Deep Learning for Detection of Complete Anterior Cruciate Ligament Tear

Peter D. Chang[1] · Tony T. Wong[2] · Michael J. Rasiej[2]

Abstract

Deep learning for MRI detection of sports injuries poses unique challenges. To address these difficulties, this study examines the feasibility and incremental benefit of several customized network architectures in evaluation of complete anterior cruciate ligament (ACL) tears. Two hundred sixty patients, ages 18–40, were identified in a retrospective review of knee MRIs obtained from September 2013 to March 2016. Half of the cases demonstrated a complete ACL tear (624 slices), the other half a normal ACL (3520 slices). Two hundred cases were used for training and validation, and the remaining 60 cases as an independent test set. For each exam with an ACL tear, coronal proton density non-fat suppressed sequence was manually annotated to delineate: (1) a bounding-box around the cruciate ligaments; (2) slices containing the tear. Multiple convolutional neural network (CNN) architectures were implemented including variations in input field-of-view and dimensionality. For single-slice CNN architectures, validation accuracy of a dynamic patch-based sampling algorithm (0.765) outperformed both cropped slice (0.720) and full slice (0.680) strategies. Using the dynamic patch-based sampling algorithm as a baseline, a five-slice CNN input (0.915) outperformed both three-slice (0.865) and single-slice (0.765) inputs. The final highest performing five-slice dynamic patch-based sampling algorithm resulted in independent test set AUC, sensitivity, specificity, PPV, and NPV of 0.971, 0.967, 1.00, 0.938, and 1.00. A customized 3D deep learning architecture based on dynamic patch-based sampling demonstrates high performance in detection of complete ACL tears with over 96% test set accuracy. A cropped field-of-view and 3D inputs are critical for high algorithm performance.

Keywords Deep learning · Machine learning · Artificial intelligence · Anterior cruciate ligament · Magnetic resonance imaging

Background

Anterior cruciate ligament tear is a common, important sports injury in adolescents and young adults. A recently published cohort study spanning over two decades discovered an incidence of 68.6 per 100,000 person-years in the general population with peak incidence in adolescents and young adults [1]. Sports injuries are the main source of ACL tears which result in surgery and, in the general population, males have a higher incidence of ACL ruptures than females [2]. This serious diagnosis often requires surgical intervention such as reconstruction or enhanced primary repair to mitigate the risk of subsequent osteoarthritis and chronic instability [3–6]. Furthermore, even with surgical repair, return to sport may be compromised [7].

Accurate, reproducible diagnosis of a complete ACL tear is important for therapeutic decision-making. While a clinical exam, including Lachman and Pivot Shift tests, by an experienced sports medicine physician is essential in post injury evaluation, magnetic resonance imaging is routinely used to confirm suspected diagnosis and to assess for concomitant injuries. MRI imaging is sensitive, specific, and accurate in diagnosing ACL tears [8], especially for an experienced, musculoskeletal-trained radiologist. However, making an accurate diagnosis may still be challenging for a non-MSK radiologist, a trainee on call, or a clinician in a rural area without access to subspecialty radiology. Accordingly, one purpose of this study is to demonstrate feasibility of a fully automated tool for detection of complete ACL tears.

Deep learning is a powerful, emerging branch of machine learning that has in recent years yielded breakthroughs in

✉ Michael J. Rasiej
rasiej@columbia.edu

1   Center for Artificial Intelligence in Diagnostic Medicine, University of California Irvine Medical Center, 101 The City Drive South, Building 55, Suite 201, Orange, CA 92868, USA

2   Department of Radiology, Columbia University Irving Medical Center, 622 West 168th St., MC 28, New York, NY 10032, USA

computer vision benchmarks [9]. The primary advantage of deep learning through convolutional neural networks (CNN) is the ability of the algorithm to learn high-order, semantically meaningful patterns in data without any explicit human programming. Instead, through repeated exposures of input data and desired output, the algorithm is able to iteratively readjust its own neural connections until an abstract, complex representation of the data is learned. This technology underlies almost all of the most recent advances in artificial intelligence over the past several years, from self-driving cars to voice and facial recognition—tasks that just a decade ago would have been impossible.

Given the potential of deep learning technology, there has been a surge of interest to apply it in the healthcare field. However, application of deep learning for MRI detection of sports injuries poses several unique challenges. First, many sports injuries such as ligament and meniscal tears are subtle abnormalities that represent only a small fraction of the overall 3D imaging volume. Second, the abnormality itself may be difficult to assess on a single 2D image slice, as the 3D orientation of the ligament fibers is an important consideration when making the diagnosis. Given these challenges, we hypothesize that a standard network classifier trained slice-by-slice will demonstrate suboptimal performance in this specific interpretative task. Accordingly, this study evaluates the incremental benefit of several customized network architectures with variations in input field-of-view (full slice, cropped slice, dynamic patch-based sampling) and dimensionality (single slice, three slices, five slices) for detection of complete ACL tears. While semi-automated detection of anterior cruciate ligament injuries using support vector machine and random forest techniques has recently been described [10], to the best of our knowledge, this study is the first to use deep learning to assess the ACL.

Methods

Patient Selection and Annotation

After IRB approval, an institutional database was queried for knee MRIs obtained between September 2013 and March 2016. Based on keyword search, patients between 18 and 40 years old with a complete ACL tear were identified. A corresponding control group of normal patients in the same age range but with no ACL pathology was identified. All ACL diagnoses were confirmed through visual inspection by a board-certified subspecialist musculoskeletal radiologist (MJR). Cases with other ACL pathology such as partial tears or mucoid degeneration were excluded.

For each exam, the coronal proton density (PD) non-fat suppressed sequence was downloaded. Each exam containing an ACL tear was manually annotated by a board-certified subspecialist musculoskeletal radiologist (MJR) to delineate (1) a bounding box for each slice containing cruciate ligaments and (2) slices containing a complete ACL tear. All ACL tears were manually annotated using 3D Slicer software (version 4.6).

Our tertiary medical center has MR scanners from General Electric and Siemens. Imaging parameters for a 3T Coronal PD fast spin echo (FSE) sequence include: field-of-view = 16 cm, TE = 20, TR = 3000, slice thickness = 3 mm, gap = 0.3 mm, echo train length = 7, flip angle = 90, frequency = 320, phase = 224, and NEX = 1–4. Imaging parameters for a 1.5T Coronal PD FSE sequence include: field-of-view = 16 cm, TE = min, TR = 3000, slice thickness = 3 mm, gap = 0.3 mm, echo train length = 7, flip angle = 90, frequency = 384–320, phase = 256, NEX = 1–4.

Image Preprocessing

All raw MRI volumes were resampled to an in-plane (coronal) resolution of 256×256 voxels, without change in the overall number of slices in each series. Subsequently, all volumes were independently normalized using a simple z-score map by subtracting the mean intensity value and dividing by the standard deviation. The histogram metrics for mean and standard deviation were calculated after excluding all outlier intensity values below the 1st percentile or above the 99th percentile.

Convolutional Neural Network

To evaluate differences in algorithm accuracy with respect to the input field-of-view, three different architectures were created comprising of several shared networks and blocks shown in Fig. 1 (A–B). First, the entire uncropped MRI slices were used in a simple CNN classifier for determining presence or absence of ACL tear on a slice-by-slice basis. This network was based on a custom ResNet-derived architecture (Fig. 1 (C)) [11]. For the second network, a two-part architecture was implemented whereby an initial localization network was used to detect and generate cropped images of the cruciate ligaments, and a subsequent classifier network was used to determine presence or absence of an ACL tear. The object localization CNN was implemented as a fully convolutional network based on U-net architecture [12], while the classifier CNN was implemented with only minor modifications to the custom ResNet-derived architecture used in the first network (Fig. 1 (D)).

The third network was identical to the second network; however, for the classification CNN, dynamically sampled randomly cropped patches without cruciate ligaments were also used as a new, third class for training (Fig. 1 (E)). Accordingly, this classification network was required to choose from one of three labels: ACL with tear, ACL without
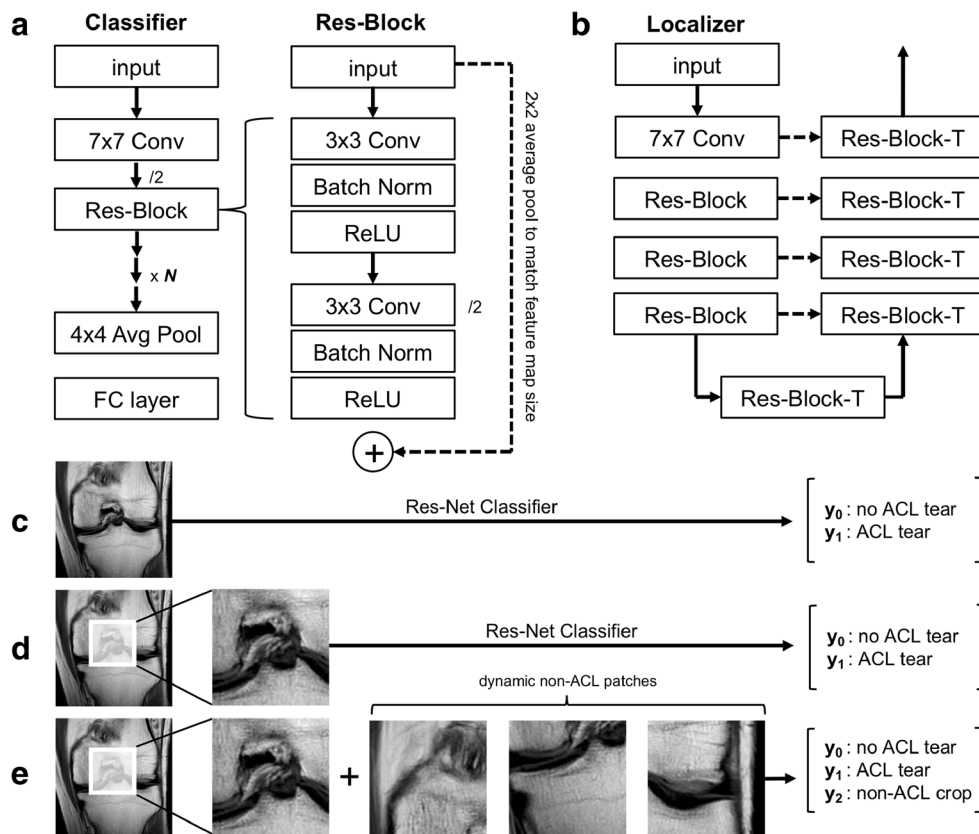
Fig. 1 Overview of network architectures. Two convolutional neural networks (classifier, localizer) and common shared operational blocks are used in various combinations to create three different algorithms for detection of ACL tear. (**a**) The classifier is defined using a single 7×7 convolutional filter with stride 2, followed by a series of residual blocks. The resulting 4×4 feature map is collapsed using an average pool operation. (**b**) The localizer is a fully convolutional U-Net–derived architecture composed primarily of the same residual blocks used by the classifier. In the expanding pathway, the strided convolutions are replaced by convolutional transpose operations to increase feature map size. (**c**) In the first algorithm, entire MRI slices were used by the classifier alone to predict ACL tear. (**d**) In the second algorithm, an initial localizer was used to generate cropped images of the cruciate ligaments, and a subsequent classifier was used to predict ACL tear. (E) In the third algorithm, dynamically sampled randomly cropped patches without cruciate ligaments were used as an additional class for training to promote image diversity

tear, and non-ACL image. Given the small number of training cases in this dataset, the addition of patches without cruciate ligaments significantly increases the diversity of training cases for network learning.

For the classifier network and the contracting pathway of the localizer network, a common shared residual block was defined by a series of 3×3 convolutions whose input and outputs were connected by a residual addition operation (Fig. 1 (A)). In each residual block, the second 3×3 convolution is applied with a stride of 2 along the image height and width to decrease corresponding feature maps by 50% along each dimension. In order to match the input and output feature maps, a 2×2 average pool is applied to the input feature map prior to addition. For the expanding pathway of the localizer network, the strided convolutions are replaced by convolutional transpose operations to expand (rather than decrease) feature map size.

The highest performing of these initial three architectures was then used as the base for experiments to evaluate differences in algorithm accuracy with respect to image dimensionality. In addition to the original 2D (single slice) input, additional networks were created using three-slice and five-slice inputs. For these 3D architectures, feature map dimensionality was decreased in the out-of-plane (anterior-posterior) direction using occasional convolutions with valid padding.

Implementation Details

The network was trained from random weights initialized using the heuristic described by He et al. [13]. The final loss function included a term for L2 regularization to prevent overfitting of data by limiting the squared magnitude of the convolutional weights. Gradients for backpropagation were estimated using the Adam optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments [14]. An initial learning rate of 0.001 was used and

annealed (along with an increase in mini-batch size) whenever a plateau in training loss was observed.

Software code for this study was written in Python 3.5 using the open-source TensorFlow r1.2 library (Apache 2.0 license) [15]. Training was performed on a GPU-optimized workstation with a single NVIDIA GeForce GTX Titan X (12GB, Maxwell architecture).

Statistics

Algorithm accuracy was assessed using per-patient binary classification of presence (one or more abnormal slices) or absence of ACL tear. Additional performance statistics are reported for sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

During the training phase, a fivefold cross-validation paradigm was used whereby 80% of the data was randomly assigned into the training cohort while the remaining 20% was used for validation. This process was then repeated five times until each exam in the entire training dataset was used for validation once. Final results below are reported for the cumulative validation set statistics across the entire training dataset. Additionally, statistics for the top-performing network are also reported for a third independent test set.

Results

Cohorts

A total of 260 patients were included in the analysis, 130 of which had an ACL tear and 130 of which were without ACL pathology. These 260 MRI volumes yielded a total of 4144 coronal PD slices, 624 slices of which contained an ACL tear. Of these, 200 cases (100 each of normal and torn ACLs) were used for initial training and validation, while a separate group of 60 cases was used as a final independent test set.

Accuracy Stratified by Field-of-View

Overall, cross-validation accuracy for detection of ACL injury was higher for networks using images cropped to the cruciate ligaments (0.720–0.765) compared to networks using uncropped full slice inputs (0.68). For the two networks with cropped inputs, addition of dynamically sampled non-ACL patches overall improved algorithm performance, with accuracy, sensitivity, specificity, PPV, and NPV of 0.765, 0.790, 0.740, 0.752, and 0.779 respectively. By comparison, without dynamic sampling, performance statistics for the same metrics were 0.680, 0.670, 0.690, 0.684, and 0.676 respectively. Full cross-validation performance statistics are shown in Table 1.

Accuracy Stratified by Dimensionality

The top-performing network based on cropped images with dynamic sampling was used as the base architecture for experiments to evaluate algorithm accuracy with respect to input image dimensionality. Overall, algorithm performance improved with incremental increase in number of input slices, with cross-validation accuracy of the five-slice network (0.915) better than the three-slice (0.865) or single-slice (0.765) models. Corresponding sensitivity, specificity, PPV, and NPV of the five-slice model was 0.940, 0.890, 0.895, and 0.937, respectively. Full cross-validation performance statistics are shown in Table 1.

Finally, the top-performing five-slice network with dynamic sampling was evaluated on the independent test set of 60 new patients. Performance statistics for corresponding accuracy, sensitivity, specificity, PPV, and NPV of this model was 0.967, 1.00, 0.933, 0.938, and 1.00, respectively.

Algorithm Training

The CNN models were trained for an average of 500 epochs with a batch size of 32. During inference, the final trained networks can generate predictions in approximately 1.4 s per patient.

Discussion

In this study, we demonstrate the feasibility of training a deep learning CNN algorithm to identify the presence of a complete ACL tear with over 96% test set accuracy. Furthermore, we explore various network architectures customized to address the unique challenges of MRI detection of sports injuries. First, we demonstrate the importance of limiting the input field-of-view to the intercondylar region for high algorithm performance. Second, we demonstrate the incremental value of contextual information of adjacent image slices in improving network classification accuracy.

Limiting the input field-of-view through image cropping improves algorithm performance in detection of subtle MRI abnormalities by reducing the image search space. Compared to non-medical image interpretation tasks which tend to rely primarily on global features, pathology on MRI is often localized to small image regions. This is especially true for MRI evaluation of musculoskeletal injuries where many relevant anatomic structures including ligaments, tendons, and menisci are relatively small or have a thin morphology. While deep learning is a very powerful technique, in theory capable of identifying even subtle imaging patterns, increasing amounts of training data are required for detection of progressively smaller image features. Due to the scarcity of high-quality annotated medical images, including the relatively small

Table 1 Network accuracies

	Slices	Set	Accuracy	Sensitivity	Specificity	PPV	NPV
Uncropped	1	CV	0.680	0.670	0.690	0.684	0.676
Cropped	1	CV	0.720	0.710	0.730	0.724	0.716
Cropped + dynamic	1	CV	0.765	0.790	0.740	0.752	0.779
	3	CV	0.865	0.890	0.840	0.848	0.884
	5	CV	0.915	0.940	0.890	0.895	0.937
	5	Test	0.967	1.00	0.933	0.938	1.00

CV, cross-validation; *Test*, test set; *PPV*, positive predictive value; *NPV*, negative predictive value

dataset in this study, cropping the MRI slices to known anatomic landmarks significantly improves algorithm performance (0.680 versus 0.720–0.765).

Another unique feature of medical cross-sectional imaging volumes is the evaluation and interpretation of 3D data. While for certain applications a simplified 2D approach may be appropriate, musculoskeletal injuries are often dependent on synthesis of 3D contextual information. This is especially true for ACL tears, where the oblique 3D orientation of the ligament fibers is a critical consideration in making the diagnosis (e.g., identifying fiber discontinuity requires assessing the trajectory of the ligament on multiple contiguous slices). This hypothesis was confirmed in our study, showing that incremental addition of extra slices for network input yielded progressive improvement in accuracy, from 0.765 (1-slice) to 0.865 (3-slice) to 0.915 (5-slice).

In general, the top-performing five-slice network architecture demonstrated very few classification errors. During initial cross-validation, only 6/100 ACL tears were missed, with 11/100 normal patients misclassified as having a ligament injury. During final test set evaluation, all 30/30 ACL tears were correctly identified, with only 2/30 normal patients misclassified as false positives. Based on visual assessment of the false negatives, some cases of missed ACL tears demonstrated intermediate signal disrupted fibers rather than a more obvious high signal gap in the fibers. In addition, in some missed cases, the tear occurred at the notch origin (Fig. 2). The notch origin tears are more difficult to detect by human readers and occur less frequently, including in our training dataset. Based on visual assessment of the false positives, some cases demonstrated intermediate signal but intact fibers along a segment of the ACL which may have reflected mild focal intrasubstance degeneration rather than a tear (Fig. 3).

Despite the overall high algorithm performance of cross-validation and test set cohorts, there remain several key limitations of this study. First, given the relatively low prevalence of complete ACL tears, the identified patient cohorts used in this study were balanced such that an equal number of injured and normal knee MRIs were used for algorithm development and testing. Given this, the overall PPV for this algorithm would be much lower in an unbalanced patient population reflective of the true prevalence of ruptured ACLs. Furthermore, because of the relatively small number of slices containing an ACL tear even in a balanced patient population (624 out of 4144 slices), classifier networks were trained with stratified sampling such that approximately an equal number of abnormal and normal slices (or patches) were present in each mini-batch. The consequence of this strategy is that, in general, most networks were slightly biased towards high sensitivity (Table 1). A high-sensitivity algorithm, however, may be desired in certain clinical use case scenarios, as highlighted in final test set performance where the algorithm did not miss any ACL tears but identified two false-positive cases.

We used coronal PD non-fat suppressed images for training and testing the network. We chose the coronal imaging plane

Fig. 2 Deep learning predictions, false negatives. Coronal PD images of the knee demonstrating false negative network predictions. Some cases of missed ACL tear demonstrated intermediate signal disrupted fibers (**a**–**c**) or tear at the notch origin of the ACL (**d**). Notch origin tears occur less frequently and are more difficult for human readers to diagnose
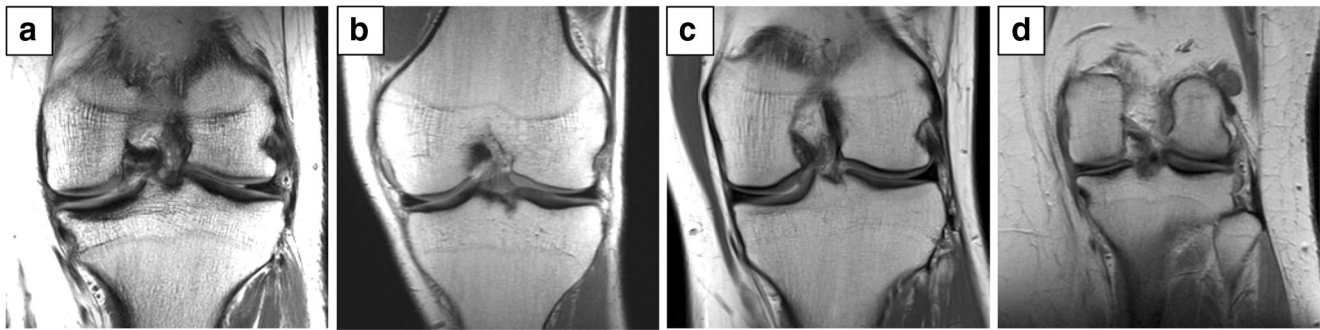
Fig. 3 Deep learning predictions, false positives. Coronal PD images of the knee demonstrating false-positive network predictions. Some false-positive cases demonstrated intermediate signal but intact fibers (**b**–**c**) which may reflect mild focal intrasubstance degeneration without a tear

because it allows the radiologist to trace ACL fibers from origin to insertion in every case regardless of the differences in obliquity of the prescribed plane. Network performance may change if it is trained on images of knees acquired in other planes. Future experiments could examine performance of the network trained on images acquired in sagittal and axial planes, with and without fat suppression.

The proposed deep learning solution for identification of complete ACL tears is presented as proof of concept for application of this new technology to MRI evaluation of musculoskeletal sports injuries. Further research will focus on application of deep learning to more subtle injuries including sprains, partial-thickness tendon and ligaments tears, chondral defects, bone contusions, and meniscal tears which would make it more clinically useful. Eventually, we hope to synthesize detection of these individual image findings using deep learning approaches to generate a coherent overall diagnosis of an injured joint. Furthermore, to improve generalizability on a variety of magnetic field strengths, scanning protocols, and MRI vendors, the work may be expanded to include multiple academic institutions and smaller community hospitals. Developing new algorithms on these larger datasets, with imaging at multiple time points and with various clinical data inputs, may also yield insight into patient outcome and prognosis based on the initial injury pattern.

While the detection of complete ACL tears is not a diagnostic challenge for subspecialized musculoskeletal radiologists, there are nonetheless several potential clinical use case scenarios for the proposed fully automated deep learning algorithm. First, an accurate diagnostic software tool may assist non-MSK trained radiologists, trainees and clinicians in low-access medical settings to evaluate injured knees for ACL ruptures, providing a "second-reader opinion" when subspecialty radiology interpretation is not readily available. Furthermore, while optimal timing of ACL reconstruction post injury remains controversial, evidence suggests that delayed surgery increases the risk of chondral and meniscal damage [16, 17]. In this context, a fully automated deep learning tool could help to triage acute knee injuries for expedited orthopedic surgical evaluation.

Deep learning technology offers tremendous potential to significantly improve the diagnostic accuracy and workflow of radiologists [18]. In this study, we demonstrate the feasibility of a high-performing CNN tool to detect complete ACL injury with over 96% test set accuracy. However, given the unique challenges of automated sports injury detection on MRI, deliberate customized network architectural choices are required for high algorithm performance, which in this study included a dynamic patch-based sampling strategy with a five-slice 3D input. Future directions include further algorithm development on expanded datasets for comprehensive evaluation of sports-related musculoskeletal pathologies.

References

1. Sanders TL, Maradit Kremers H, Bryan AJ, Larson DR, Dahm DL, Levy BA, Stuart MJ, Krych AJ: Incidence of Anterior Cruciate Ligament Tears and Reconstruction: A 21-Year Population-Based Study. Am J Sports Med 44:1502–1507, 2016

2. Gianotti SM, Marshall SW, Hume PA, Bunt L: Incidence of anterior cruciate ligament injury and other knee ligament injuries: A national population-based study. J Sci Med Sport 12:622–627, 2009

3. Perrone GS, Proffen BL, Kiapour AM, Sieker JT, Fleming BC, Murray MM: Bench-to-bedside: Bridge-enhanced anterior cruciate ligament repair. J Orthop Res 35:2606–2612, 2017

4. Louboutin H, Debarge R, Richou J, Selmi TAS, Donell ST, Neyret P, Dubrana F: Osteoarthritis in patients with anterior cruciate ligament rupture: A review of risk factors. The Knee 16:239–244, 2009

5. Satku K, Kumar VP, Ngoi SS: Anterior cruciate ligament injuries. To counsel or to operate? J Bone Joint Surg Br 68:458–461, 1986

6. Fridén T, Zätterström R, Lindstrand A, Moritz U: Disability in anterior cruciate ligament insufficiency: An analysis of 19 untreated patients. Acta Orthop Scand 61:131–135, 1990

7. Myklebust G, Bahr R: Return to play guidelines after anterior cruciate ligament surgery. Br J Sports Med 39:127–131, 2005

8. Mink JH, Levy T, Crues, III JV: Tears of the anterior cruciate ligament and menisci of the knee: MR imaging evaluation. Radiology 167:769–774, 1988

9. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L:

ImageNet Large Scale Visual Recognition Challenge. arXiv: 1409.0575v3 [cs.CV]:1–43, 2015

10. Štajduhar I, Mamula M, Miletić D, Ünal G: Semi-automated detection of anterior cruciate ligament injury from MRI. Comput Methods Programs Biomed 140:151–164, 2017

11. He K, Zhang X, Ren S, Sun J: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90:770-778, 2016

12. Ronneberger O, Fischer P, Brox T: U-net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention 2015. Springer. https://doi.org/10.1007/978-3-319-24574-4_28:234-241, 2015.

13. He K, Zhang X, Ren S, Sun J: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV): IEEE Computer Society. arXiv:1502.01852v1 [cs.CV]:1026–1034, 2015

14. Kingma DP, Ba JL: Adam: A Method for Stochastic Optimization. Published as a conference paper at ICLR 2015. arXiv:1412.6980v9 [cs. LG]:1–15, 2017

15. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467v2 [cs. DC]:1–19, 2016

16. Granan L-P, Bahr R, Lie SA, Engebretsen L: Timing of Anterior Cruciate Ligament Reconstructive Surgery and Risk of Cartilage Lesions and Meniscal Tears: A Cohort Study Based on the Norwegian National Knee Ligament Registry. Am J Sports Med 37:955–961, 2009

17. Kennedy J, Jackson MP, O'Kelly P, Moran R: Timing of reconstruction of the anterior cruciate ligament in athletes and the incidence of secondary pathology within the knee. J Bone Joint Surg Br 92:362–366, 2010

18. Kruskal JB, Berkowitz S, Geis JR, Kim W, Nagy P, Dreyer K: Big Data and Machine Learning-Strategies for Driving This Bus: A Summary of the 2016 Intersociety Summer Conference. J Am Coll Radiol 14:811–817, 2017