Diagnostic and Interventional Imaging (2019) **100**, 243—249

ORIGINAL ARTICLE / *Computer developments*

Artificial intelligence to diagnose meniscus tears on MRI

V. Roblot[a,*], **Y. Giret**[b,c], **M. Bou Antoun**[a], **C. Morillot**[b], **X. Chassin**[b], **A. Cotten**[e], **J. Zerbib**[a], **L. Fournier**[a,d]

[a] *UMR-S970, Department of Radiology, Hôpital Européen Georges-Pompidou, Assistance Publique—Hôpitaux de Paris, Université Paris-Descartes, 75015 Paris, France*
[b] *CentraleSupélec, Université Paris Saclay, 91190 Gif-sur-Yvette, France*
[c] *Foodvisor, 75011 Paris, France*
[d] *Laboratoire de Recherche en Imagerie, LRI, PARCC-HEGP, UMR 970, Inserm/université Paris Descartes, Sorbonne-Paris cité, 75015 Paris, France*
[e] *Department of Musculoskeletal Radiology, Lille University Hospital, 59037 Lille, France*

KEYWORDS
Artificial intelligence (AI);
Meniscus tear;
Magnetic resonance imaging (MRI);
Region convolutional neuronal networks (RCNN);
Convolutional neuronal network (CNN)

Abstract
Purpose: The purpose of this study was to build and evaluate a high-performance algorithm to detect and characterize the presence of a meniscus tear on magnetic resonance imaging examination (MRI) of the knee.
Material and methods: An algorithm was trained on a dataset of 1123 MR images of the knee. We separated the main task into three sub-tasks: first to detect the position of both horns, second to detect the presence of a tear, and last to determine the orientation of the tear. An algorithm based on fast-region convolutional neural network (CNN) and faster-region CNN, was developed to classify the tasks. The algorithm was thus used on a test dataset composed of 700 images for external validation. The performance metric was based on area under the curve (AUC) analysis for each task and a final weighted AUC encompassing the three tasks was calculated.
Results: The use of our algorithm yielded an AUC of 0.92 for the detection of the position of the two meniscal horns, of 0.94 for the presence of a meniscal tear and of 083 for determining the orientation of the tear, resulting in a final weighted AUC of 0.90.
Conclusion: We demonstrate that our algorithm based on fast-region CNN is able to detect meniscal tears and is a first step towards developing more end-to-end artificial intelligence-powered diagnostic tools.

* Corresponding author.
E-mail address: victoire.roblot@aphp.fr (V. Roblot).

https://doi.org/10.1016/j.diii.2019.02.007
2211-5684/

Introduction

Machine learning software has the potential to make workflow more efficient for medical professionals. Indeed, manufacturers of radiological equipment have started integrating artificial intelligence (AI) tools into their medical imaging software systems. This approach is limited by the need to access large numbers of patient data and images that provide learning material for AI algorithms.

The French Radiology Society (SFR) organized a data challenge to detect meniscus tears using a dataset from magnetic resonance imaging (MRI) examination of the knees in October 2018 during the *Journées Francophones de la Radiologie*. The meniscus tear is one of the most frequent cartilage injuries of the knee, and MRI is a useful method for detecting meniscus tears as it has high sensitivity and specificity for that task [1,2].

The purpose of this study was to build and evaluate a high-performance algorithm to detect and characterize the presence of a meniscus tear on MRI of the knee.

Material and methods

Data collection

We formed a multidisciplinary team comprised of three radiologists from the Georges Pompidou hospital of the Assistance Publique—Hôpitaux de Paris (AP—HP) and three engineers including one student from Centrale-Supélec School. Images were collected after obtaining signed informed consent from patients, anonymization, and authorization by the French regulatory authorities (i.e., *Commission nationale de l'informatique et des libertés*).

The training dataset was composed of 1123 MR images of the knee, and the test dataset was composed of 700 images. All these MR images were obtained from 41 hospitals in France. The dataset contained T2-weighted images of the right or left knee passing through the anterior and posterior horns of the medial or lateral meniscus in Nifti format [3,4] with a matrix resolution of 256×256 and isotropic voxels of $0.332\,mm^2$. Only normal menisci or those with abnormal grade 3 high meniscal signal intensity according to Stoller's classification (abnormal hyperintensity that extends to at least one articular surface, superior or inferior) were included. Along with training images, we also had annotation data in a comma-separated values (CSV) file with information about the type of meniscus (lateral or medial), the presence of tear and, if any, its location (anterior or posterior), and orientation (horizontal or longitudinal), which were used as ground truth. Fig. 1 illustrates images from the data set.

The objective of the challenge was to build a high-performance algorithm to detect the presence of a tear in the meniscus, the position of the tear and its orientation. Meniscus tear includes two main mechanisms: degenerative (longitudinal) or post-traumatic, and their diagnosis is based on two important imaging criteria including an abnormal shape of the meniscus and high signal intensity unequivocally in contact with the surface of the meniscus on T2-weighted MRI images. We studied the high signal intensity to detect the tear.

We separated the main task into three sub-tasks:
- meniscus detection (the position of both horns);
- meniscus classification (with or without tear);
- tear classification (orientation of the tear).

A first ''brute-force'' version of the algorithm could have been used to train a neural network to perform a binary classification on the entire image (torn meniscus vs. no torn meniscus). Instead, we decided to follow the steps doctors take to perform the diagnosis without an algorithm: first to find the horns in the image, then diagnose each of them as torn or not. The horn represented a small part of the total image and the tear an even smaller one (Fig. 1). The main information to detect the meniscus tear was the high intensity signal on the horn. However, the limitation of considering only the high intensity signal was that the images were often noisy (Fig. 2).

Convolutional neural network

All three steps of our algorithm were based on convolutional neural network (CNN), and more specifically on fast-region CNN (RCNN) [5] and faster-RCNN [6] backbones.

A CNN is an accumulation (more or less deep) of layers classified into four main categories: convolutional layers, rectification layers, pooling layers, and loss layers [7]. The learning process of a CNN lies in two different steps: the forward pass (computes the loss), and the backward pass (computes the gradient of this loss). The backward pass begins with the loss and computes the gradient with respect to the output. The gradient with respect to the rest of the model is computed layer-by-layer through the chain rule. The weight parameters of the network are usually updated using a stochastic gradient descent (SGD). Using a batch of images (usually around 100) randomly selected from a dataset, we evaluate a loss that is back-propagated to update the weight parameters of the network according to Eq. (1):

$$w = w - \alpha \frac{\partial l}{\partial w} \tag{1}$$

To train a CNN, this operation is repeated thousands of times until we reach convergence. Thus, a larger dataset would result in better CNN training.

RCNN

Object detection tasks recognize an object, and object localization tasks evaluate the coordinates of a bounding box in which the object is situated in the image. The RCNN combines a region proposal algorithm with a CNN [8]. The object detection pipeline consists of three main modules:
- a region proposal algorithm, which extracts category-independent regions of interest in a given image;
- a large and deep CNN that generates a fixed-length feature vector from each region proposal;
- a set of one-versus-all linear support vector machine (SVM) classifiers, which classify each region proposal (Fig. 3).

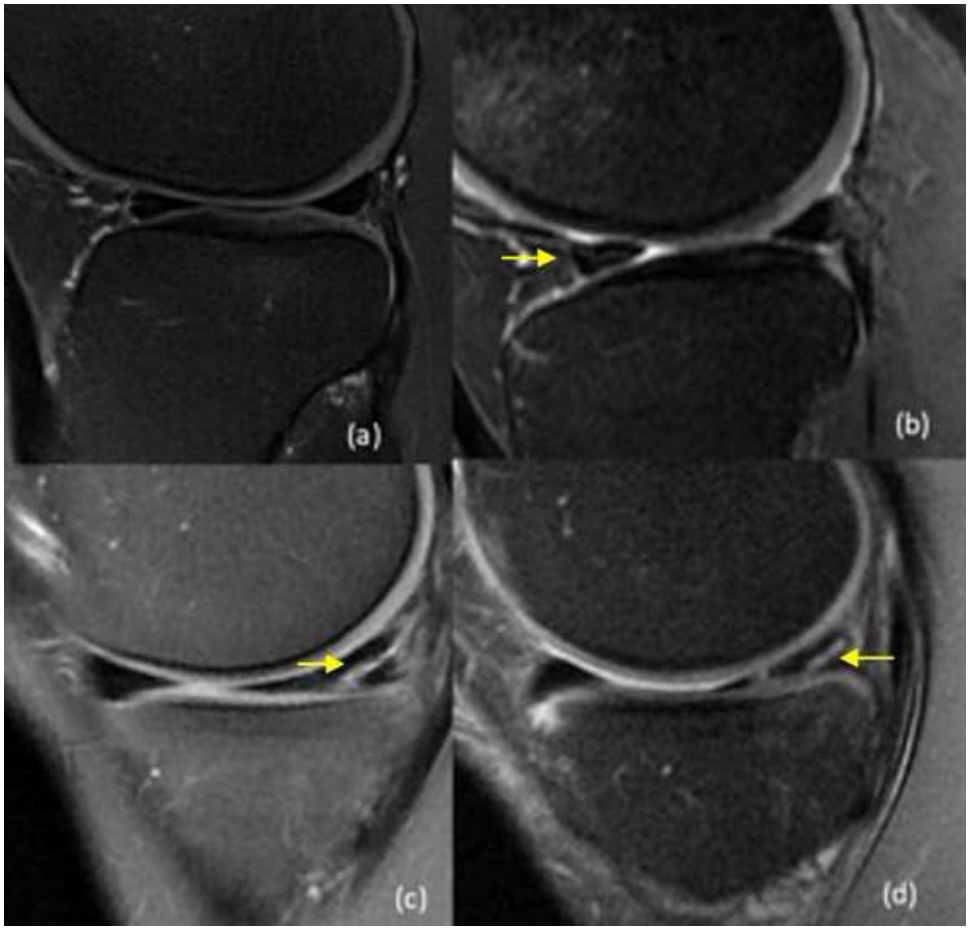

Figure 1. T2-weighted MR images in the sagittal plane through the medial or lateral meniscus show the different types of meniscal images included in the training and validation data sets. (a) Lateral meniscus without tear, (b) lateral meniscus with horizontal tear in the anterior horn (arrow), (c) medial meniscus with vertical tear in the posterior horn (arrow)and (d) medial meniscus with vertical tear in the posterior horn (arrow).

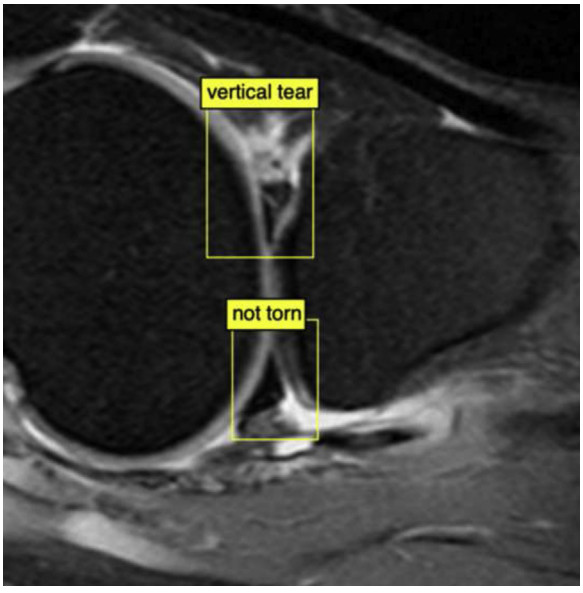

Figure 2. T2-weighted MR image of the knee including bounding boxes with their classification. The image was originally obtained in the sagittal plane and further flipped horizontally and rotated 90° counter-clockwise for analysis purposes.

Fast RCNN

To overcome the limitations of RCNN, a fast RCNN was used [5]. This ''accelerated method'' forwards the whole image in the net and extracts the region proposals only after the last convolutional layer and before the first fully connected layer. This is done through a region of interest (ROI)-pooling layer; given the coordinates of the ROI, and the feature maps supplied by the last convolutional layer of the network (typically an array of size $H \times W \times C$), the ROI-pooling layers output max-pooled feature maps with fixed spatial size $H' \times W'$ and the original C channels ($H' \leq H$ and $W' \leq W$). Thus, instead of forwarding each region proposal through the net one by one, we forward the full image only once, reducing computational time considerably.

Faster RCNN

Since convolutions are shared across region proposals, fast RCNN achieves near real-time rates using deep networks, but this does not include the time spent on region proposals. Region proposal computational time is thus the bottleneck region of interest of these methods. To tackle this barrier, a region proposal network (RPN) that shares full-image convolutional features with the detection network, has

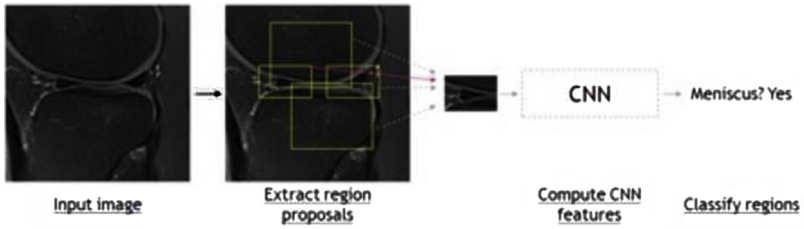

Figure 3. Diagram shows the region convolutional neural network pipeline.

been developed thus enabling nearly cost-free region proposals[6]. A RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position [9]. A RPN takes an image of any size as input and outputs a set of rectangular region proposals that are most likely to contain an object. Each region proposal comes with an objectness score.

RPN can be considered a fully convolutional network, but the main advantage of the RPN described in faster RCNN [7] is that its convolutional layers are shared with the fast RCNN [5] object detection architecture. Thus, both networks can share a large part of their computation. After the last shared convolution layer, a feature map is generated on which they slide a convolution layer of 3×3, followed by two sibling fully-connected layers. A box-regression layer outputs the deltas that must be applied to the coordinates of reference boxes (called anchors) [5], and a classification layer that computes the probability of being an object. At each node of the feature map, κ region proposals are simultaneously predicted corresponding to different scale and aspect ratio. Therefore, for a feature map of dimension $H \times W$ there is a total of $H \times W \times k$ reference boxes. Here, we used 3 scales and 3 aspect ratios, yielding $k = 9$ reference boxes.

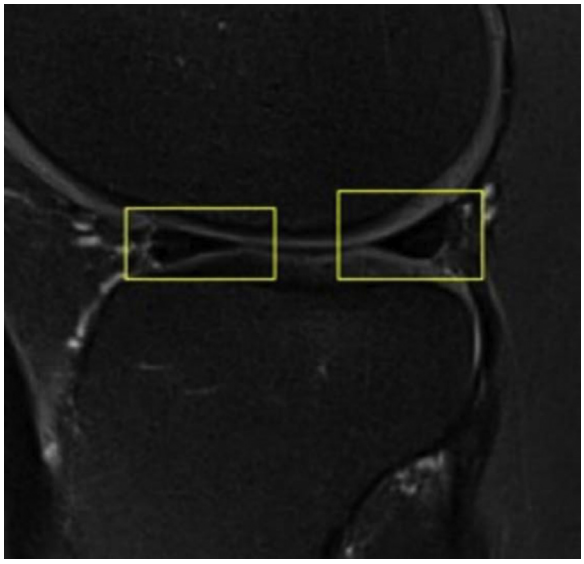

Figure 4. T2-weighted MR image of the knee in sagittal plane with superimposed rectangular regions of interest generated by the faster region convolutional neural network including menisci. Bounding boxes of automatically-generated regions are in yellow.

Training

Dataset

Our training dataset consisted of 1123 images. To train the three steps of our algorithm, we annotated each image with a bounding box surrounding each horn of the meniscus (Fig. 4). Then we added a label of either not torn, horizontal tear, or vertical tear to each bounding box (Fig. 5). We had a total of 2246 menisci with 1948/2246 (87%) non-torn menisci; 298/2246 (13%) with a meniscal tear; 183/272 (61%) menisci with a horizontal tear; and 115/272 (39%) menisci with a vertical tear.

Fast RCNN for meniscus classification

In fast RCNN [5], the selective search algorithm agnostically generates many regions of interest [10]. A label is then assigned to each of these regions. A region that has an intersection-over-union (IOU) higher than a set threshold with any ground-truth box is given the label of the ground truth box. A region with a maximum IOU over all the ground truth boxes of the image below a second set threshold is labeled as ''background''. We typically set the first threshold to 0.3 and the second to 0.1. Instead of using the selective search algorithm to generate the regions of interests, we randomly generate boxes around the ground truth boxes (Fig. 4). This produced similar regions as selective search with the same diversity in size and aspect ratio, but with the advantage of being instantaneous. After labeling regions of interest, we fed them with corresponding images to the network and trained it using SGD and a softmax loss. In order to augment the size of our dataset, we used both the original images and their vertically-flipped version during the training. We trained both classification networks: meniscus classification and tear classification along 16 epochs (Fig. 6).

Faster RCNN for meniscus detection

To train the RPN, we assigned a binary class label (object or not) to each anchor. We assigned a positive label to two kinds of anchors: the anchors that had the highest IOU overlap with a given ground-truth box and an anchor that had an IOU overlap higher than 0.7 with any ground-truth box. On the contrary, we assigned a negative label to an anchor if its IOU ratio was lower than 0.3 for all ground-truth boxes. It is to be noted that with label assignment, there were anchors that were neither positive nor negative. Those anchors were ignored and did not contribute to the training objective. The generated regions were then used to train the fast RCNN part of the network. Each region was assigned a label in the same way as described above. Two methods exist to train the faster RCNN network [6]. The first method consists of

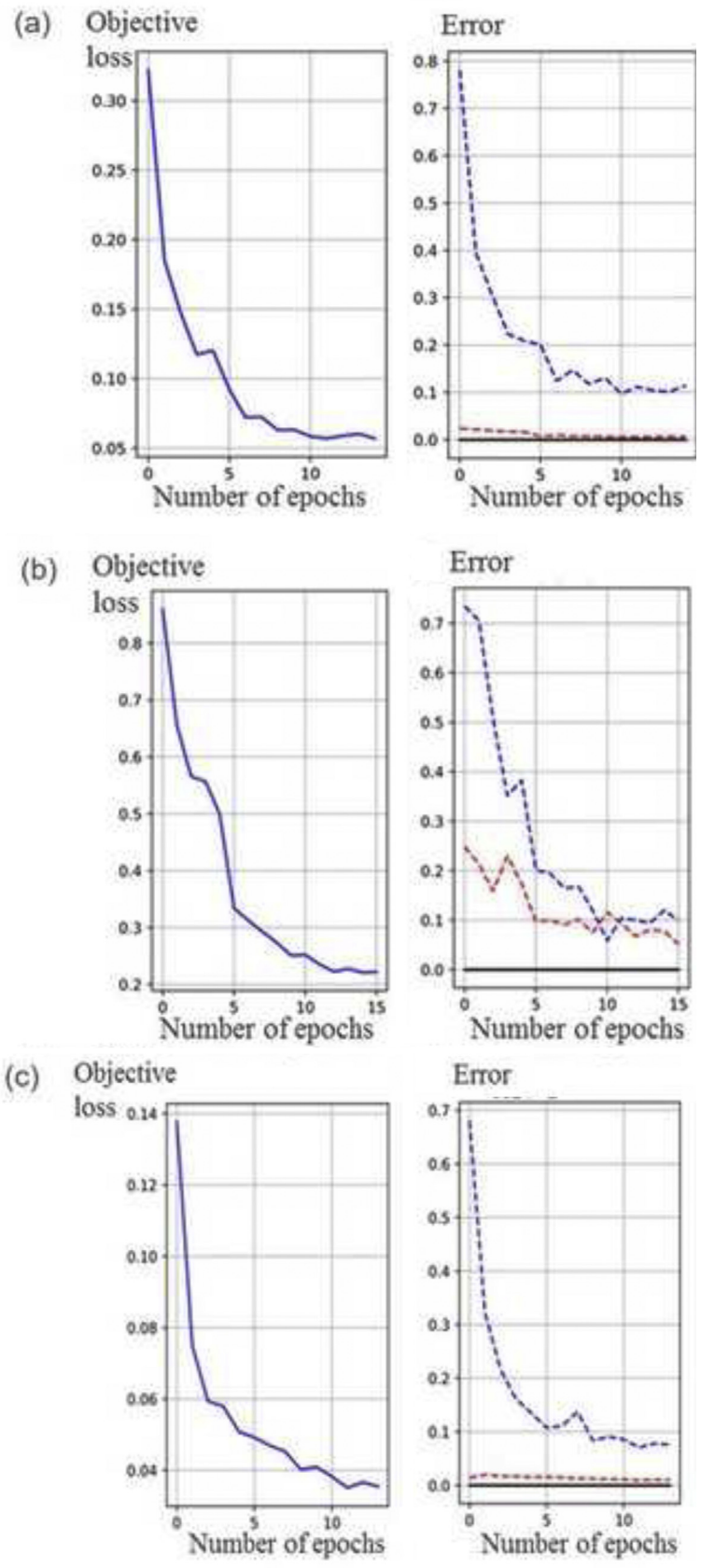

Figure 5. Diagram shows full algorithm pipeline. CNN indicates convolutional neural network. RCNN indicates region convolutional neural network.

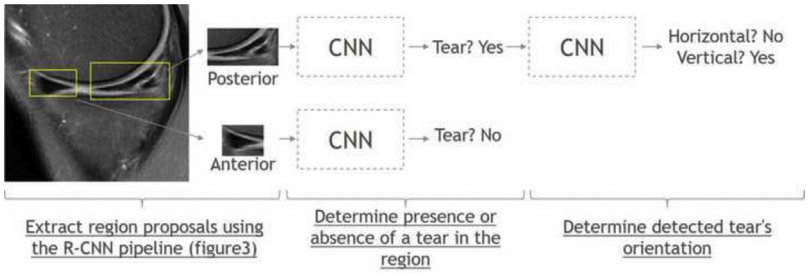

Figure 6. Graphs show training curves of objective loss (left) and error (right) according to number of epochs for (a) the detection of meniscal tear, (b) the location of the tear and (c) the direction of the tear tasks.

learning the RPN and fast RCNN one after another, and the second method consists in learning everything altogether. We chose the latter and trained the CNN along 14 epochs.

Inference

Analysis pipeline description

Once the three specific algorithms were trained, we built an analysis pipeline that took an image as input and gave the diagnosis as output in terms of absence/presence of a torn meniscus. In case of presence of torn meniscus, the output also specified the nature of the tear as horizontal or vertical. First, we transferred the image into the faster RCNN network trained to detect menisci in the image. It output a fixed number of regions (500) with the probability of being a meniscus. We kept only the regions with a probability of 0.5 and higher. Then we applied the non-maximum suppression (NMS) algorithm to remove overlapping regions. Finally, as there should be two menisci per image, we forced the algorithm to output two regions. Each region was then classified using the fast RCNN network trained to distinguish torn and normal menisci (Fig. 5). The input was the image with the extracted regions and the outputs were the probabilities for each region of being a torn or a normal meniscus (the sum of the two probabilities were equal to 1). In a classical binary classification algorithm, the class with the maximum probability would be classified as the meniscus. In this case, this strategy performed poorly and a threshold was set instead on the probability of being a torn meniscus. A threshold of 0.15 achieved the best results. This parameter was shown to affect the whole pipeline (see following section). Finally, for each meniscus classified in the precedent step as being torn, the fast RCNN network trained to classify tears as being horizontal or vertical was used (Fig. 5). The outputs of the network were once again two probabilities (summing to 1): the probability of being a vertical tear and the probability of being a horizontal tear. In this case, the meniscus was attributed to the class with the maximum probability.

Precision and recall

In a detection task, precision and recall are the two main metrics to follow. A high recall algorithm could be an algorithm that classifies all menisci as being torn, but it would have low precision. On the contrary, we could build an algorithm that only classifies a meniscus as torn when it is 100% sure, which would have a very high precision, but a low recall. Therefore, algorithms need to have a good balance between both recall and precision. However, in the case of medical diagnosis, one side may be favored more than another. If this is used as a cleaning tool for doctors, it would be better to favor recall to be sure to detect all torn menisci, with the doctor only removing false positives. The 0.15 threshold set in the torn versus normal classification step was the main parameter influencing the performance of our algorithm; the higher the threshold, the more it favored precision.

Algorithm assessment

The performance metric defined by the challenge was based on area under the curve (AUC). [11]. Considering three distinct tasks, which included detecting the presence, position (anterior of posterior horn), and the orientation (longitudinal or vertical) of the tear, this score was defined as in Equation 2:

$$score = (0.4 \times AUC_{Presence}) + (0.3 \times AUC_{Position}) + (0.3 \times AUC_{Orientation}) \quad (2)$$

Results

On the training dataset of 1123 images, the algorithm obtained a score of 0.95. On the final datasets of 700 images, the following AUCs were obtained: $AUC_{Presence} = 0.94$, $AUC_{Position} = 0.92$, and $AUC_{Orientation} = 0.83$, reaching a total score of 0.90 according to Equation 2.

Next, we merged the meniscus classification and the tear classification phases. Instead of having two binary classifiers, we trained a multi-label classifier with 3 classes, which included not torn, vertical tear, and horizontal tear. Thus, we had only one classification step during the inference phase. However, this did not perform better than the previously described method. Finally, we also tried to merge all three phases. During the training of the faster R-CNN algorithm, instead of only detecting the meniscus, we tried to detect and directly classify the meniscus. Instead of having only two classes (background, meniscus), there were four classes (background, meniscus-not-torn, meniscus-horizontal-tear, meniscus-vertical-tear). During the inference phase, there was only one step. Again, this did not perform better than the previously described method.

Discussion

The deep learning tool developed for this data challenge yielded good performance for diagnosing the presence or absence of a meniscal tear, and localization and direction of the tear when present. Our algorithm was trained on 1123 images but a larger dataset would improve the performance of the algorithm. As large numbers of annotated image data are necessary to develop and test deep learning algorithms, this experience is proof of concept that creating datasets is an achievable endeavor. Since annotation and determination of ground truth is one of the limits to obtaining large datasets, one of the solutions may be to mine medical records and radiology reports through natural language processing techniques [12]. This task could be simplified if radiologists wrote their reports in a consistent and structured format. For example, in this data challenge, each MR image of the knee was annotated with specific diagnoses referenced in the CSV file to help teams be more efficient, and work faster on the creation of the algorithm.

There is only one other study using deep learning networks for diagnosis of meniscal tear [13]. Bien et al. had a similar training dataset of 1370 MRI examinations, but had a greater percentage of examinations containing a meniscal tear (508/1370, 37%) versus 13% in our dataset. Bien et al. used all MR images to train the network, and ground truth was obtained from the majority vote of musculoskeletal radiologists. The network (MRNet) was tailored to detect general abnormalities and specific diagnoses (among which meniscal tears was only one example). The model also included a convolutional neural network that differed from ours in that they used three images where we used only one; they had a full image classification approach while we detected regions of interest first and then classified them; they did not classify the orientation of the tear as horizontal versus vertical; and they used a shallower network (AlexNet vs. VGG) [14]. Performance for the diagnosis of meniscal tear was an AUC of 0.847, slightly lower than our model. It is noteworthy that the model was optimized preferentially for the diagnosis of anterior cruciate ligament tears rather than meniscal tears.

The first limitation of our study relates to the specific dataset created for this challenge, which contained only two T2-weighted MR images in the sagittal plane for each patient whereas MRI examination of the knee usually includes around 100 images. Moreover, images were pre-processed to have the same matrix and voxel size. To generalize and bring any algorithm similar to the one we built to a clinical application, these steps would need to be incorporated into the workflow and a real end-to-end diagnostic tool would need to be developed. Secondly, we strictly analyzed only normal menisci or those with abnormal grade 3 high meniscal signal intensity. There was no abnormal high meniscal signal intensity grade 1 or grade 2. Failure to include these types of lesion, which likely differs from the radiologist's usual practice, limits the applicability of our algorithm, resulting in an added difficulty. Finally, this tool is part of ''narrow artificial intelligence (AI)'' addressing a very specific task in imaging. There is current research focusing on other uses besides computer-aided diagnosis, for example to speed up the time of the examination.

In conclusion, our algorithm, based on RCNN, demonstrated an AUC overall score of 0.9 to detect the presence of meniscal tears, their position, and orientation, and paves the way to develop more end-to-end AI-powered diagnostic tools to help radiologists. Our findings suggest that AI may improve the sensitivity and specificity of diagnoses in radiology, by optimizing workflow efficiency, quality, reduce errors, and inter-observer variability.

Ethical statement

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humanss.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] Lecouvet F, Van Haver T, Acid S, Perlepe V, Kirchgesner T, Vande Berg B, et al. Magnetic resonance imaging (MRI) of the knee: identification of difficult-to-diagnose meniscal lesions. Diagn Interv Imaging 2018;99:55—64.
[2] Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003;226:837—48.
[3] M. Jenkinson. NIfTI-1 Data Format; available from: https://nifti.nimh.nih.gov/nifti-1 [Accessed 23 March, 2019].
[4] Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 2016;264:47—56.
[5] Girschik R. Fast R-CNN. arXiv. org, ArXiv :150408083. Cornell University; 2015.
[6] Ren S, Girschik R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv.org, ArXiv :150601497. Cornell University; 2015.
[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv.org, ArXiv :14091556. Cornell University; 2014.
[8] Girschik R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv.org, ArXiv :13112524. Cornell University; 2013.
[9] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640—51.
[10] Uijilings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int J Comput Vision 2013:104.
[11] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299—310.
[12] Jurafsky D, Martin JH. Speech and Language Processing. 2nd ed. Upper Saddle River, N.J: Prentice Hall; 2008.
[13] Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15:e1002699.
[14] Chen Y, Krishna T, Emer JS, Sze V. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid State Circuits 2017;52:127—38.