“MRNet paper summary”

This is a summary exploring the major highlights of the paper titled “*Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet*” published in October 23, 2018. The paper is based on the work of: Nicholas Bien , Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng and Matthew P. Lungren.

Motivation:

The paper explores the implementation and outcomes of a deep learning model (MRNet) that given the knee MRI of a patient, diagnoses the patient identifying whether his knee has any abnormalities. It further specifies the diagnoses to detect ACL tears and meniscal tears. This studies the impact of machine learning in the medical field, “*We wanted to determine whether a deep learning model could improve the diagnostic accuracy, specificity, or sensitivity of clinical experts, including general radiologists and orthopedic surgeons.*”(A quote from the paper).

Data set used:

The data set is 1370 MRI exams manually labeled to have 1104 abnormal exams of which 319 are ACL tears and 508 are meniscal tears. Both ACL tears and meniscal tears occurred concurrently in 194 exams total. Three MRI series were extracted for use in this study, which are sagittal plane T2-weighted series, coronal plane T1-weighted series and axial plane PD-weighted series. Each MRI exam had a total of s slices in each series where the series may have different number of slices (MRIs from different perspectives).

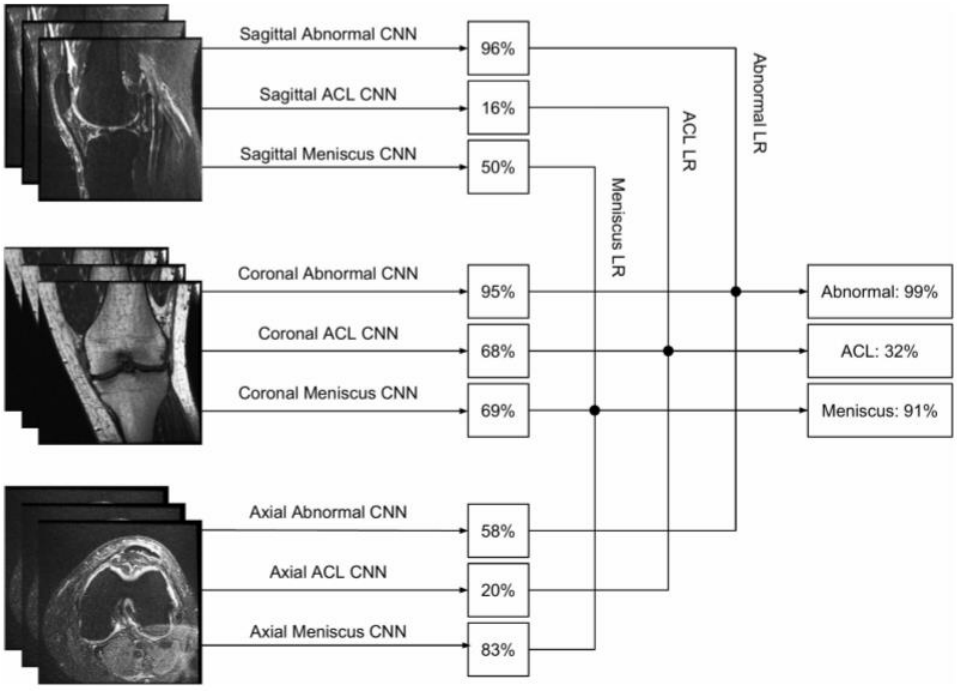
Splitting the data set:

The data set was split as follows:

- Training set: 1130 exams.
- Tuning set for model validation: 120 exams.
- Validation set for model testing: 120 exams.

Model overview:

Here we show a quick overview for the model used then it is further elaborated in other sections. The model consists of three main parts, first of which is 9 feature extractors (CNNs) that extract features of a certain MRI series for a certain diagnosis (as axial-ACL for example). The second part is a binary classifier that given the extracted features, classifies the input as either having the diagnosis or not with a number between zero and one (probability). The final part is a logistic regression that takes the probability output of each series and gives a single output for each exam corresponding to each diagnosis. *This image – taken from the paper – shows the high-level view of the model:*



Building block:

- The main building block is a CNN mapping a three dimensional MRI series of s slices (s x 3 x 256 x 256) to a probability showing a certain diagnosis for this series.
- The used feature extractor was based on AlexNet, obtaining an sx256x7x7 tensor.
- That tensor is passed to a global average-pooling layer to obtain an sx256 matrix that is passed to a max-pooling layer to obtain a 256 dimensional vector to get the best representative slice to extract features.
- That vector is then passed to a fully connected layer and a sigmoid activation function to get the probability for this certain diagnosis.
- The loss function used in optimization and backpropagation was binary cross-entropy loss (as this block is a binary classifier).

CNN block tuning:

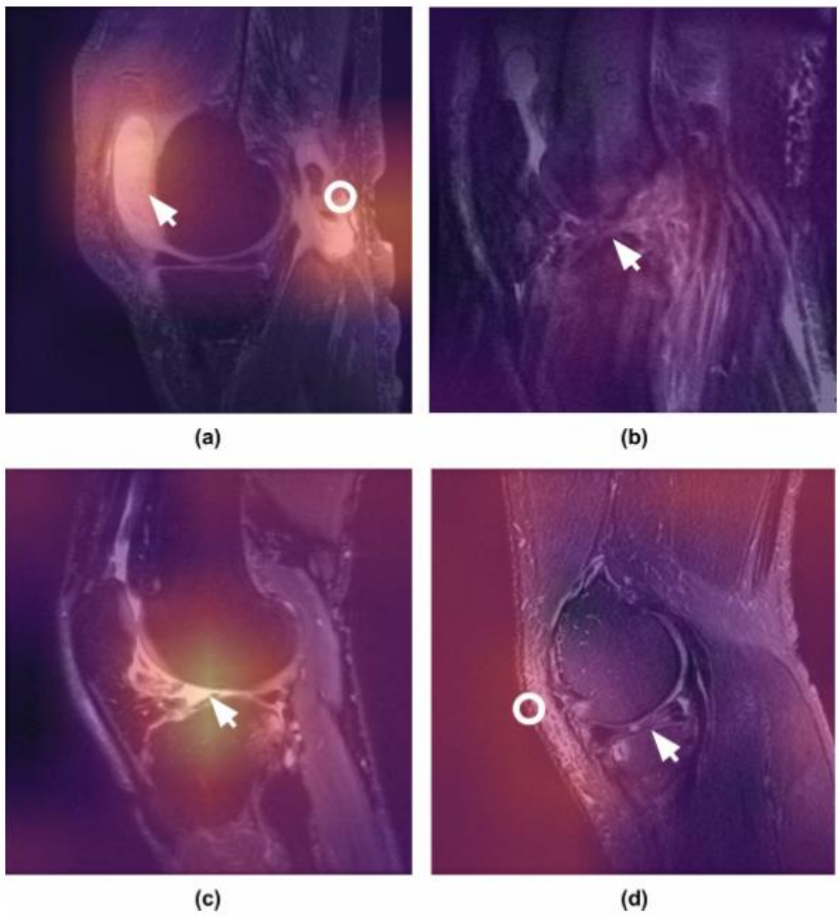
- Data augmentation on the input was used as follows:
    - Random rotation between -25 and 25 degrees.
    - Random shift between -25 and 25 pixels.
    - Horizontal flip with 50% probability.
- Transfer learning was used by initializing the weights to the values optimized on ImageNet data.

Combining the predictions:

- A logistic regression was trained to weight the prediction of the three series and give a single output.
- The idea is that a logistic regression would identify the most beneficial series for a certain diagnosis and give its prediction more weight than other series.
- Three logistic regression models were required as the output contains three predictions.

Localizing the symptoms:

This part in the paper is in fact very interesting as it may be answering the question “Do machine learning algorithms really understand?” according to the findings shown in the paper the answer may tend more to be yes. A class activation mapping (CAM) was generated for each image input after feature extraction to identify the parts in the image having the highest effect on the model’s prediction. This was done by mapping the CAM to a 256x256-color scheme and overlaying it on the original image, where the brightest areas show the most influential regions on the model predictions. A result that could be found astonishing is that these regions corresponded to actual dysfunctional regions in the patient’s knee as ruptures and effusions. An interesting note is that the model was not explicitly trained to localize the symptoms but only to predict and classify certain diagnosis, the model was able to specify these abnormalities based only on contrast to regular knee MRIs given in the training input. *This image shows the said localizations:*



Model evaluation:

- The first evaluation was on the internal validation data set. It was labeled by majority voting of three board-certified MSK radiologists as reference standard labels. Then 7 general radiologists and 2 orthopedic surgeons were divided to two groups, group 1 and group 2. Each of the two groups labeled the validation set with and without the assistance of the model with a washout period of 10 days between each review.
- The second evaluation was on an external dataset from Stajduhar et al. containing only sagittal exams (917 total) labeled for ACL injuries. Only the MRNet block for sagittal-ACL was required for evaluation in this case. First, MRNet was applied without retraining on the external data set, then it was optimized using the training and validation (tuning) sets.

Model results and clinical implications:

- Statistical analysis showed that model assistance made clinical experts more accurate in diagnoses especially considering the **specificity in identifying ACL tears**, which was stated in several parts in the paper as one of the main results of the model. Other results were not as significant statistically include increased accuracy in ACL detection and increased sensitivity in meniscal tear detection.
- As claimed by the authors, this study was the first of its kind to explore providing outputs of deep learning models in assisting radiologists and non-radiologists in task of image interpretation.
- The model was able to identify abnormalities that were not present in the training data set and correctly classify the input as abnormal. However, it is stated in the paper that more work is needed to see whether subtler abnormalities need specific training data.
- The model also had great accuracy when tested using the external data set where after optimization by training the model on the data set, it provided higher accuracy than the best model Stajduhar et al. recorded. It also took only 30 minutes to train the model and 2 minutes to evaluate it on the dataset.

Model statistics:

- In detecting abnormalities, ACL tears and meniscal tears, the model achieved AUC of 0.937, 0.965 and 0.847 respectively.
- On the external dataset in detection of ACL tears the model achieved AUC of 0.824 without additional training and 0.911 after training.
- The inter-rater agreement measured by Fleiss kappa score on the internal validation (testing) set among the three MSK radiologists was 0.508 for abnormality detection, 0.8 for ACL tears detection and 0.745 for meniscal tears detection.
- The model achieved specificity for abnormality detection of 0.714 lower than that of general radiologists, which is about 0.844. It achieved a sensitivity of 0.879 and accuracy of 0.85 while general radiologists achieved 0.905 and 0.894 respectively.
- The model achieved specificity in detection of ACL tears of 0.968 while general radiologists achieved 0.933, but radiologists achieved significantly higher sensitivity in detection of ACL tears at 0.906 in contrast to 0.759 achieved by the model.
- Radiologists achieved significantly higher specificity in detecting meniscal tears at a specificity of 0.892 opposed to 0.741 achieved by the model
- When clinical experts where provided model assistance, there was a significant increase in the expert's specificity in identifying ACL tears at 4.8%, as the validation set contained 62 exams, this meant that potentially 3 fewer patients sent to surgery for suspected ACL tear.

For further information about statistical methods used, refer to the paper's "Statistical methods" section.

Conclusion:

The results shown in the paper clearly highlight the effect of deep learning on achieving a high performance in clinical classification, and the potential of rolling out false identifications preventing patients from having unnecessary surgeries. The study also showed how deep learning models as this clearly localizes symptoms and was able enough to achieve "*state of the art"* results on an external data set. The model still needs to be tested if it could be generalized to an external dataset with more MRI series and similar MRI protocol.

Abbreviations used:

- ACL: anterior cruciate ligament.
- AUC: area under the receiver operating characteristic curve.
- CAM: class activation mapping.
- CNN: convolutional neural network.
- PD: proton density.
- MRI: magnetic resonance imaging.