Physica Medica 83 (2021) 64–71

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Meniscal lesion detection and characterization in adult knee MRI: A deep learning model approach with external validation

B. Rizk [a,*], H. Brat [b], P. Zille [c], R. Guillin [d], C. Pouchy [c], C. Adam [c], R. Ardon [c], G. d'Assignies [c]

[a] Centre d'Imagerie de Fribourg, Groupe 3R, Rue du Centre 10, 1752 Villars-sur-Glâne, Switzerland
[b] Institut de Radiologie de Sion, Groupe 3R, Rue du Scex 2, Sion, Switzerland
[c] Incepto Medical, 128 rue La Boétie, Paris, France
[d] Department of Medical Imaging, Centre Hospitalier Universitaire de Rennes, Rennes, France

ARTICLE INFO

Keywords:
Meniscus
Deep learning
Knee
Magnetic Resonance Imaging

ABSTRACT

Purpose: Evaluation of a deep learning approach for the detection of meniscal tears and their characterization (presence/absence of migrated meniscal fragment).
Methods: A large annotated adult knee MRI database was built combining medical expertise of radiologists and data scientists' tools. Coronal and sagittal proton density fat suppressed-weighted images of 11,353 knee MRI examinations (10,401 individual patients) paired with their standardized structured reports were retrospectively collected. After database curation, deep learning models were trained and validated on a subset of 8058 examinations. Algorithm performance was evaluated on a test set of 299 examinations reviewed by 5 musculoskeletal specialists and compared to general radiologists' reports. External validation was performed using the publicly available MRNet database. Receiver Operating Characteristic (ROC) curves results and Area Under the Curve (AUC) values were obtained on internal and external databases.
Results: A combined architecture of meniscal localization and lesion classification 3D convolutional neural networks reached AUC values of 0.93 (95% CI 0.82, 0.95) for medial and 0.84 (95% CI 0.78, 0.89) for lateral meniscal tear detection, and 0.91 (95% CI 0.87, 0.94) for medial and 0.95 (95% CI 0.92, 0.97) for lateral meniscal tear migration detection. External validation of the combined medial and lateral meniscal tear detection models resulted in an AUC of 0.83 (95% CI 0.75, 0.90) without further training and 0.89 (95% CI 0.82, 0.95) with fine tuning.
Conclusion: Our deep learning algorithm demonstrated high performance in knee menisci lesion detection and characterization, validated on an external database.

1. Introduction

Knee conditions are common in clinical practice and Magnetic Resonance Imaging (MRI) is the non-invasive method of choice to depict internal joint lesions. MRI detection of meniscal tear correlated to arthroscopic findings shows variable diagnostic performances in systematic reviews [1,2,3], with sensitivity, specificity and accuracy ranging from respectively 83.0 to 93.3%, 69.0 to 88.4% and 81.0 to 86.3% medially, and from 62.0 to 79.3%, 88.0 to 95.7% and 77.0 to 88.8% laterally. Sensitivity and specificity of MRI tear migration are respectively of 69% and 94% for notch fragment and 71% and 98% for recess fragments [4].

Beyond prescription appropriateness, clinically significant diagnostic errors may impact active patients, with unnecessary interventions or treatment delays. The development of automated machine learning based tools may assist and increase diagnostic performances of general radiologists. Deep learning (DL) models have been proposed in medical imaging over recent years for an increasing number of tasks and with improving performances, fueled by strong collaborative efforts between radiologists and data scientists. Machine learning based knee injuries

Abbreviations: MRI, Magnetic Resonance Imaging; DL, Deep learning; CNN, Convolutional Neural Networks; ACL, Anterior Cruciate Ligament; SFR, Société Française de Radiologie (French Radiology Society) ; MSK, MusculoSKeletal; AI, Artificial Intelligence; PD, Proton Density; FS, Fat Suppressed; NLP, Natural Language Processing; IoU, Intersection over Union; ReLU, Rectified Linear Unit; GRU, Gated Recurrent Unit; CBOW, Continuous Bag of Words; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve; DICOM, Digital Imaging and COmmunications in Medicine; CI, Confidence Interval.

* Corresponding author.
E-mail addresses: benrizk@gmx.net, benoit.rizk@groupe3r.ch (B. Rizk).

https://doi.org/10.1016/j.ejmp.2021.02.010
Received 30 November 2020; Received in revised form 31 January 2021; Accepted 16 February 2021

detection models (usually focused on anterior cruciate ligament (ACL), meniscal or cartilage lesions) from MRI imaging have been proposed in the literature [5]. Bien et al. [6] used aggregated 2D convolutional neural networks (CNN) to detect both general abnormalities and specific diagnoses (ACL and meniscal tears) from knee MRI examinations and published their dataset, MRNet. Pedoia et al. [7] performed automatic segmentation of cartilage and menisci using 2D U-Net architectures, followed by automatic detection and severity grading of meniscal and cartilage lesion using a 3D CNN. A data challenge organized by the French Radiology Society whose goal was to identify a meniscal tear on MRI on a given dataset led to 2 published articles by the winning teams [8,9]. Finally, Fritz et al. [10] compared musculoskeletal radiologists with a deep convolutional neural network-based model for the detection of meniscal tears using surgery as standard of reference.

In the recent literature, these artificial intelligence (AI) applications remain mostly experimental and few studies provide external validation which could enhance robustness, generalizability and safety of clinical implementation of these tools in the assessment of patients in a real-world production setting.

By adding to the literature a well-powered externally validated algorithm for the detection and characterization of meniscal tears, our study aims to bridging the gap of bringing AI into routine radiologist practice.

2. Materials and methods

2.1. Database creation

We retrospectively collected 11,353 knee examinations from 10,401 adult patients who underwent knee MRI examinations between 2009 and 2018 from 11 medical imaging centers in Switzerland. Our multicentric institution has a general consent form signed by each patient to allow or refuse retrospective data analysis for research purposes. MRI images and reports used for the database were anonymized with removal of personal information. Patients under the age of 16 (N = 309) and those with a known past knee surgical history (N = 2189) were excluded (Fig. 1), leaving 8058 examinations with coronal and sagittal proton density (PD) fat suppressed (FS)-weighted images. Images were obtained from 13 MRI scanners, distributed mainly among Philips Panorama 1 Tesla (54.0%) and Philips Ingenia 3 Tesla (36.3%) equipment (Table 1). The content of the corresponding radiological structured standardized reports was extracted using Natural Language Processing (NLP) algorithms. The population consisted in 48.1% of female and 51.9% of male patients, with a mean age of 44.8 years (range 16–89) and a mean weight of 74.3 kg (range 38–186).

2.2. Meniscal localization

A random subset of 1000 examinations was manually annotated by two data scientists, trained by a senior radiologist to recognize menisci on 50 MR examinations. 3D bounding boxes normalized in the range [0,1] were placed around medial and lateral menisci without segmentation, using an in-house annotation tool. Using 3D bounding boxes instead of more advanced types of annotations (e.g. dense segmentations of the menisci) for the meniscal localization task offers several advantages: (i) 3D dense segmentation annotations are extremely time-consuming to obtain, while drawing a 3D bounding box englobing the area of interest is much faster; (ii) Deep learning architectures performing dense segmentations (such as 3D U-Net or V-net) are computationally expensive, while predicting 3D bounding box coordinates can be achieved using a standard CNN architecture with a multi-dimensional output (2 sets of 3 scalar coordinates for each bounding box).

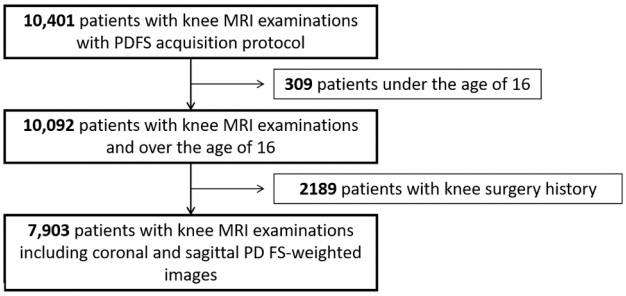

Fig. 1. Database curation.

Table 1
Study population and distribution.

Statistic	Database
Number of patients	7903
Female / Male ratio (%)	48.1 / 51.9
Mean age (years) (range)	43.6 (16–120)
Mean weight (kg) (range)	74.3 (38–186)
Total number of examinations	8058
Number of examinations on Philips Panorama 1 T system (%)	4348 (54.0)
Number of examinations on Philips Ingenia 3 T system (%)	2929 (36.3)
Number of examinations on GE ONI MSK Extreme 1.5 T system (%)	392 (4.9)
Number of examinations on GE Optima MR430s 1.5 T system (%)	330 (4.1)
Number of examinations on GE Signa Pioneer 3 T system (%)	53 (0.7)
Number of examinations on GE Signa HDxt 1.5 T system (%)	4 (0.0)
Number of examinations on SIEMENS Skyra 3 T system (%)	2 (0.0)

This annotated database was used as a training set for two coronal and sagittal CNN-based localization models to extract bounding boxes coordinates around both menisci in a given MRI series. Both coronal and sagittal CNN-based meniscus localization models contained 4 convolution blocks made of layers of (16,8,16)/(16)/(128,32,32)/(64,128,8,128) and (8)/(64,32,32,8)/(8,16,128)/(8,32) convolution kernels, respectively. Each convolution layer was followed by a rectified linear unit (ReLU) activation and a batch normalization step. Max-pooling (factor 2) was applied after each convolution block, and global average pooling followed by a ReLU activation to output the final localization results made of 12 coordinates (2 sets of 3 coordinates representing upper-left and lower-right corners for each meniscal bounding box). Both coronal and sagittal models were trained using an Adam optimizer, L1 regression loss, with an initial learning rate of 1e-5 and for 41 and 35 epochs, respectively. No dropout was applied for any of the networks. No data augmentation techniques have been used during the training phase of these networks.

The performance of the models was evaluated on a test set of 100 examinations annotated by a musculoskeletal radiologist with 10 years of experience. Intersection over Union (IoU) evaluation metric was used to measure the localizer model accuracy. Suppose we have two bounding boxes denoted by A and B, respectively. Denote $|I| = |A \bigcap B|$ the intersection between A and B, and $|U| = |A \bigcup B|$ the union of A and B.

According to Rezatofighi et al. [11], the IoU is the ratio defined as:

$$IoU = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|I|}{|U|}$$

2.3. Meniscal tear detection

Using an in-house text annotator tool, another random subset of 2611 examinations was manually labelled from radiological reports by a team of 4 trained data scientists trained and assisted by 2 experienced (17 and 15 years) radiologists for absence / presence of medial and lateral meniscal tear, according to the following key-words: tear, lesion, flap, bucket-handle, parrot beak, cleavage, morphology distortion, free fragment [12]. This labelled database was used as a training set for a bidirectional Gated Recurrent Unit citation Neural Network (GRU)-based NLP model [13] to extract keywords and to label the entire database. A ten-fold cross validation was used for performance analysis.

We then fed the recurrent neural network with word embeddings computed with Word2Vec algorithm [14]. Word representations were obtained using the Continuous Bag of Words (CBOW) architecture on our own reports database. The Bidirectional GRU made a prediction after processing each embedded word from the report. NLP model performance for medial and lateral meniscal tear detection from reports was evaluated by ROC curves, AUC serving as a quantitative performance indicator.

Meniscal crops produced by the localization models were resized to a common size of 64x64x64 across volumes, and then fed with the NLP found labels (both hand annotated and NLP-inferred annotations) into a CNN-based, meniscal tear detection model, common for both coronal and sagittal series. Both medial and lateral CNN-based meniscal tear detection models contained 3 convolution blocks made of layers of (32,32,32)/(32,32,32)/(16,64,128) and (32,32,32)/(32,32,32)/(128,32) convolution kernels, respectively. Each convolution layer was followed by a ReLU activation and a batch normalization step. Maxpooling (factor 2) was applied after each convolution block, and global average pooling followed by a sigmoid activation to output the final binary classification result. Both medial and lateral models were trained using an Adam optimizer, L1 regression loss, with an initial learning rate of 1e-5 and for 21 and 15 epochs, respectively. No dropout was applied for any of the networks. No data augmentation techniques have been used during the training phase of these networks. Meniscal tear detection pipeline is illustrated in Fig. 2.

The database (N = 8058) was divided into 3 non-overlapping splits for training (N = 6221), validation (N = 1538) and testing (N = 299). The meniscal tear detection model's final results were aggregated within examinations using the average prediction scores across sagittal series and coronal series.

The performance of this model was evaluated on a test set of 299 examinations annotated with an in-house DICOM image annotator tool by a team of 5 musculoskeletal (MSK) radiologists, classifying for each meniscus the status of presence/absence of tear and migration. Inter-observer variability was calculated using Kappa scores. Mismatches (N = 82) were reviewed by 2 MSK radiologists in consensus. Distribution of meniscal tears on training/validation sets and on test set are provided in Table 2. Demographics statistics between training/validation sets and test set are provided in Table 3.

2.4. Deep learning models interpretation

To gain some insight into which areas of the image are the most discriminative for our meniscal tear detection network, we used a noisy perturbation-based model. Gaussian noise was successfully applied to overlapping patches within the image. By comparing the prediction score from the original image and the ones obtained by the perturbated images, we computed a heatmap highlighting areas that influences the most the prediction when perturbated. We then applied a simple threshold to the resulting heatmap (only keeping values above the 99th percentile), as well as a gaussian filter for visual ease. Examples of resulting heatmaps can be seen in Fig. 3.

Table 2
Descriptive statistics for training/validation and testing sets for tear detection and migrated tear characterization.

Statistic	Training and validation sets	Test set
Number of examinations for meniscal tear detection	7759	299
Number of annotated examinations (%)	–	299 (100)
Number of overlapping annotated examinations (%)	–	176 (59)
Number of medial meniscal tear (%)	2607 (33.6)	171 (57.2)
Number of lateral meniscal tear (%)	846 (10.9)	89 (29.8)
Number of examinations for meniscal tear characterization	1133	299
Number of migrated medial meniscal tear (%)	453 (40.2)	77 (25.8)
Number of migrated lateral meniscal tear (%)	141 (12.5)	21 (7.0)

2.5. Meniscal tear characterization

Meniscal tear characterization was defined as presence or absence of a migrated meniscal fragment. Radiological reports from a random subset of 1133 examinations were manually labelled by a team of 4 trained data scientists and 2 experienced radiologists, according to following keywords: free fragment, displaced, migrated, flap, bucket-handle. These labels, combined with meniscal crops produced by the localization model previously described, were used to feed two CNN-based migrated meniscal tears detection models (one for coronal series, and one for sagittal series).

The medial coronal and sagittal, lateral coronal and sagittal meniscal tear characterization models contained 4 convolution blocks made of convolution layers of (32,32)/(64,64)/(32,128)/(32,32), (32,32)/(32,32,32)/(32,16)/(128,16), (32,32)/(64,64,64)/(16,16)/(64), (32,32)/(64,64)/(32,128)/(32,32) convolution kernels, respectively. Each convolution layer was followed by a ReLU activation and a batch normalization step. Maxpooling (factor 2) was applied after each convolution block, and global average pooling followed by a sigmoid activation to output the final binary classification result. The medial coronal, medial sagittal, lateral coronal and lateral sagittal meniscal tear characterization models were trained using an Adam optimizer, L1 regression loss, with an initial learning rate of 1e-5 and for 49, 38, 50 and 50 epochs, respectively. No dropout was applied for any of the networks. No data augmentation techniques have been used during the training phase of these networks. Meniscal tear characterization pipeline is illustrated in Fig. 4.

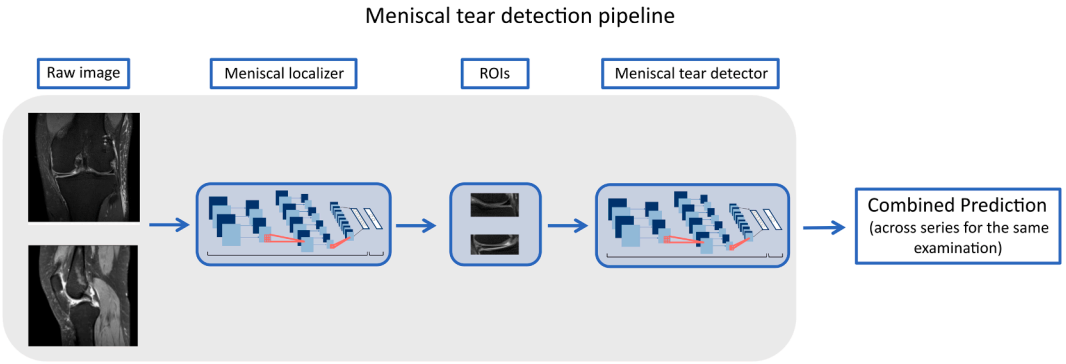

Fig. 2. Meniscal tear detection pipeline.

Table 3
Study population and distribution along splits.

Statistic		Training and validation sets	Test set	P-value
Number of examinations		7759	299	
Female / Male ratio (%)		48.1 / 51.9	50.8/ 49.2	0.347
Mean age (years) (range)		43.4 (16–120)	47.7 (16–105)	<0.001
Mean weight (kg) (range)		74.3 (38–186)	74.5 (38–178)	0.732
Number of examinationsper manufacturer				<0.001
	Philips Panorama 1 T system (%)	4343 (59.9)	5 (1.7)	
	Philips Ingenia 3 T system (%)	2643 (34.1)	286 (95.6)	
	GE ONI MSK Extreme 1.5 T system (%)	390 (5.0)	2 (0.7)	
	GE Optima MR430s 1.5 T system (%)	324 (4.2)	6 (2.0)	
	GE Signa Pioneer 3 T system (%)	53 (0.7)		
	GE Signa HDxt 1.5 T system (%)	4 (0.0)		
	SIEMENS Skyra 3 T system (%)	2 (0.0)		

The database (N = 1432) was divided into 3 non-overlapping groups of training (N = 898), validation (N = 235) and testing (N = 299). Distribution of migrated meniscal tear on training/validation sets and on test set are described in Table 4.

The models' final results were aggregated within examinations using the average prediction scores across sagittal series and coronal series. At last, we combined both the meniscal tear detection and characterization pipelines for evaluation on the test set: migration prediction was only performed when the prediction score of the meniscal tear detection model was above a defined threshold. Sensitivity, specificity and accuracy of meniscal tear detection and characterization in the radiological reports are compared to deep learning performances.

2.6. Meniscal tear detection external validation

Our combined CNN meniscal tear detection model was then validated on publicly available MRNet dataset from Bien et al. [6]. This database is composed of 1250 knee MRI examinations (1130 subdivided into 80/20% splits for training/validation, 120 for testing) annotated by 3 MSK radiologists. It contains the following sequences: coronal T1 weighted, coronal T2 FS, sagittal PD weighted, sagittal T2 with fat saturation, and axial PD weighted with fat saturation, performed exclusively with GE MRs.

Since no distinction between medial or lateral meniscal tear is possible from the available labels in the external database, we merged predictions of both our algorithms (medial and lateral menisci) into a single global tear prediction.

Performances of our models were measured using ROC curves and AUC values. In addition, we also provide performances after using the training set of the MRNet dataset to fine-tune (equivalently, retrain) our models on these additional data samples.

2.7. Statistical analysis

Performance metrics for the localization models were IoU values and their associated standard deviations. Performance metrics for the classification models included AUC, sensitivity, specificity and accuracy values as well as their respective confidence intervals. These confidence intervals were calculated using bootstrap [15] method with replacement. Once the model training processed is performed, successive random draws of prediction values of the statistics of interest are used to compute its resampled distribution. Quantilized values of that resampled distribution for a given α level provide its confidence interval. In this work, we used n = 10000 for each confidence interval calculation. We would like to stress the fact that the training process is only carried out once, and not for each bootstrap sample.

2.8. Computational tools

All training experiments were undertaken using the following software packages: Python 3.6, Keras 2.2.5, Tensorflow 1.15.0, Scikit-learn 0.22.1, and Numpy 1.19.1.

In addition, calculations were ran using Amazon Web Service cloud-based P3 instances, using customized Intel Xeon processors running at 2.7 GHz, and NVIDIA Tesla V100 GPUs with 16G of memory.

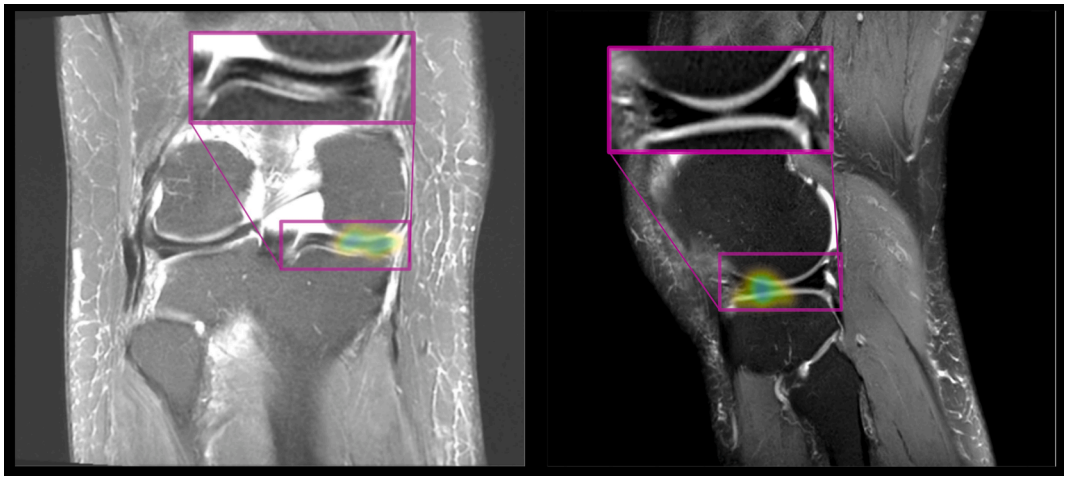

Fig. 3. Examples of perturbation-based feature interpretation heatmaps for our meniscal tear detector. Left: the resulting heatmap properly overlaps with a meniscal tear. Right: the heatmap doesn't correspond to a meniscal tear.

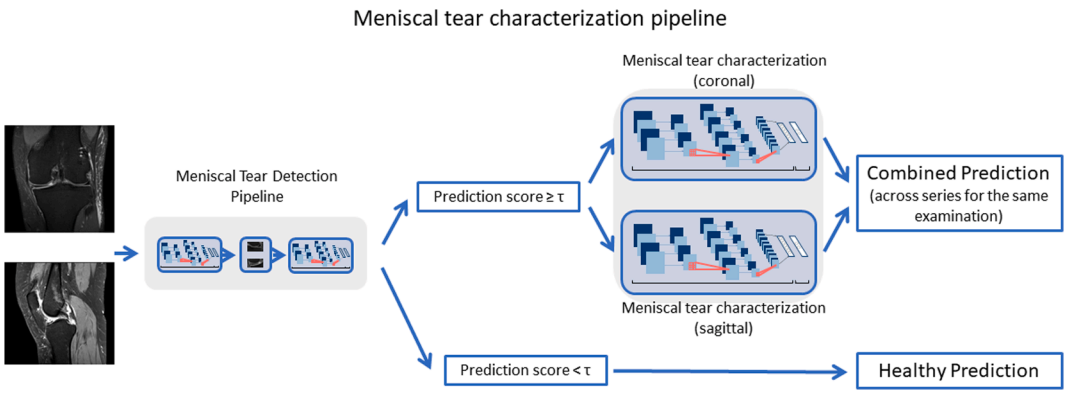

Fig. 4. Meniscal tear characterization pipeline.

Table 4
Migrated meniscal tears prevalence along splits.

Statistic	Training and validation sets	Test set
Number of examinations for meniscal tear characterization	1133	299
Number of migrated medial meniscal tear (%)	453 (40.2)	77 (25.8)
Number of migrated lateral meniscal tear (%)	141 (12.5)	21 (7.0)

3. Results

3.1. Meniscal localization

The meniscal localization pipeline resulted in IoU values for coronal and sagittal series of 0.85 ± 0.12 (lateral) / 0.81 ± 0.15 (medial) and 0.82 ± 0.15 (lateral) / 0.82 ± 0.15 (medial), respectively (Fig. 5). No significant statistical effect has been observed for differences in IoU between scanners (Philips Panorama 1 T vs. Philips Ingenia 3 T) or sex.

3.2. Meniscal tear labels extraction with NLP

The meniscal tear label NLP extraction model resulted in AUC, specificity and sensitivity values for medial/lateral meniscus of 0.99 (95% CI 0.97, 1.00)/ 0.98 (95% CI 0.97, 1.00), 0.99 (95% CI 0.98, 1.00)/ 0.99 (95% CI 0.82, 1.00) and 0.99 (95% CI 0.82, 1.00)/ 0.98 (95% CI 0.82, 1.00), respectively.

3.3. Meniscal tear detection

Kappa scores for inter-observer variability regarding presence/absence of tear and migration are reported in Table 5. On the testing set, AUC, sensitivity, specificity and accuracy values for medial/lateral meniscal tear detection models were 0.93 (95% CI 0.82, 0.95)/0.84 (95% CI 0.78, 0.89), 0.89 (95% CI 0.84, 0.93)/0.67 (95% CI 0.57, 0.77), 0.84 (95% CI 0.76, 0.90)/0.88 (95% CI 0.84, 0.92) and 0.87 (95% CI 0.83, 0.90)/0.82 (95% CI 0.78, 0.86), respectively (Fig. 6).

Table 5
Inter-annotators Kappa score for all graded items.

	Medial meniscus tear	Lateral meniscus tear	Medial meniscus migrated tear	Lateral meniscus migrated tear
Kappa score	0.86 (95% CI 0.83, 0.89)	0.77 (95% CI 0.71, 0.93)	0.83 (95% CI 0.81, 0.86)	0.93 (95% CI 0.91, 0.95)

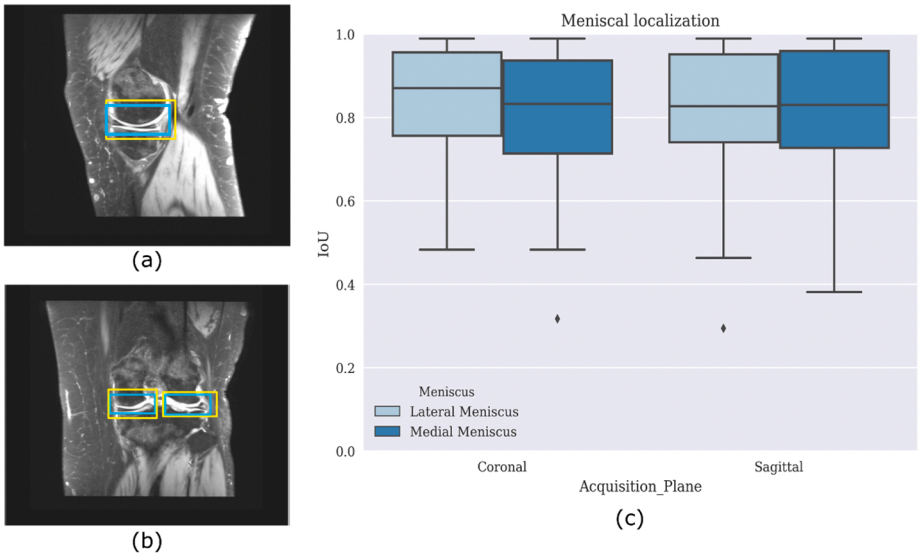

Fig. 5. Meniscal localization results. (a-b) Meniscal bounding boxes predictions (blue) compared to hand drawn (yellow) boxes. (c) Box diagram of meniscal localization algorithms predictions.

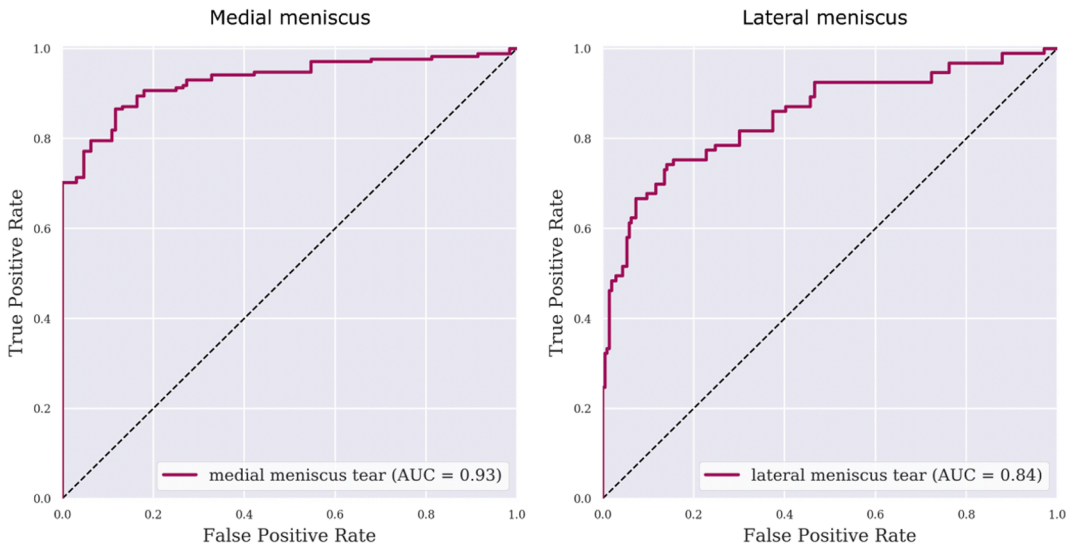

Fig. 6. Meniscal tear detection ROC curves.

3.4. Meniscal tear characterization

On the testing set, AUC, sensitivity, specificity and accuracy values for medial/lateral meniscal tear migration characterization models were 0.91 (95% CI 0.87, 0.94)/0.95 (95% CI 0.92, 0.97), 0.80 (95% CI 0.69, 0.89) /0.57 (95% CI 0.33, 0.80), 0.85 (95% CI 0.80, 0.89)/0.95 (95% CI 0.93, 0.98) and 0.83 (95% CI 0.79, 0.88)/0.93 (95% CI 0.90, 0.96), respectively (Fig. 7). Sensitivity, specificity and accuracy of meniscal tear detection and characterization in the radiological reports, compared to expert MSK annotators, are presented in Table 6.

3.5. Meniscal tear detection external validation

Our full pipeline, including localization and classification models, resulted in AUC, sensitivity, specificity and accuracy values for meniscal tear detection without/with finetuning of 0.83 (95% CI 0.75, 0.90)/0.89 (95% CI 0.82, 0.95), 0.77 (95% CI 0.65, 0.88)/0.81 (95% CI 0.69, 0.91), 0.84 (95% CI 0.75, 0.92) /0.87 (95% CI 0.78, 0.94) and 0.81 (95% CI 0.73, 0.88) / 0.84 (95% CI 0.78, 0.90), respectively (Fig. 8).

Table 6
First reviewer (radiological report) performances for all graded items.

First reviewer performances	Sensitivity	Specificity	Accuracy
Medial meniscus tear	0.98 (95% CI 0.96, 1.0)	0.85 (95% CI 0.79, 0.90)	0.92 (95% CI 0.90, 0.95)
Lateral meniscus tear	0.75 (95% CI 0.66, 0.84)	0.97 (95% CI 0.95, 0.99)	0.92 (95% CI 0.89, 0.95)
Medial meniscus migrated tear	0.37 (95% CI 0.27, 0.48)	0.95 (95% CI 0.90, 0.98)	0.71 (95% CI 0.65, 0.77)
Lateral meniscus migrated tear	0.27 (95% CI 0.0, 0.55)	1.0 (95% CI 1.0, 1.0)	0.87 (95% CI 0.79, 0.95)

4. Discussion

Using a real-world large dataset of adult knee-MRI, our algorithms achieved high and stable externally validated performances in detecting meniscal tears. According to published literature and confirmed by our data, human performances are limited for meniscal fragment detection. Our study bridged the gap to clinical routine fueled by strong

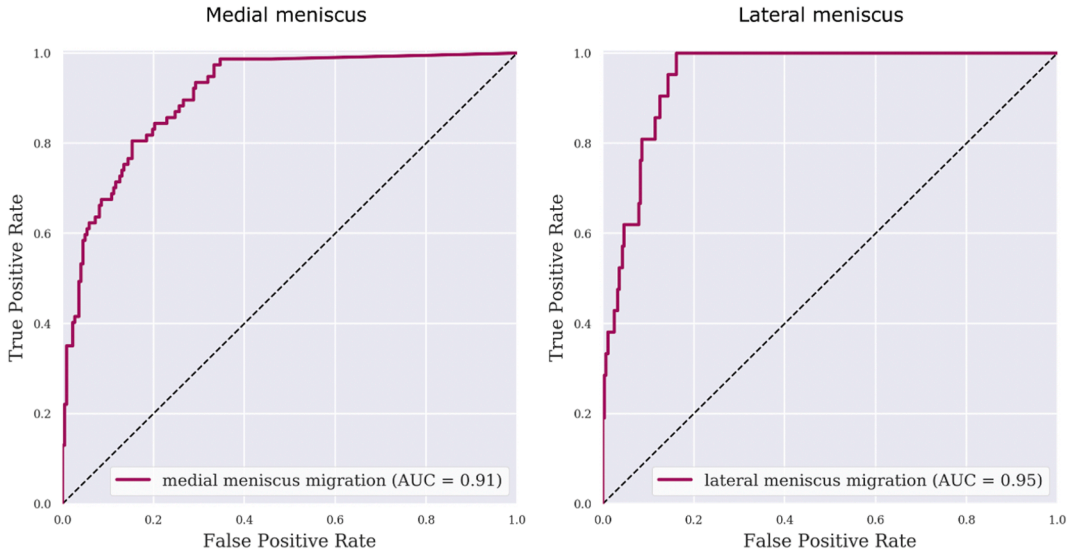

Fig. 7. Meniscal tear characterization ROC curves (left: medial meniscus, right: lateral meniscus).

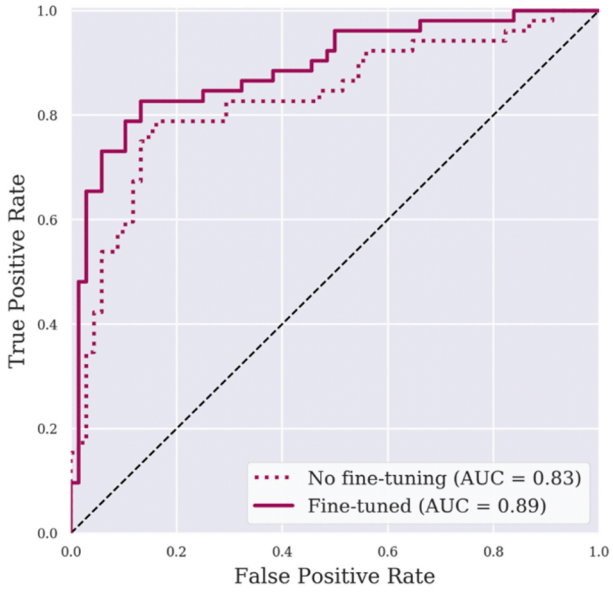

Fig. 8. Meniscal tear detection model performances on MRNet external dataset, with and without fine-tuning.

performances in diagnosing meniscal fragment migration and, as a result, supporting useful patient clinical decision.

With tremendous advances in the field of deep learning in the last decade, AI applications focusing on menisci are on the rise and shifting progressively and rapidly from automatic segmentation and computer-aided detection methods to proof-of-concept meniscal tear classifiers, with implementation of AI models in a production clinical setting as a near-future perspective.

Pedoia et al. [7] used knee MRI examinations to evaluate a binary meniscal lesion detection task and a severity score classifier using the Whole-Organ Magnetic Resonance Imaging Score (WORMS) (mild/moderate versus severe). This proof of concept fully automated deep-learning pipeline achieved a sensitivity of 81.98% and a specificity of 89.81% for meniscal lesion detection with AUC of 0.89 on test set. Ground truth was annotation by board-certified radiologists. Dataset was smaller than in our study, as they used 1478 examinations with 10 times augmentation techniques to increase their training set. Their population including only subjects at various stages of osteoarthritis and after ACL injury and reconstruction does not represent accurately clinical routine.

Teams competing in a data challenge organized by the French Radiology Society in 2018 used fast-region CNN [8] or mask-region-based CNN [9] to classify menisci between healthy and torn, and categorize orientation and location of tears used as reference standard a single annotated sagittal T2 image dataset. The two winning teams obtained AUC of 0.94 for the meniscal tear detection task [7] and a weighted AUC score of 0.906 for all three tasks [8]. However, meniscal tear detection does not rely only on a single sagittal MRI image in a real world setting and these results could not be used in clinical practice.

Bien et al. [6] developed a deep learning model for detecting general abnormalities, ACL and meniscal tears using a 1370 knee MRI dataset performed with GE scanners. Reference standards labels being majority vote of 3 MSK radiologists, their MRNet model achieved sensitivity, specificity, accuracy and AUC of respectively 0.710, 0.741, 0.725 and 0.847 for overall meniscal tear detection in an internal validation test set of 120 examinations. Algorithm specificity was lower compared to general radiologists (0.892). MRNet was validated externally for ACL tear but not for meniscal tear due to lack of available dataset. After fine tuning, our model outperforms the performance of the MRNet model on his own test database by 4.3% of AUC value.

More recently, Fritz et al. [10] used a study design flowchart and data science methodology similar to ours. Their model showed sensitivity, specificity, accuracy and AUC of respectively 84%, 88%, 86%, 88.2% medially, and 58%, 92%, 84%, 78.1% laterally. They achieved a similar specificity but lower sensitivity in comparison with MSK radiologists. They did not test the model on external data to fully validate it clinically, as described as "best practice'' in the checklist for AI in Medical Imaging published in Radiology: Artificial Intelligence [16].

Limitations of our study include meniscal tear labels extraction from radiological reports without surgical correlation, but internal validation on a subset labelled by expert MSK radiologists and external validation advocate for robustness.

Dataset imbalance may explain the inferior overall performances on lateral meniscal tear detection and characterization. A larger amount of data including lateral meniscal tear in training dataset may further increase model performances laterally.

Performances of a human reader assisted by the model was not performed, but as MSK radiologists noticed some clinically relevant lesions like meniscal root tears were sometimes overlooked by general radiologists, we are confident model assistance could lower error rate in radiological report.

Knee MRI analysis is a complex task and an AI tool solely focused on a small subset of all potential internal lesions of the knee is unsure to add value to the patient care. Therefore, in our opinion, further work needs to be done to cover broader structures analysis of knee components in a structured and standardized way before implementing efficiently these tools in clinical practice.

Further studies are also needed on deep learning algorithms interpretability to support professional confidence and efficient implementation, but active participation of radiologists in the building of these models and strong partnership with data scientists are keys to support early adoption in clinical routine.

5. Conclusions

Deep learning models can efficiently detect and characterize meniscal tears, while maintaining robustness when confronted to external data. This opens perspectives for generalization and might result in clinical applications as part of a more complex machine learning system adding value and augmenting human reading of knee MRI.

References

[1] 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(3):837–48. https://doi.org/10.1148/radiol.2263011892.

[2] Nikolaou VS, Chronopoulos E, Savvidou C, Plessas S, Giannoudis P, Efstathopoulos N, et al. MRI efficacy in diagnosing internal lesions of the knee: a retrospective analysis. J Trauma Manag Outcomes 2008;2(1). https://doi.org/10.1186/1752-2897-2-4.

[3] Crawford R, Walley G, Bridgman S, Maffulli N. Magnetic resonance imaging versus arthroscopy in the diagnosis of knee pathology, concentrating on meniscal lesions and ACL tears: A systematic review. Br Med Bull 2007;84:5–23. https://doi.org/10.1093/bmb/ldm022.

[4] Vande Berg BC, Malghem J, Poilvache P, Maldague B, Lecouvet FE. Meniscal tears with fragments displaced in notch and recesses of knee: MR imaging with arthroscopic comparison. Radiology 2005;234(3):842–50. https://doi.org/10.1148/radiol.2343031601.

[5] Garwood ER, Tai R, Joshi G, Watts VGJ. The use of artificial intelligence in the evaluation of knee pathology. Semin Musculoskelet Radiol 2020;24(1):21–9. https://doi.org/10.1055/s-0039-3400264.

[6] 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;27(15(11)):e1002699. https://doi.org/10.1371/journal.pmed.1002699.

[7] Pedoia V, Norman B, Mehany SN, Bucknor MD, Link TM, Majumdar S. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. J Magn Reson Imaging 2019;49(2):400–10. https://doi.org/10.1002/jmri.v49.210.1002/jmri.26246.

[8] Roblot V, Giret Y, Bou Antoun M, Morillot C, Chassin X, Cotten A, et al. Artificial intelligence to diagnose meniscus tears on MRI. Diagn Interv Imaging 2019;100(4):243–9. https://doi.org/10.1016/j.diii.2019.02.007.

[9] Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging 2019;100(4):235–42. https://doi.org/10.1016/j.diii.2019.03.002.

[10] Fritz B, Marbach G, Civardi F, Fucentese SF, Pfirrmann CWA. Deep convolutional neural network-based detection of meniscus tears: Comparison with radiologists and surgery as standard of reference. Skeletal Radiol 2020;49(8):1207–17. https://doi.org/10.1007/s00256-020-03410-2.

[11] Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 658–66.

[12] Nguyen JC, De Smet AA, Graf BK, Rosas HG. MR imaging-based diagnosis and classification of meniscal tears. Radiographics 2014;34(4):981–99. https://doi.org/10.1148/rg.344125202.

[13] Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling 2014;arXiv:1412.3555.

[14] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space 2013;arXiv:1301.3781.

[15] DiCiccio TJ, Efron B. Bootstrap confidence intervals. Statist Sci 1996:189–212.

[16] Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029. https://doi.org/10.1148/ryai.2020200029.