the network was trained solely on AIS patient's radiographs, 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.

Conclusion.

An automatic Risser grading method was developed using a convolutional neural network, a deep learning approach. In addition, we evaluated inter-observer variability at our institution. Our automatic method was able to perform within the known inter-observer 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.

References

1. Izumi Y. The accuracy of Risser staging. Spine. 1995 Sep 1;20(17):1868–71.

2. Reem J, Carney J, Stanley M, Cassidy J. Risser sign inter-rater and intra-rater agreement: is the Risser sign reliable? Skeletal Radiology. 2009 Apr;38(4):371–5.

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. Journal of Pediatric Orthopaedics. 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 Aug;470(8):2335–8.

5. Weinstein SL, Dolan LA, Cheng JC, Danielsson A, Morcuende JA. Adolescent idiopathic scoliosis. The Lancet. 2008 May 3;371(9623):1527–37.

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. 1988 Dec 1;13(12):1371–7.

7. Dhar S, Dangerfield PH, Dorgan JC, Klenerman L. Correlation between bone age and Risser's sign in adolescent idiopathic scoliosis. Spine. 1993 Jan;18(1):14–9.

8. Shuren N, Kasser JR, Emans JB, Rand F. Reevaluation of the use of the Risser sign in idiopathic scoliosis. Spine. 1992 Mar;17(3):359–61.

9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 27;521(7553):436–44.

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 Mar;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. Radiology: Artificial Intelligence. 2019;1(1):e180001.

12. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:14091556

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, BOE. 2017 Feb 1;8(2):1203–20.

14. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017 Dec 1;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:180208760

16. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. 2009.

17. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: OSDI. 2016. p. 265–283.

18. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement. 1973 Oct;33(3):613–9.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.

20. Kotikalapudi R, contributors. keras-vis [Internet]. GitHub; 2017. Available from: https://github.com/raghakot/keras-vis

21. Sabour S. Reliability of the Sanders Classification Versus the Risser Stage; Avoid Misinterpretation. Journal of Pediatric Orthopaedics. 2018 Jan;38(1):e29.

22. Minkara A, Bainton N, Tanaka M, Kung J, DeAllie C, Khaleel A, et al. High Risk of Mismatch Between Sanders and Risser Staging in Adolescent Idiopathic Scoliosis: Are We Guiding Treatment Using the Wrong Classification? Journal of Pediatric Orthopaedics. 2018 Jan;1.

23. Nault M-L, 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: The Journal of Bone and Joint Surgery-American Volume. 2010 May;92(5):1073–1081.

24. Hresko MT, Troy M, Miller P, Price N, Talwalkar V, Zaina F, et al. Risser plus sign: a new grading system to classify skeletal maturity in scoliosis patients.

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. 1994 Jul 15;19(14):1562–72.

26. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Medical Image Analysis. 2017 Feb;36:41–51.

27. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully Automated Deep Learning System for Bone Age Assessment. J Digit Imaging. 2017 Aug 1;30(4):427–41.

28. Torres F, Bravo MA, Salinas E, Triana G, Arbeláez P. Bone age detection via carpogram analysis using convolutional neural networks. In International Society for Optics and Photonics; 2017 [cited 2017 Dec 15]. p. 1057217.

Table 1: Pairwise Kappa value of the observers, the ground truth and the proposed automatic grading method

Observers	*Obs 1*	*Obs 2*	*Obs 3*	*Obs 4*	*Obs 5*	*Obs 6*	*AGM*	*GT*
Obs 1	1.00	0.62	0.53	0.55	0.71	0.68	0.64	0.63
Obs 2		1.00	0.50	0.50	0.62	0.55	0.54	0.57
Obs 3			1.00	0.59	0.57	0.65	0.58	0.52
Obs 4				1.00	0.53	0.58	0.57	0.49
Obs 5					1.00	0.66	0.69	0.60
Obs 6						1.00	0.60	0.52
AGM							1.00	0.72
GT								1.00

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

Table 2:Pairwise percentage of agreement for the observers, the ground truth and the proposed automatic grading method

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

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

Figure 1: **(a)** Distribution of the Risser grade in the radiographic database and visual illustration of Risser stage. The expert test set consisted of 200 images to assess rater's variability. The holdout set was used to test the model. The training- validation set was used to train and validate the model. **(b)**

Figure 2: Feature extraction and classification workflow with convolutional neural networks. The output of the proposed method is the Risser grade (0-5).

Figure 3: **(a)** Performance of each observer (Obs) in grading the test set. **(b)** Performance of all the observer 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.

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 column 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.

Figure 5: **(a)** Sample radiographic images correctly classified by the automatic grading method (top) and misclassified by one grade (2nd row) and two grades (3rd row). **(b)** Sample radiographic images with ground truth (top), observer's assigned Risser stage (2nd row), automatic grading method (AMG) classification (3rd row) and visualization of the model's attention (4th row).

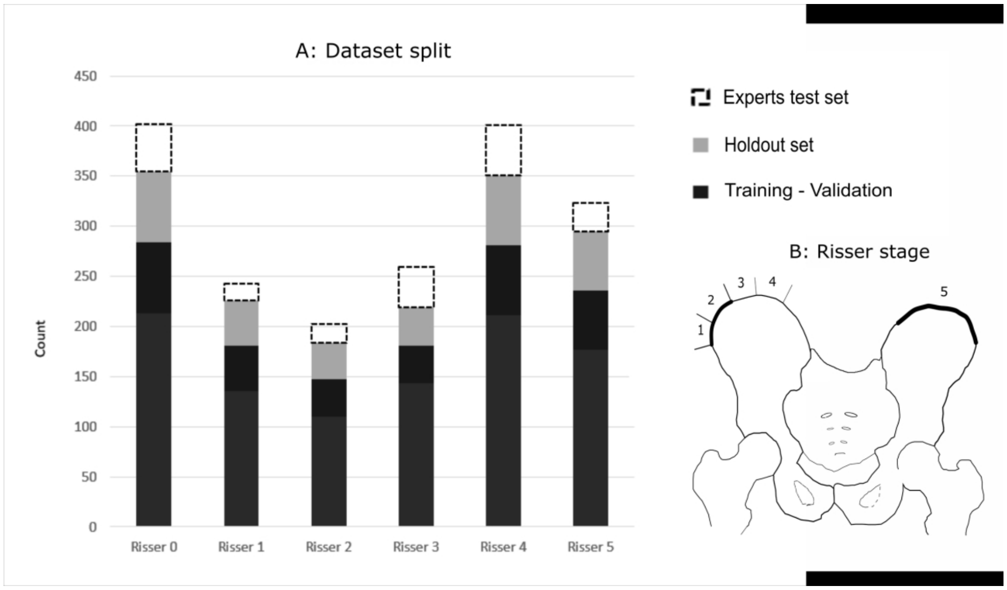

Figure 1: (a) Distribution of the Risser grade in the radiographic database and visual illustration of Risser stage. The expert test set consisted of 200 images to assess rater's variability. The holdout set was used to test the model. The training- validation set was used to train and validate the model.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

190x111mm (300 x 300 DPI)

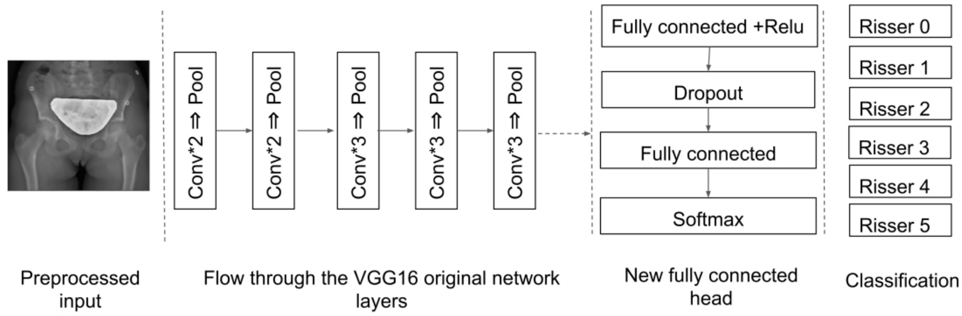

Figure 2: Feature extraction and classification workflow with convolutional neural networks. The output of the proposed method is the Risser grade (0-5)

254x91mm (375 x 375 DPI)

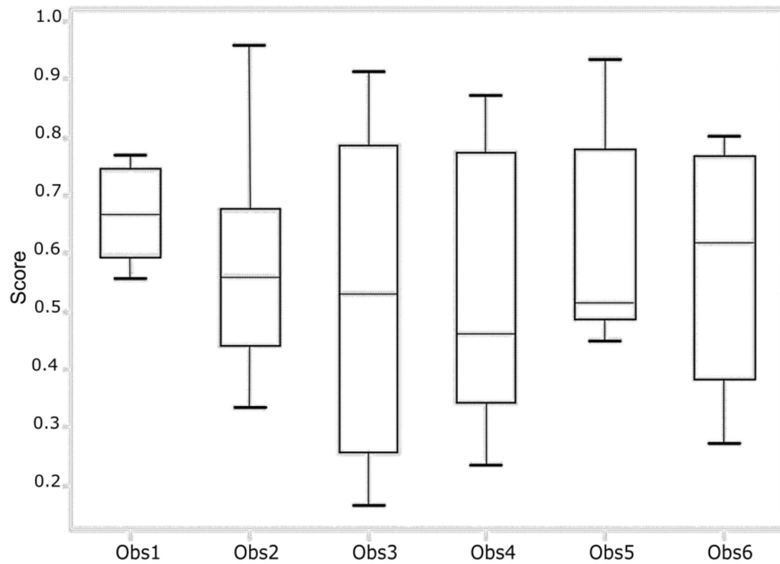

Figure 3a: Performance of each grader in grading the test set.Obs: Observers grading the imagesScore: Fraction of answers in agreement with the ground truth

235x176mm (280 x 280 DPI)

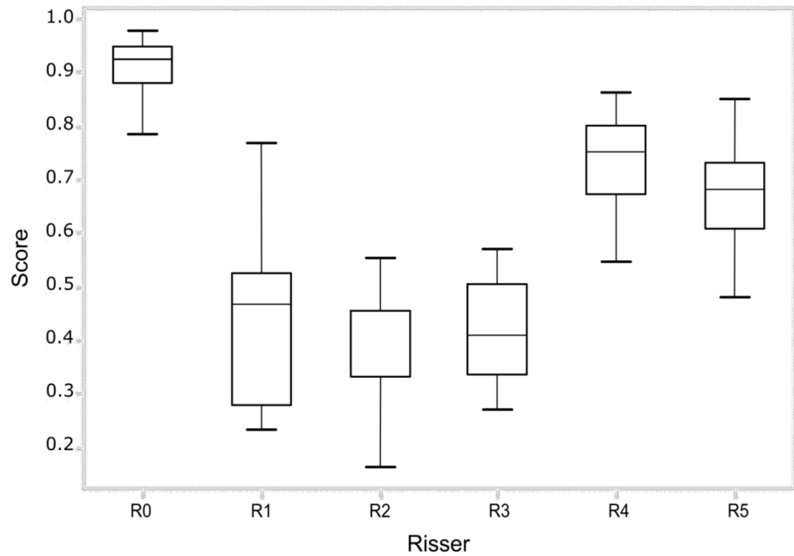

Figure 3b: Performance of all the graders for each Risser stage.R: Risser stage Score: Fraction of answers in agreement with the ground truth

222x153mm (282 x 282 DPI)

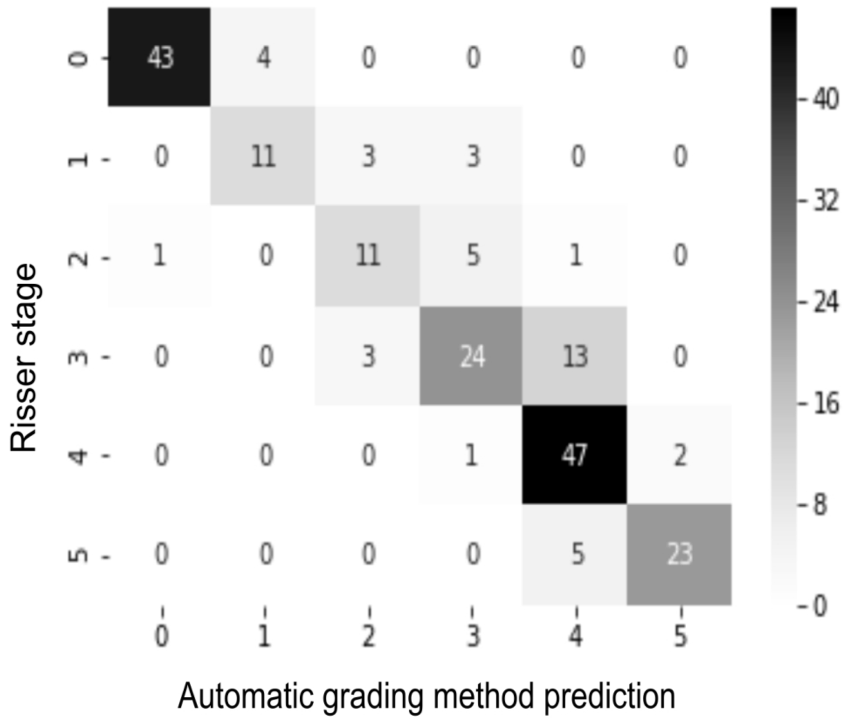

Figure 4a: Confusion matrix for the automatic grading method.
The rows of the matrix show the values indicated by the observer while the column show the ground truth.
The values on the diagonal of the matrix illustrate the number of samples correctly classified by Risser grade.
The values above and below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

143x113mm (282 x 282 DPI)

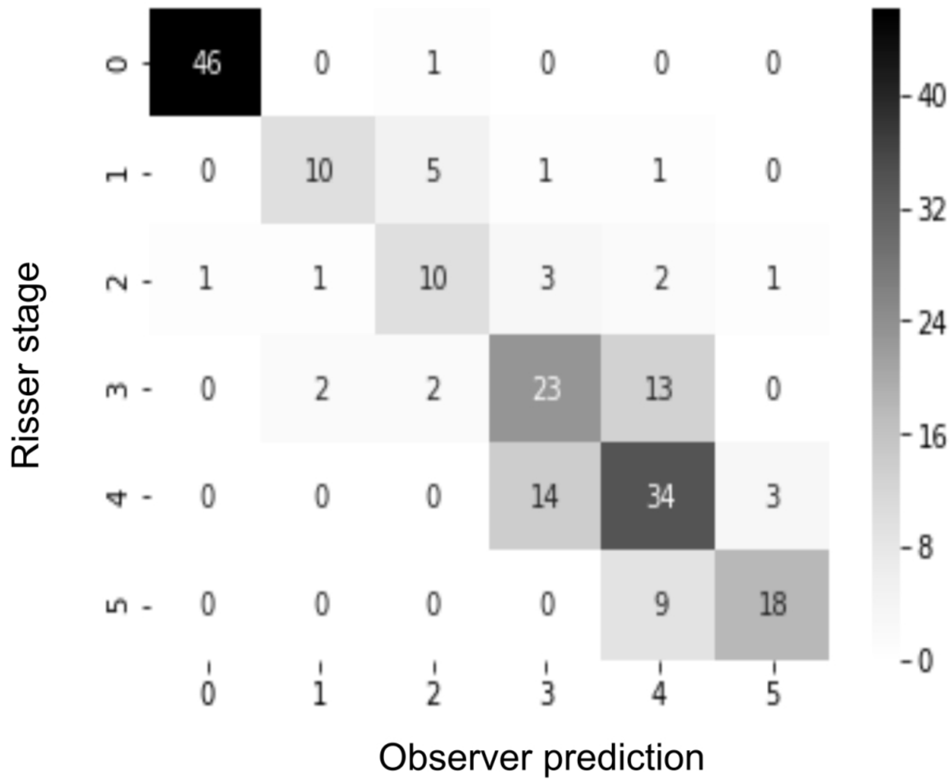

Figure 4b: Confusion matrix for one of the observers.
The rows of the matrix show the values indicated by the observer while the column show the ground truth.
The values on the diagonal of the matrix illustrate the number of samples correctly classified by Risser grade.
The values above and below each value of the diagonal show misclassified samples.

131x110mm (282 x 282 DPI)

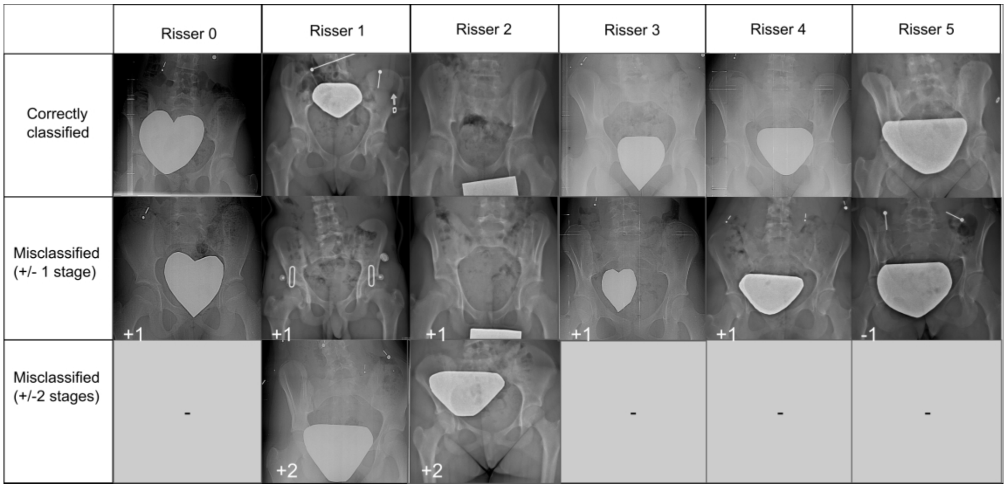

Figure 5a: Sample radiographic images correctly classified by the automatic grading method (top) and misclassified by one grade (2nd row) and two grades (3rd row)

254x122mm (375 x 375 DPI)

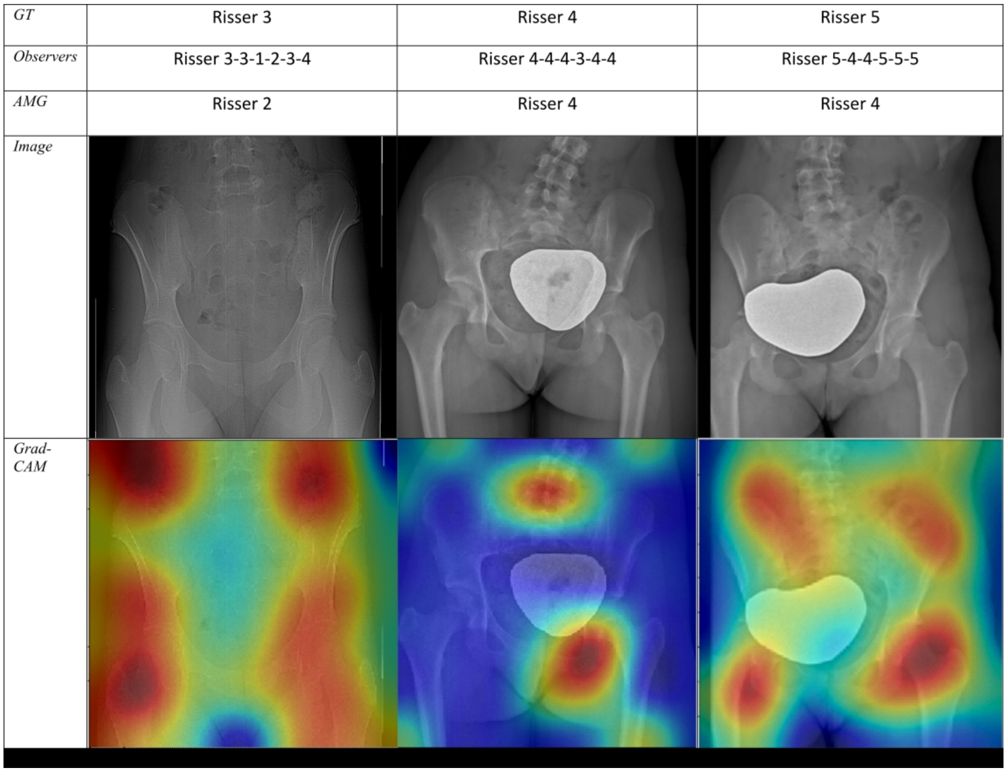

Figure 5b: Sample radiographic images with groud truth (top), observer’s assigned Risser stage (2nd row), automatic grading method (AMG) classification (3rd row) and visualization of the model's attention using Grad-CAM (4th row).

203x155mm (375 x 375 DPI)