European Radiology
https://doi.org/10.1007/s00330-022-08923-z

MUSCULOSKELETAL

# Deep learning to detect anterior cruciate ligament tear on knee MRI: multi-continental external validation

Alexia Tran[1] · Louis Lassalle[2] · Pascal Zille[3] · Raphaël Guillin[4] · Etienne Pluot[5] · Chloé Adam[3] · Martin Charachon[3] · Hugues Brat[6] · Maxence Wallaert[3] · Gaspard d'Assignies[3,7] · Benoît Rizk[6]

Received: 30 October 2021 / Revised: 28 April 2022 / Accepted: 30 May 2022


## Abstract

**Objectives** To develop a deep-learning algorithm for anterior cruciate ligament (ACL) tear detection and to compare its accuracy using two external datasets.

**Methods** A database of 19,765 knee MRI scans (17,738 patients) issued from different manufacturers and magnetic fields was used to build a deep learning–based ACL tear detector. Fifteen percent showed partial or complete ACL rupture. Coronal and sagittal fat-suppressed proton density or T2-weighted sequences were used. A Natural Language Processing algorithm was used to automatically label reports associated with each MRI exam. We compared the accuracy of our model on two publicly available external datasets: MRNet, Bien et al, USA (PLoS Med 15:e1002699, 2018); and KneeMRI, Stajduhar et al, Croatia (Comput Methods Prog Biomed 140:151-164, 2017). Receptor operating characteristics (ROC) curves, area under the curve (AUC), sensitivity, specificity, and accuracy were used to evaluate our model.

**Results** Our neural networks achieved an AUC value of 0.939 for detection of ACL tears, with a sensitivity of 87% (0.875) and a specificity of 91% (0.908). After retraining our model on Bien dataset and Stajduhar dataset, our algorithm achieved AUC of 0.962 (95% CI 0.930–0.988) and 0.922 (95% CI 0.875, 0.962) respectively. Sensitivity, specificity, and accuracy were respectively 85% (95% CI 75–94%, 0.852), 89% (95% CI 82–97%, 0.894), 0.875 (95% CI 0.817–0.933) for Bien dataset, and 68% (95% CI 54–81%, 0.681), 93% (95% CI 89–97%, 0.934), and 0.870 (95% CI 0.821–0.913) for Stajduhar dataset.

**Conclusion** Our algorithm showed high performance in the detection of ACL tears with AUC on two external datasets, demonstrating its generalizability on different manufacturers and populations.

**Summary** This study shows the performance of an algorithm for detecting anterior cruciate ligament tears with an external validation on populations from countries and continents different from the study population.

**Key Points**

- *An algorithm for detecting anterior cruciate ligament ruptures was built from a large dataset of nearly 20,000 MRI with AUC values of 0.939, sensitivity of 87%, and specificity of 91%.*
- *This algorithm was tested on two external populations from different other countries: a dataset from an American population and a dataset from a Croatian population. Performance remains high on these two external validation populations (AUC of 0.962 and 0.922 respectively).*

**Keywords** Deep learning · Knee injuries · Magnetic resonance imaging · Artificial intelligence

*Alexia Tran and Louis Lassalle are joint first authors.

✉ Alexia Tran
tranalexia@yahoo.com

[1] Department of Radiology, Hôpital Européen Georges Pompidou, Assistance Publique-Hôpitaux de Paris; Université de Paris, 20 Rue Leblanc, 75015 Paris, France

[2] Réseau d'imagerie Sud Francilien, Evry, France

[3] Incepto Medical, Paris, France

[4] Department of Radiology, Centre Hospitalier Universitaire de Rennes, Rennes, France

[5] Department of Radiology, Radiologie B, Hôpital Cochin, Assistance Publique-Hôpitaux de Paris; Université de Paris, Paris, France

[6] Institut de Radiologie de Sion, Groupe 3R, Sion, Switzerland

[7] Department of Radiology, Centre Hospitalier Départemental Vendée, La Roche-sur-Yon, France

Published online: 21 June 2022

**Abbreviations**

| | |
|---|---|
| ACL | Anterior cruciate ligament |
| CNN | Convolutional neural networks |
| IoU | Intersection over Union |
| NLP | Natural Language Processing |
| PD | Proton density |
| ReLU | Rectified linear unit |
| SD | Standard deviation |

## Introduction

Anterior cruciate ligament (ACL) ruptures are frequent, accounting for more than 50% of knee injuries [1], and severe since they may lead to knee instability with the risk of progression to osteoarthritis. Magnetic resonance imaging (MRI) is the gold standard non-invasive examination and recommended first-line procedure for diagnosing ACL tears and identifying surgical candidates [2]. According to a recent meta-analysis, MRI is able to provide an appreciable diagnostic performance with a sensitivity of 87%, a specificity of 90%, and an area under the curve (AUC) of 0.93 [3]. However, this diagnostic performance may depend on radiologists' experience and field of practice [4]. Artificial intelligence (AI) has already proven its usefulness in medical imaging [5, 6], especially in musculoskeletal imaging [7]. Deep-learning methods for ACL tear detection have been developed so far either on datasets of less than 10,000 studies or without external validation and often on a single manufacturer images [8–10].

Our goal was to develop a deep-learning tool for ACL tear detection using a large dataset and to compare its accuracy using two external datasets (MRNet, USA [8]; and KneeMRI, Croatia [9]).

## Materials and methods

### Dataset

To create our dataset, we retrospectively included 19,765 knee MRI studies from 12 imaging centers, performed between 2009 and 2020, 15% (2,965) of which showed a partial or complete ACL tear. Natural Language Processing was used to automatically label reports associated with each MRI exam and determine which exam showed an ACL rupture. Our multi-centric institution has a general consent form signed by each patient to allow or refuse retrospective data analysis for research purposes. Only patients older than 16 years were included. The most frequent indications for MRI were the assessment of acute or chronic pain and trauma. The included population consisted of 17,738 patients, 1,744 (9.8%) of which had at least two magnetic resonance (MR) MRI examinations, with an average of 2.2 MR scans. Mean age was 44 years with a standard deviation (SD) of 17 years and a female/male ratio of 48% (8,514) / 52% (9,224). Of these 17,738 patients, 10,122 patients have been previously reported [10]. This prior article dealt with the development of a deep learning algorithm for the detection of meniscal tears and their characterization (presence/absence of migrated meniscal fragment) whereas in this article we study the performance of an algorithm for detecting anterior cruciate ligament tears.

Examinations were issued from several manufacturers (Philips Healthcare; GE Healthcare; and Siemens Healthcare) and different magnetic fields (1 Tesla, 1.5 Tesla, 3 Tesla). Since knee MRI examinations were acquired in clinical routine, all had at least standard sequences: coronal, axial, and sagittal (either 2D or 3D with 2D reformatted images) fat-suppressed proton density-weighted or fat-suppressed T2-weighted sequences. MRI characteristics and acquisition parameters are detailed in Table 1. For the development of our algorithm, we only used coronal and sagittal fat-suppressed proton density (PD)–weighted or fat-suppressed T2-weighted sequences.

Our dataset has been randomized into a training set (70%, 13,836 examinations) used to fit the parameters of the model, a validation set (20%, 3,953 examinations) to provide an evaluation of the fitted model and to optimize the model's hyperparameters, and a test set (10%, 1,976 examinations) to provide an evaluation of the final model. In each set, 15% of the MRIs showed an ACL tear (2,075 in the training set, 593 in the validation set and 296 in the test set). All exams and images corresponding to the same patients were in the same split (training, validation, and test). The flowchart of the dataset is presented in Fig. 1. Statistical comparisons across splits for all the demographic variables were done. We relied on $chi^2$ square tests of independence for categorical variables and one-way ANOVA for continuous ones. No statistical differences were observed across splits at 0.05 alpha level.

### Ground truth: Natural Language Processing on MRI reports

We developed a Natural Language Processing (NLP) algorithm to automatically label reports associated with each MRI exam. To do so, 2643 reports were chosen at random in the database, the large sample size ensuring that the subset is representative of our population.

These 2643 reports were manually annotated with one of the 2 possible labels "torn ACL" (746) or "normal ACL" (1897). We split these 2643 reports into 2 sets: a set of 2511 reports to perform 5-fold cross-validation and a set of 132 reports that we used as an aggregated test set. We used a bidirectional gated recurrent unit neural network fed by sentences from results and conclusion parts. Each word of the input sentences was represented by its corresponding embedding which was computed with the Word2Vec algorithm

**Table 1** Age and sex distribution of the study population, manufacturers, and MRI magnetic fields in our database and in each set. *Patient age was missing data in about 52% of the MRI examinations.** $p$ values between training set, validation set, and test set. A chi-square test was performed for categorical data (sex, ACL tear prevalence, manufacturer, magnetic field strength). A one-way ANOVA test was performed for non-categorical data (age, repetition-time, echo-time). No statistical difference was found between these sets. *** Some examinations contain several sagittal and/or coronal sequences

| | All dataset (n = 19,765) | Training set (n = 13,836) | Validation set (n = 3,953) | Test set (n = 1,976) | $p$ ** |
|---|---|---|---|---|---|
| Study population | | | | | |
| Mean age (standard deviation)* | 44 (18) | 44 (18) | 44 (18) | 43 (18) | 0.438 |
| Female | 48% | 48% | 47% | 47% | 0.199 |
| ACL tear prevalence | 15% | 15% | 15% | 15% | 0.729 |
| Distribution by manufacturer | | | | | |
| Philips | 84% | 84% | 84% | 85% | 0.092 |
| GE | 12% | 12% | 12% | 12% | |
| Siemens | 4% | 4% | 4% | 3% | |
| Distribution by magnetic field | | | | | |
| 3 Tesla | 32% | 33% | 33% | 31% | 0.380 |
| 1.5 Tesla | 17% | 17% | 16% | 17% | |
| 1 Tesla | 51% | 50% | 51% | 53% | |
| Mean repetition time (standard deviation) | | 2779 msec (929 msec) | 2756 msec (858 msec) | 2785 msec (853 msec) | 0.317 |
| Mean echo time (standard deviation) | | 28 msec (11 msec) | 28 msec (10 msec) | 28 msec (10 msec) | 0.404 |
| Number of PD-weighted or T2-weighted sequences in each set*** | | | | | |
| Number of coronal PD-weighted or T2-weighted sequence | 21,763 | 15,203 | 4,397 | 2,163 | |
| Number of sagittal PD-weighted or T2-weighted sequence | 24,489 | 17,078 | 4,999 | 2,412 | |

[11, 12]. We selected the parameters (hidden layer size, word embedding size) providing the best average area under the receiver operating characteristic curve (AUC) on the 5 test splits. To assign a label to the network predictions (normal ACL or torn ACL), we concatenated all the predictions obtained on the 5 test folds and selected the threshold maximizing the F1 Score.

To obtain a final model, we trained a bidirectional gated recurrent unit neural network with the best parameters on the 2643 reports we used for cross-validation, using only 10% of the reports as a validation set.

## Deep convolutional neural network

Our deep-learning model consisted of two parts: a first model to locate the ACL and a second to classify ACL as normal or torn. A schematic illustration of our model can be seen in Fig. 2.

### Metal artifact detection

We developed a deep convolutional neural network detecting metallic artifacts on sequences in proton density with fat saturation or in T2 with fat saturation, which allowed us to

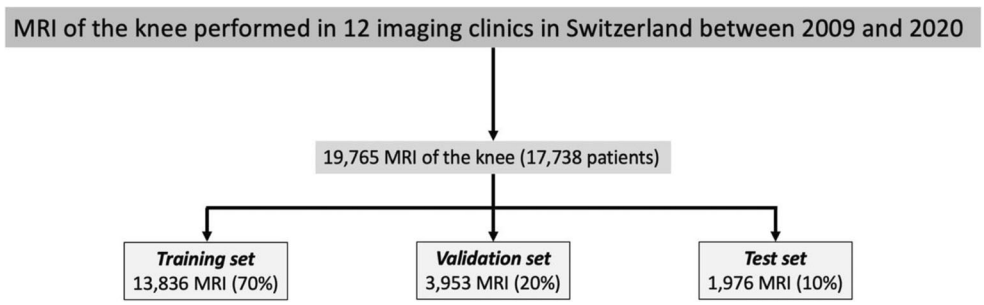


**Fig. 1** Flow chart. We retrospectively included 19,765 knee MRI studies (17,738 patients) from 12 imaging centers performed between 2009 and 2020, 15% (2,965) of which shows a partial or complete ACL tear. The dataset has been randomized into a training set (70%, 13,836 examinations) used to fit the parameters of the model, a validation set (20%, 3,953 examinations) to provide an evaluation of the fitted model, and to optimize the model's hyperparameters and a test set (10%, 1,976 examinations) to provide an evaluation of the final model. In each set, 15% of the MRIs showed an ACL tear (2,075 in the training set, 593 in the validation set, and 296 in the test set)

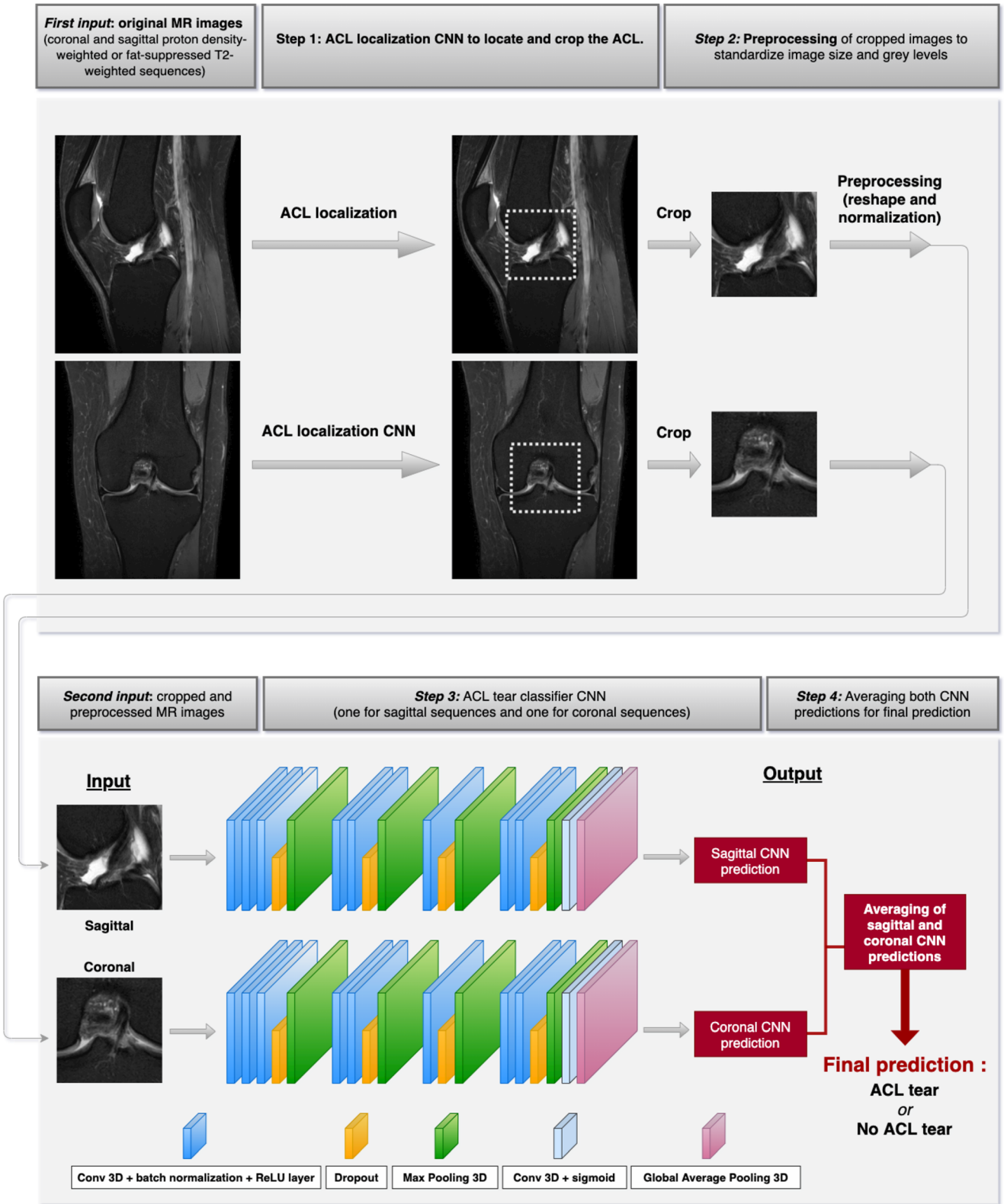


**Fig. 2** Illustration of our model. The first models used for ACL localization were built by linear regression from a meniscus localization CNN that we had previously built. These models were connected in a cascaded fashion to an ACL tear classifier which was composed of two CNNs, one for sagittal view and one for coronal view. The architecture of these CNNs is described in the figure. The final prediction was the average of the two CNNs

exclude examinations with metallic artifacts. We used the same database of 19,765 knee MRI scans. From this database, we selected exams containing words meaning that the patient had been operated on, with a keyword search in the report database. We found 1,200 exams. After looking at each MRI one by one, we labeled these exams according to whether they contained metallic artifacts. We then had a database of 1025 MRIs containing metal artifacts, to which we added 3000 examinations without metal artifacts. This database was divided into training set (70%), validation set (20%), and test set (10%). We then trained a fully convolutional network that we optimized through several hyperparameter tuning steps. The final AUC of our algorithm was 0.985. With this metal artifact detection model, all exams that contained metal artifact were excluded.

### ACL localization

Two coronal and sagittal localization models were used to extract bounding box coordinates around ACL. These models were built by linear regression from a meniscus localization CNN that we had previously built [10]. The performance of the models was evaluated on a test set of 197 examinations annotated by a data scientist after training and under the supervision of a radiologist. Intersection over Union (IoU) evaluation metric was used to measure the accuracy of our ACL localizer model.

ACL crops produced by the localization models were resized to a common size of 64 × 64 × 64 across volumes and paired with the NLP found labels ("torn ACL" or "normal ACL"). They were used as input for the ACL tear classifier.

### ACL classification

The ACL tear classifier was made of two convolutional neural networks (CNN): one for the sagittal plane and one for the coronal plane. Both CNNs were composed of five convolution blocks. For the first four convolution blocks, each convolution layer was activated by a rectified linear unit (ReLU) function and followed by a batch normalization step. There was a layer of MaxPooling 3D between each block. For the final convolution block, the convolution layer was activated by a sigmoid function and followed by a Global Average Pooling 3D layer. To have only one prediction by knee MRI exam, we averaged the output of the two CNNs. Keras deep learning library (keras.io) and a TensorFlow backend (www.tensorflow.org) were used to develop our CNNs. Training was performed on a NVIDIA V100 graphic processing unit.

### Training of the CNN

The two CNNs were trained with our training set of 13,864 examinations, all containing sagittal and coronal sequences. This training set allowed us to compute the weights of the CNN. Then, we used the validation set to tune the hyperparameters. We finally tested our model accuracy using the test set.

### Heatmap generation

Heatmaps were generated using an in-house method [13]. This method had two different image generators, a similar and adversarial. The studied model classifies the input image with a certain label, healthy or pathological. On the one hand, the similar generator will produce an image close to the original image that will be classified by the studied model with the same label. On the other hand, the adversarial generator will produce an image that will be classified with the opposite label. The heatmap is then defined as the absolute difference between the two generated images. The main advantage of this method is that the generators are built so that the difference between generated images only captures meaningful information for the classification model while producing images close to the training database. Heatmaps are much less corrupted by noise or non-meaningful features.

## Double external validation

To evaluate the generalizability of our deep learning model and compare its performance with pre-existing models, we tested it on two external validation datasets: KneeMRI and MRNet.

KneeMRI is a publicly available dataset collected by Stajduhar et al [9]. It contains 917 sagittal PD-weighted examinations acquired with a Siemens Avanto 1.5-T scanner at Clinical Hospital Centre Rijeka, Croatia, from 2007 until 2014. All these MRI examinations were labeled based on both report and additional reading by a radiologist. The dataset consists of 690 non-injured ACL (≈ 75%), 172 partially injured (≈ 20%) and 55 completely ruptured (≈ 5%) cases. Two models were built in the publication: one model capable of detecting injured ACL cases (partially and completely ruptured), differentiating them from normal (healthy) cases, and one model capable of automatically detecting completely ruptured cases. Their models had an area under the curve of 0.894 for the injury-detection problem (partially and completely ruptured ACL) and 0.943 for the complete-rupture-detection problem. Like the authors, we merged partially injured and completely ruptured ACL into one category: ACL tear. We randomly split the 917 examinations into three sets: 60% of the examinations were in the training set, 20% of the examinations were in the validation set, and 20% of the examinations were in the test set. The prevalence of ACL tears was the same in all three sets. We decided to adopt a "60/20/20" split to follow Bien et al external validation experiments setup on KneeMRI dataset and harmonize subsequent comparisons as

much as possible. We first tested our model on the KneeMRI test set without re-training, then we did retraining and fine-tuning schemes for adapting our deep learning model to KneeMRI training and validation set. This procedure is used to obtain increased performances on new datasets. Starting from the original model weights (obtained using our internal dataset), we ran an additional training experiment using the data from the KneeMRI dataset. This training experiment was run using the same parameters as the one used for the internal dataset, apart from the learning rate whose value was set to one-tenth of the original one. Using a lower learning rate is common practice when running fine-tuning experiments, in order to provide more control over the speed at which the model's weights adapt to the new dataset. The weights corresponding to the lowest loss value on the validation split of the KneeMRI dataset were retained for final evaluation on the test split. The objective was to classify the examinations as ruptured or unruptured ACL. In the rest of the article, this dataset will be referred to as the Stajduhar dataset.

We used another publicly available dataset of 1,370 examinations collected at Stanford University Medical Center between 2001 and 2012 by Bien et al [8]. Each examination contained the following sequences: coronal T1-weighted, coronal T2 with fat saturation, sagittal proton density (PD)–weighted, sagittal T2 with fat saturation, and axial PD-weighted with fat saturation. Examinations were performed using GE scanners, 775 (56%) with a 3-T magnetic field, the remainder with a 1.5-T magnetic field. MRIs were labeled using radiological reports. Bien et al provide the training set and the validation set of MRNet in free access on its website. However, the test set was not freely available on the site. So, we used the MRNet validation set as a test set and split the training set into a training set and a validation set. These splits were comparable to each other for age and sex as well as prevalence of ACL tears. Bien et al developed MRNet, a convolutional neural network for classifying MRI series and combined predictions from 3 series per exam using logistic regression. In detecting abnormalities, ACL tears, and meniscal tears, this model achieved area under the receiver operating characteristic curve (AUC) values of 0.937 (95% CI 0.895, 0.980), 0.965 (95% CI 0.938, 0.993), and 0.847 (95% CI 0.780, 0.914), respectively, on the internal validation set [8]. This dataset will be further referred to as the Bien dataset.

### Statistics

Statistical analysis was performed by using MATLAB (version 2013a; MathWorks) and MedCalc (version 14.8; MedCalc Software). We evaluated the performance of our deep learning model by measuring sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC). 95% confidence intervals were calculated for sensitivity, specificity, accuracy, and AUC. These confidence intervals were calculated using the bootstrap method with replacement [14]. Once the model training process is completed, successive random draws of prediction values from the original samples are used to compute a resampled distribution of the model's predictions. Quantized values of that resampled distribution for a given $\alpha$ 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. A threshold was determined using the point on the ROC curve that optimized the Youden index. We calculated the AUC before and after the exclusion of examinations containing metallic artifacts in the test set of our database.

## Results

### Model performance

The NLP model achieved an AUC of 0.984 (95% CI: 0.946–1.0) and the ACL localizer an IoU of 0.72 (95%CI: 0.70–0.73) in our database.

The ACL classifier CNN had an AUC of 0.939 (95%CI: 0.918–0.956) in the hold-out test set of 1,971 examinations for ACL tear detection. After excluding examinations containing metallic artifacts in the test set ($N = 103$), we obtained an AUC of 0.941(95%CI: 0.922–0.959).

The threshold that optimized Youden's index was 0.38. Using this threshold, our model achieved a sensitivity of 87% (95% CI 84–92%, 0.875) and a specificity of 91% (95% CI 89–92%, 0.908). Its accuracy was 0.902 (95% CI 0.889–0.914).

Heatmaps were generated to better understand which areas of the image were the most discriminating for our model. We applied a threshold to the heatmap in order to keep only the values above the 99th percentile. Examples of resulting heat maps are shown in Fig. 3.

### External validation

Without retraining, our model obtained an AUC of 0.941 (95% CI: 0.897–0.978) on the Bien dataset and 0.860 (95%CI: 0.829–0.892) on the Stajduhar dataset. After retraining our model on the Bien dataset, external validation on their test set resulted in an AUC of 0.962 (95% CI 0.930–0.988). By choosing the threshold that optimized Youden's index, which was 0.37, sensitivity was 85% (95% CI 75–94%, 0.852), and specificity was 89% (95% CI 82–97%, 0.894). The accuracy was 0.875 (95% CI: 0.817–0.933). After retraining our model on the Stajduhar dataset, external validation on the Stajduhar test set resulted in an AUC of 0.922 (95% CI 0.875, 0.962). Choosing a threshold of 0.60 which optimized

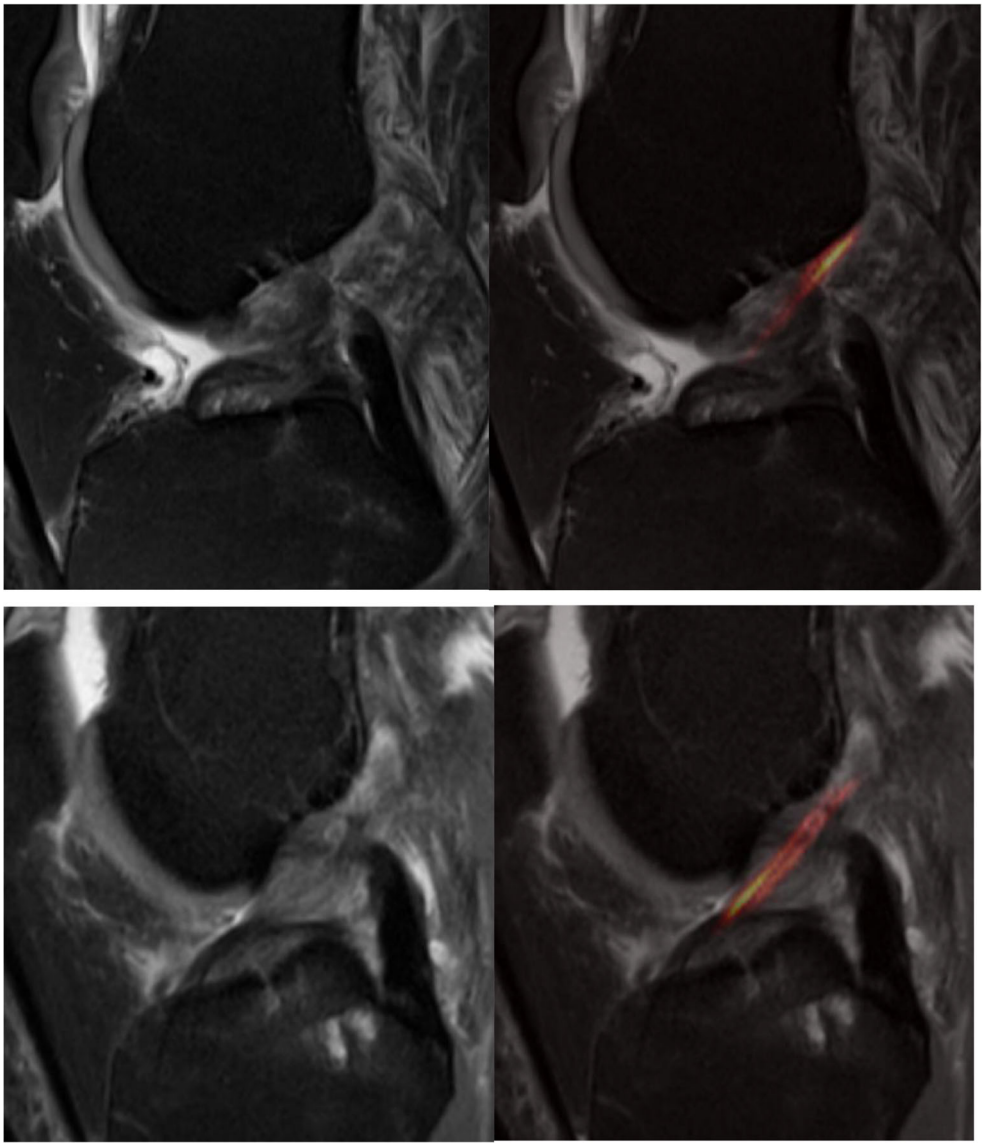

**Fig. 3.** Examples of torn ACL detected by the algorithm with corresponding Heatmaps. Cropped sagittal proton density–weighted MR image showing two ACL complete tear. Heatmap showing the high-probability areas in the ACL on which our model based its interpretation of an ACL tear

Youden's index, sensitivity was 68% (95% CI 54–81%, 0.681) and specificity was 93% (95% CI 89–97%, 0.934). The accuracy was 0.870 (95% CI: 0.821–0.913) (Fig. 4).

Table 2 compares our performance with that of Bien et al and Stajduhar et al.

## Discussion

Our study demonstrates the ability of a deep learning model to accurately detect an ACL tear on MRI, both on an internal and on two external validation sets. To our knowledge, this is the first study with such a size of training dataset and with two external validations. Our model was competitive compared to similar models from the literature. Although our AUCs were calculated on different sets, the population and prevalence of the two sets are comparable.

Liu et al [15] developed a deep learning–based diagnosis system to detect ACL tears with a sensitivity and specificity of 0.96 and 0.96 respectively (AUC 0.98). Our model had a sensitivity and specificity of 0.87 and 0.91 respectively (AUC 0.939). Liu et al's dataset was a smaller dataset of 350 MRIs that were all performed on the same 3.0-T imaging unit using the same imaging protocol, whereas our algorithm was trained using a large 19,765 knee MRIs, multicentric (12 centers) dataset of patients examined by several different MRI scanners with different magnetic field strengths (1/1.5/3 T) and different MRI protocols. The variety of our training set may explain the performances on external datasets, which further improved when retraining the model specifically for those datasets.

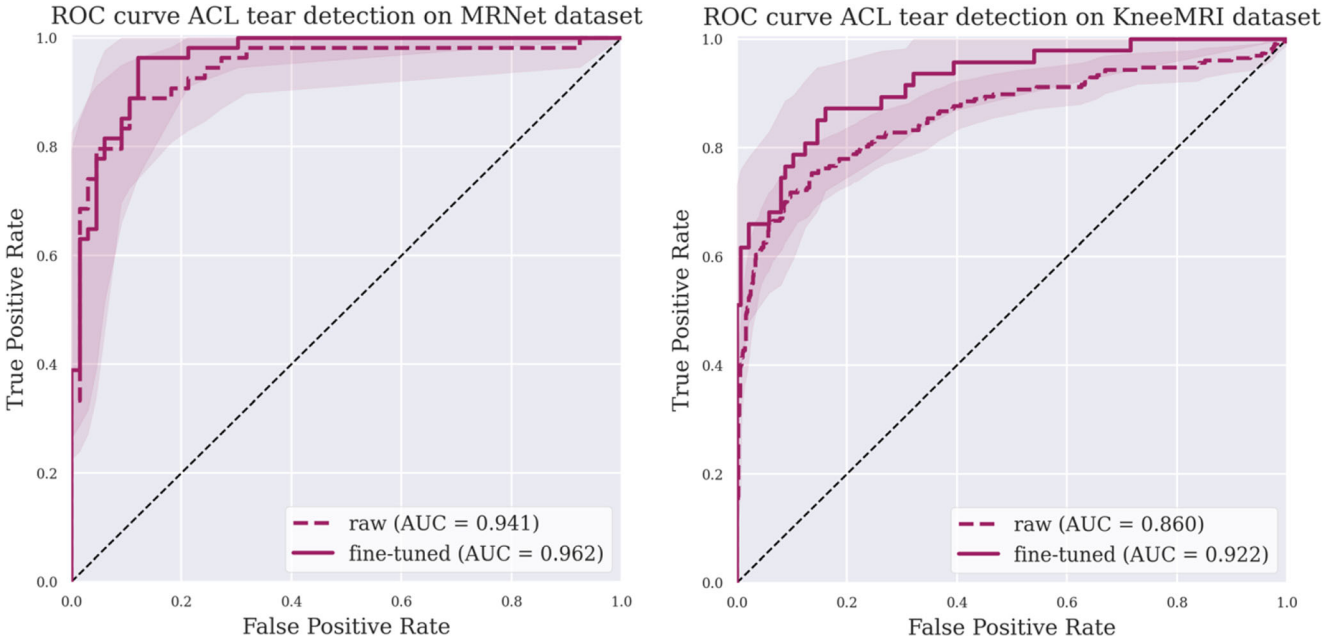


**Fig. 4** AUC before and after fine tuning our model on external datasets. The area in light purple corresponds to confidence intervals

Bien et al [8] developed a deep learning model (MRNet) for detecting general abnormalities and specific diagnoses (ACL and meniscal tears) on MRI examinations. Their model performance for ACL tear detection achieved a sensitivity of 0.759, a specificity of 0.968, and an AUC of 0.965 (95% CI 0.938, 0.993). Their most beneficial series was the coronal T1-weighted sequence which may seem surprising since that sequence is rarely used in clinical practice to detect ACL rupture. They also had a smaller dataset (1,370 examinations) and all examinations were performed using the same manufacturer (GE).

The Stajduhar dataset was developed for semi-automated detection of ACL injury on MRI [9]. Using manually selected regions of interest, support vector machine, and random forests model, they obtained an AUC of 0.894 for the ACL injury-detection model and 0.943 for complete ACL rupture diagnosis on their dataset. We could only validate externally on the sagittal plane available in the Stajduhar dataset and it prevented pooling predictions across acquisition planes which could boost the model's performance.

Recently, several teams have developed algorithms for the detection of ACL lesions with high performance. Germann et al [16] built a deep convolutional neural network for the detection of surgically proven ACL tears. The DCNN had a sensitivity of 0.96 and a specificity of 0.93, and an AUC of 0.935. Zhang et al [17] evaluated the diagnostic performance of their deep learning approach for ACL lesion detection using arthroscopy as the gold standard. The sensitivity and specificity of their deep learning architecture were 0.957 and 0.976 respectively (AUC 0.960). This interest in building an automatic classification tool for ACL injuries shows both its feasibility and usefulness.

The use of NLP had an important place in our annotation strategy. Our NLP model had a high AUC of 0.98 showing that data mining from radiological reports could lower radiologists' annotation costs in the field of medical imaging AI. This is consistent with a proof of concept recently published by Pinto dos Santos et al [18]. Several deep learning models detecting ACL tears have been previously published with high accuracies but with smaller datasets and sometimes without external validation.

The area of explainability of deep learning models has not yet produced stable and definitive methods. In our work, we chose to use an algorithm recently introduced in [13] to

**Table 2** Comparison of the performance of our model, Bien et al's model (MRNet) and Stajduhar et al's model

| | Our model's AUC | Bien et al's AUC | Stajduhar et al's AUC |
|---|---|---|---|
| Our test set* | 0.941 | | |
| Bien et al dataset | 0.962 on validation set** | 0.965 on test set | |
| Stajduhar et al dataset | 0.922 | 0.911 | 0.894 |

*on examinations that do not contain metallic artifacts

**the test set we used is not the same as the test set of Bien et al. We used Bien et al's validation set

generate heatmaps, which not only relies on the model itself but also makes use of the training database to enforce coherent results. The main advantage of this method is that the generators are built so that the difference between generated images only captures meaningful information for the classification model while producing images close to the training database. In particular, heatmaps are much less corrupted by noise or non-meaningful features.

Our study has limitations. First, ACL labels were extracted from radiological reports without surgical correlation and reports can sometimes be inaccurate. However, arthroscopy is the gold standard with the limitations of potential verification bias. Furthermore, our internal and external validations showed the robustness of our AI models, using a 3D architecture and a localization algorithm to narrow the FOV and allow the classifier to focus on ACL, in line with the results of Chang et al [19]. Second, we did not distinguish partial-thickness from full-thickness ACL tears, which may be difficult for the algorithm as it implies the detection of subtle nuances of contour detection and signal alterations. Distinguishing partial and full ACL tears is also a difficult issue for radiologists reading MRI, even at 3T [20]. Such a difficulty is due to the ACL anatomy and its lesion mechanism, and to the information extent required for appropriate clinical decisions. In such a scenario, any static imaging method faces serious limitations, in spite of MR field strength. Lastly, no comparison was made with the performances of radiologist alone or assisted with the algorithm. Our algorithm had a sensitivity of 87% (95% CI 84–91%) and specificity of 91% (95% CI 89–92%). A recent review of radiologists' performances for diagnosing ACL tears using arthroscopy as the gold standard showed a sensitivity of 87% (95 % CI 77–94 %) and a specificity of 93 % (95 % CI 91–96 %) [21]. Although it is not possible to make a direct comparison because of the different gold standards, the performance of our algorithm seems interesting. Long-term clinical benefits of AI-model to detect ACL tear are still to prove, but in our opinion, radiologists and also clinicians could improve their diagnostic performances when assisted by the algorithm, especially for radiologists with less experience in musculoskeletal imaging. Bien et al [8] proved that providing model predictions significantly increased clinical experts' specificity in identifying ACL tears. It may reduce subjectivity, variability, and errors due to distraction and fatigue.

To conclude, high detection accuracy for ACL tear was achieved over large sample size. Moreover, under similar experimental conditions, the model achieved competitive performances compared to similar models from the literature, demonstrating its robustness. These experimental results suggest the potential for clinical application of deep learning–based approach to assist radiologists when diagnosing knee injuries. Further studies are needed to evaluate the value of such a tool for clinical practice.

**Funding** The authors state that this work has not received any funding.

## Declarations

**Guarantor** The scientific guarantor of this publication is Pascal Zille.

**Conflict of interest** The authors of this manuscript declare relationships with the following companies:

- A.T. was an intern at Incepto Medical, but at the time of submission of the article disclosed no relevant relationships.
- P.Z., C.A., M.C., and M.W. are employed by Incepto Medical.
- G.A. is the founder and Chief Medical Officer of Incepto Medical.
- R.G. disclosed no conflict of interest.
- H.B. disclosed no conflict of interest.
- B.R. disclosed no conflict of interest.
- L.L. has a consulting activity for Incepto Medical.

**Statistics and biometry** No complex statistical methods were necessary for this paper.

**Informed consent** Written informed consent was obtained from all subjects (patients) in this study.

**Ethical approval** Institutional Review Board approval was obtained.

**Study subjects or cohorts overlap** Some study subjects have been previously reported in Rizk B, Brat H, Zille P, Guillin R, Pouchy C, Adam C, et al Meniscal lesion detection and characterization in adult knee MRI: a deep learning model approach with external validation. *Physica Medica*. 2021 Mar;83:64–71.

**Methodology**

- retrospective
- diagnostic study
- multicenter study

## References

1. Musahl V, Karlsson J (2019) Anterior cruciate ligament tear. N Engl J Med 380:2341–2348
2. Shea K, Carey J (2015) Management of anterior cruciate ligament injuries. J Am Acad Orthop Surg 23:e1–e5
3. Li K, Du J, Huang L, Ni L, Liu T, Yang H (2017) The diagnostic accuracy of magnetic resonance imaging for anterior cruciate ligament injury in comparison to arthroscopy: a meta-analysis. Sci Rep 7:7583
4. Challen J, Tang Y, Hazratwala K, Stuckey S (2007) Accuracy of MRI diagnosis of internal derangement of the knee in a non-specialized tertiary level referral teaching hospital. Australas Radiol 51:426–431
5. Singh R, Kalra M, Nitiwarangkul C et al (2018) Deep learning in chest radiography: detection of findings and presence of change. PLoS One 13:e0204155

6. Chilamkurthy S, Ghosh R, Tanamala S et al (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392:2388–2396
7. Gyftopoulos S, Lin D, Knoll F, Doshi A, Rodrigues T, Recht M (2019) Artificial intelligence in musculoskeletal imaging: current status and future directions. AJR Am J Roentgenol 213:506–513
8. Bien N, Rajpurkar P, Ball R et al (2018) Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 15:e1002699
9. Štajduhar I, Mamula M, Miletić D, Ünal G (2017) Semi-automated detection of anterior cruciate ligament injury from MRI. Comput Methods Prog Biomed 140:151–164
10. Rizk B, Brat H, Zille P, Guillin R, Pouchy C (2021) Adam C, et al Meniscal lesion detection and characterization in adult knee MRI: A deep learning model approach with external validation. Phys Med 83:64–71
11. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781. Available from: http://arxiv.org/abs/1301.3781
12. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:14123555. Available from: http://arxiv.org/abs/1412.3555
13. Beaufils P, Hulet C, Dhénain M, Nizard R, Nourissat G, Pujol N (2009) Clinical practice guidelines for the management of meniscal lesions and isolated lesions of the anterior cruciate ligament of the knee in adults. Orthop Traumatol Surg Res 95(6):437–442
14. DiCiccio T, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11:189–228
15. Liu F, Guan B, Zhou Z et al (2019) Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. Radiol Artif Intell 1:180091
16. Germann C, Marbach G, Civardi F et al (2020) Deep convolutional neural network-based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. Invest Radiol 55:499–506
17. Zhang L, Li M, Zhou Y, Lu G, Zhou Q (2020) Deep learning approach for anterior cruciate ligament lesion detection: evaluation of diagnostic performance using arthroscopy as the reference standard. J Magn Reson Imaging 52:1745–1752
18. Pinto dos Santos D, Brodehl S, Baeßler B et al (2019) Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. Insights Imaging 10:93
19. Chang P, Wong T, Rasiej M (2019) Deep learning for detection of complete anterior cruciate ligament tear. J Digit Imaging 32:980–986
20. Van Dyck P, Vanhoenacker F, Gielen J et al (2010) Three tesla magnetic resonance imaging of the anterior cruciate ligament of the knee: can we differentiate complete from partial tears? Skelet Radiol 40:701–707
21. Phelan N, Rowland P, Galvin R, O'Byrne J (2015) A systematic review and meta-analysis of the diagnostic accuracy of MRI for suspected ACL and meniscal tears of the knee. Knee Surg Sports Traumatol Arthrosc 24:1525–1539