# Evaluating the impact of reduced feature space on deep-learning-assisted diagnosis for knee magnetic resonance imaging.

**Katelyn K. Bechler, BS[1], Elin Lovisa Byman, BS[2], Shreya Narayan, BS[2], Asha Thomas, BS[1]**
**[1]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA; [2]Department of Electrical Engineering, Stanford University, Stanford, CA**

## Introduction

The anterior cruciate ligament (ACL) is the most commonly injured ligament in the human body, affecting more than 200,000 individuals in the United States alone.[3, 4] The ACL is located on the knee joint and connects the end of the thigh bone (femur) to the top of the shin bone (tibia). ACL tears are severe injuries that are typically sports-related, although they present in non-athletes and the broader human population as well. Early diagnosis and treatment is crucial for preventing long term osteoarthritis.[4] The current "gold-standard" for ACL tear diagnosis is arthroscopy, a surgical procedure, but because of its invasive nature, Magnetic Resonance Imaging (MRI) is preferentially used for ACL tear visualization and diagnosis. MRI-based diagnosis has a high negative predictive value and has been shown to have "high accuracy and good consistency with arthroscopic diagnosis".[5] However, MRI is also a time-intensive process that is subject to diagnostic error and variability. As the experience of the clinicians labeling and identifying the image is a key component to the image interpretation, we lack a standardized and consistent approach for labeling these images and possible ACL tears.

Using artificial intelligence (AI) to classify medical images has been a longstanding area of research in the medical community. Due to the inevitable variability in human classification, we recognize the usefulness of a standardized and accurate approach to labeling diagnostic images. Recent advances in machine learning, and specifically deep learning, have been at the forefront of novel models used to label and classify images. This is largely due to the way deep learning exploits "hierarchical feature representations learned solely from data, instead of handcrafted features mostly designed based on domain-specific knowledge".[6] As utilizing deep learning approaches to analyze multi-image series has become increasingly popular, there exists an opportunity to use these approaches in order to prioritize high-risk patients and assist in clinical diagnoses. Per a systematic review of AI methods to detect ACL tears in medical imaging, three categories of methods have been studied: 1) machine learning, 2) deep learning with transfer learning, and 3) custom-made deep-learning networks. The machine learning models used algorithms such as support vector machines, random forests, and K-nearest neighbors. Though some had a high AUC (0.9 or above), there were others that found no statistically significant differences when comparing the diagnostic capability between the AI model and radiologists. Of the two models cited using deep learning with transfer learning, Bien et al. (2018) found that radiologists achieved significantly higher sensitivities for ACL tears than the AI model, while Azcona et al. (2020) developed a model using transfer learning combined with a carefully tuned data augmentation strategy that achieved an AUC of 0.96.[1, 2] Models using custom-made deep-learning networks achieved results with high accuracy, though not all were significantly better than radiologist labeling. As with any model, there are limitations of deep learning approaches such as data imbalance, model generalizability, verification bias, and ground-truth subjectivity.[4]

We aim to expand upon a previously developed deep learning model, MRNet, by reducing its feature space to improve efficiency. MRNet is a CNN model that maps the MRI series, composed of sagittal, coronal, and axial views, to a probability. It uses AlexNet for feature extraction, a global average

pooling layer for feature reduction, max pooling for calculating the probability, and binary cross-entropy loss for optimization. The highest performing model in the original MRNet publication achieved an AUROC of 0.965.[1] We evaluate whether we can achieve similar or better accuracy using just one MRI view (sagittal, coronal, or axial) to reduce computational complexity, cost, and time.

## Methods

*Dataset*

The dataset used in this study is derived from the MRNet dataset[1], which consists of 1,370 knee MRI exams, of which 1,250 have publicly available labels. The exams were obtained at Stanford University Medical Center between 2001 and 2012 with GE scanners and a standard knee MRI coil. Each exam consists of an axial Proton Density (PD)-weighted series, a coronal T1-weighted series and a sagittal T2-weighted series, which are each available as 256 x 256 NumPy array files.[1] T1 and T2 weighted series are two different sequences of MRI scans. T1-weighted series are produced by using short TR (repetition time) and short TE (time to echo), whereas T2-weighted series are produced by using long TR and long TE.

For this analysis, only the 1,130 exams with ACL tear labels were considered (Figure 1). Each exam is labeled with the presence (208 exams) or absence (922) of ACL tears, determined by the majority vote of three musculoskeletal radiologists. These exams were split into a training (904 exams), validation (226 exams), and test set (120 exams).

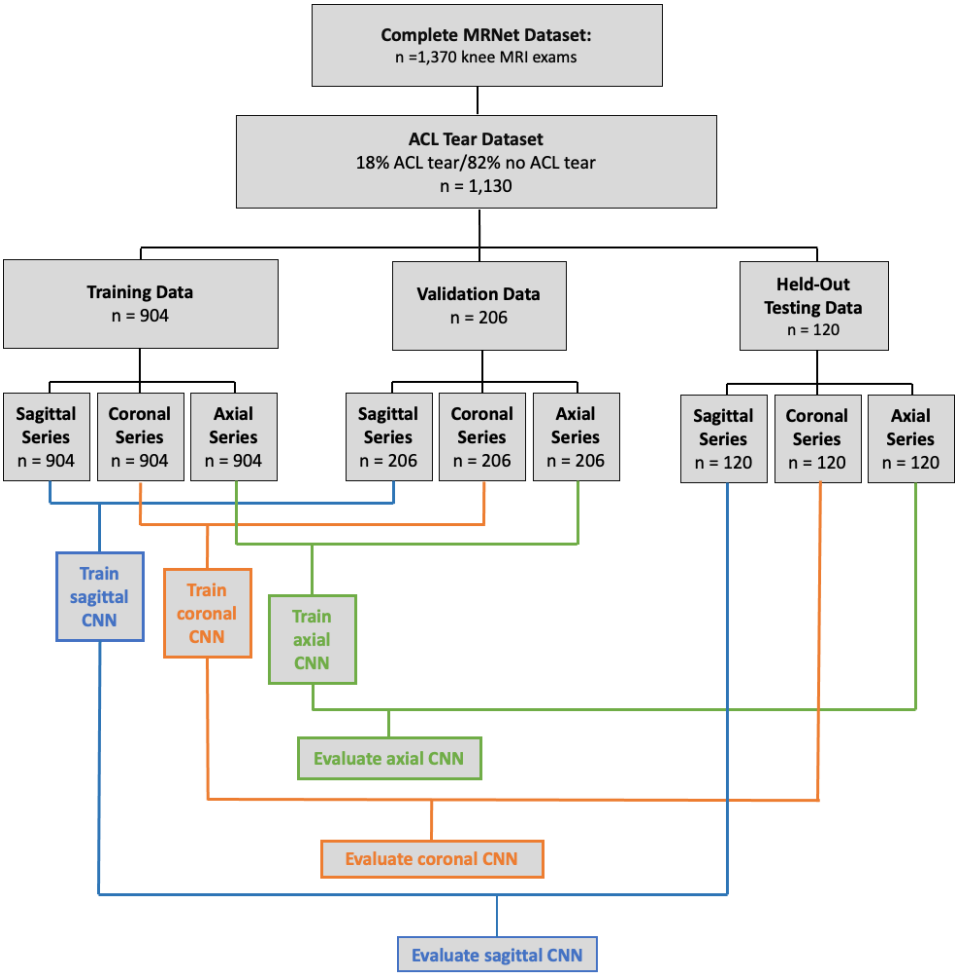



Figure 1. The MRNet dataset was filtered to include only exams with ACL tear labels. These exams were split into training, validation, and test sets, so that all series for each exam were in only one set. Three CNNs were trained using only their one series type (sagittal CNN, coronal CNN, and axial CNN) and evaluated with their respective series data in the held-out test set.

*Image Preprocessing*

NumPy array files available in the MRNet dataset were converted to a 3D array (1 x 256 x 256), normalized using min-max normalization and standardized using z-score normalization. A fourth dimension was then for three color channels, resulting in a final input tensor with dimensions, *s* x 3 x 256 x 256.

*Model*

MRNet is a CNN that maps a series of MRI images to a probability of the presence of an ACL tear. In its original form, MRNet trains three separate CNNs for each series type (sagittal, coronal, and axial), and then trains a logistic regression to weight predictions from each CNN to generate a single output probability associated with the MRI exam. We adapted MRNet to make the same prediction using only one MRI series type. Our adapted MRNet took as input one MRI series type with dimensions, *s* x 3 x 256 x 256, where *s* was the number of slices in the series (Figure 2). First, AlexNet was used for feature extraction to produce an *s* x 256 x 7 x 7 tensor. Next, global average pooling reduced these features to *s* x

256 and max pooling produced a 256-dimensional vector. This vector was passed to a fully connected layer and sigmoid activation function to produce a prediction between 0 and 1.

We trained three models, one with the sagittal series, one with the coronal series, and one with the axial series. Each model was trained using 50 epochs with a learning rate starting at $10^{-5}$. The learning rate was decreased by a factor of 0.3 if the validation loss did not reduce by more than $10^{-4}$ for five consecutive epochs. In order to minimize overfitting to the training data, the validation loss after each epoch was compared to each other and the model weights associated with the smallest validation loss were used. Weight decay was also used during training to further reduce overfitting to the training data.

*Evaluation*

For each series-specific MRNet, the model with the lowest validation loss during training was selected to evaluate with the corresponding series in the held-out test set (Figure 1). The area under the receiving operator curve (AUROC) was selected to evaluate the models to allow for a fair comparison between the original MRNet and our adapted MRNet. Furthermore, this metric's ability to capture both sensitivity and specificity is useful for emphasizing the importance of detecting ACL tears without false positives to enable the correct treatment as quickly as possible.

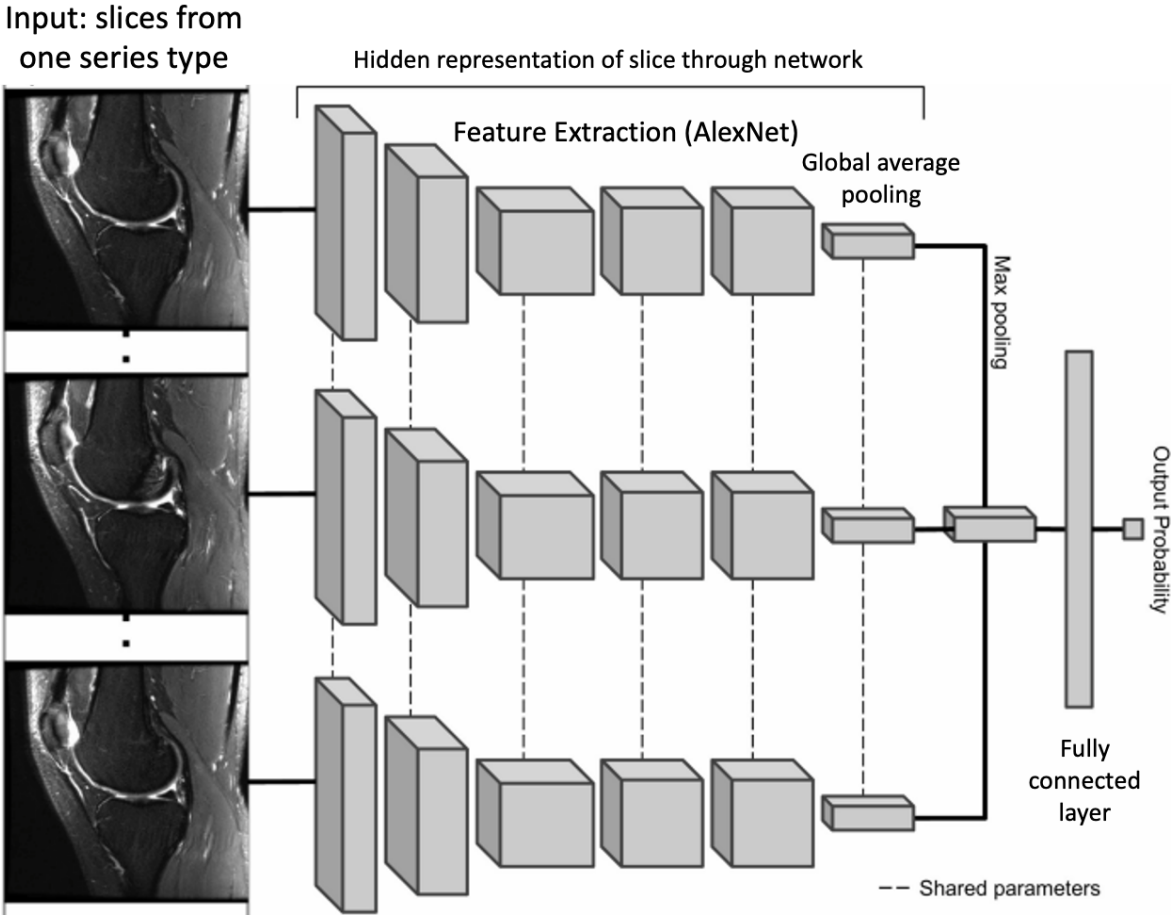



Figure 2. Adapted MRNet architecture based on figure by Bien et al. (2018).[1] Adapted MRNet was trained for each of three series types: sagittal, coronal, and axial.

*Class Activation Mappings (CAMs)*

Class Activation Mappings (CAM) were used to visualize the regions used by the CNN to predict the output class. The CNN, as described above, produces a final output layer after putting the image through various layers of convolution and global average pooling. The CNN decides which class an input image is classified as by feeding features from the final layer to a fully connected layer with sigmoid activation[9]. To produce the CAM, we project the output layer weights for each of these features back into the convolution maps that form the final convolution layer of the CNN[9]. We then apply a colormap (effectively a 'heatmap') that shows in color which areas of the image were weighted highly. As we are trying to detect ACL tears, we expect a color representing higher weight for the output classes to be found around the area of the ACL tendon.

**Results**

*Model Performance*

13 epochs were required to train both the axial and sagittal MRNet models, while 20 epochs were required to train the coronal MRNet model. The models with the lowest validation loss for each series varied in performance, with the coronal model producing the highest validation AUROC and the sagittal model producing the lowest validation AUROC (Table 1). When these models were evaluated on the held-out test set, however, the axial model was the highest-performing and the coronal model was the lowest-performing. While the MRNet trained with all three series had the highest test AUROC, the axial MRNet's test AUROC was only 0.022 points worse, despite the decrease in training time by almost five hours (Table 1).

Table 1. Performance of series-specific MRNets on reserved test set.

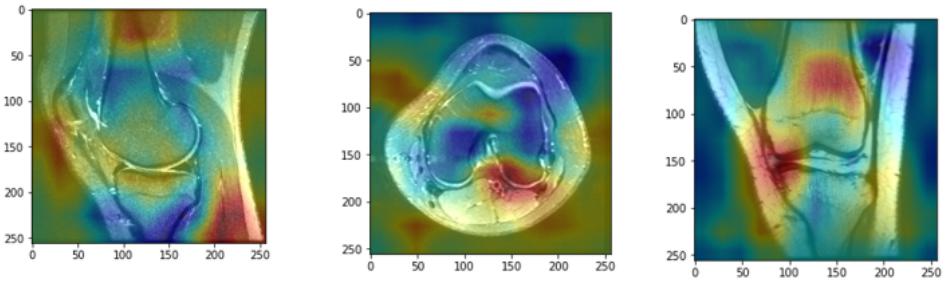
| **Series type** | **Train AUROC** | **Validation AUROC** | **Test AUROC** | **Training time** |
|---|---|---|---|---|
| Axial | 0.998 | 0.863 | 0.943 | 68 minutes |
| Coronal | 0.994 | 0.900 | 0.867 | 61 minutes |
| Sagittal | 0.998 | 0.836 | 0.932 | 65 minutes |
| All[1] | Not reported | Not reported | 0.965 | 6 hours |

*Class Activation Mappings*

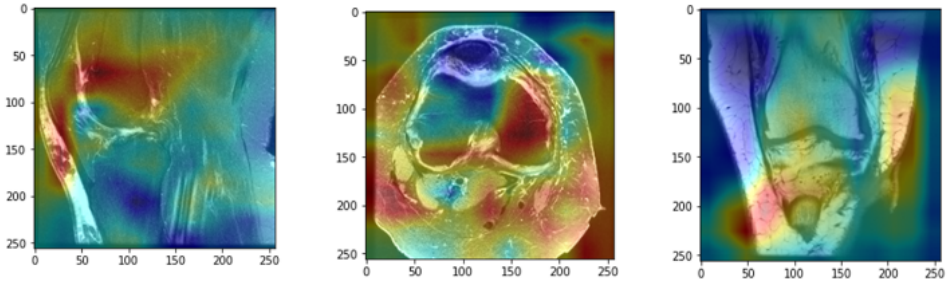
Class activation mappings (CAMs) of eight different patients are shown in Figure 3. These maps show the areas of the image that most indicated whether the patient would be classified as having an ACL tear or not. Blue regions are associated with regions of the image used for classification. In patients without an ACL tear, there are no clear color patterns, indicating that the model did not find an abnormality near the ACL. However, in the patients with an ACL tear, blue regions are consistently found near the ACL in axial and sagittal images. In coronal images, however, the images are skewed with incomplete or inaccurate representations of the patient's leg and knee.

Sagittal | Axial | Coronal

Patient without an ACL tear (Patient 0048)

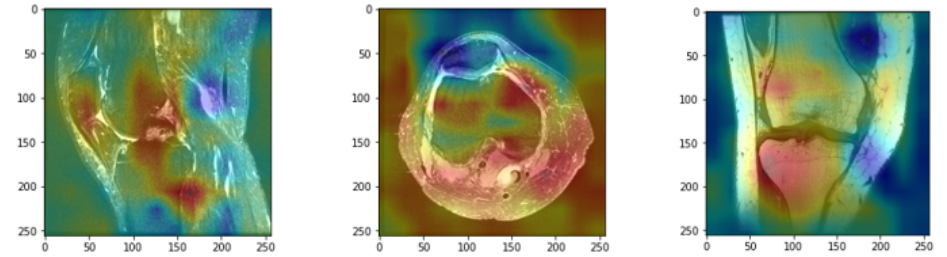


Patient with an ACL tear (Patient 0049)

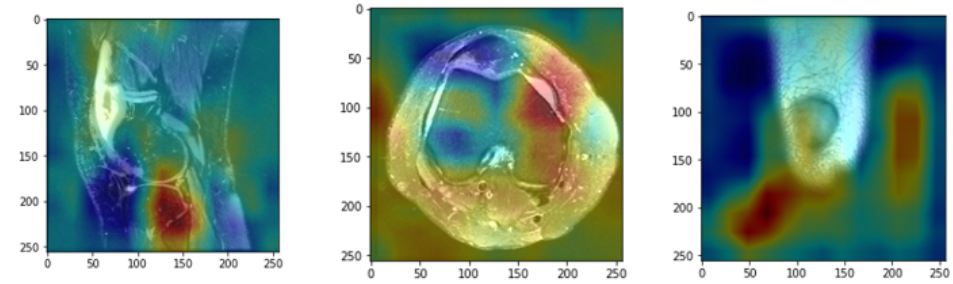


Sagittal | Axial | Coronal

Patient without an ACL tear (Patient 0000)



Patient with an ACL tear (Patient 0062)



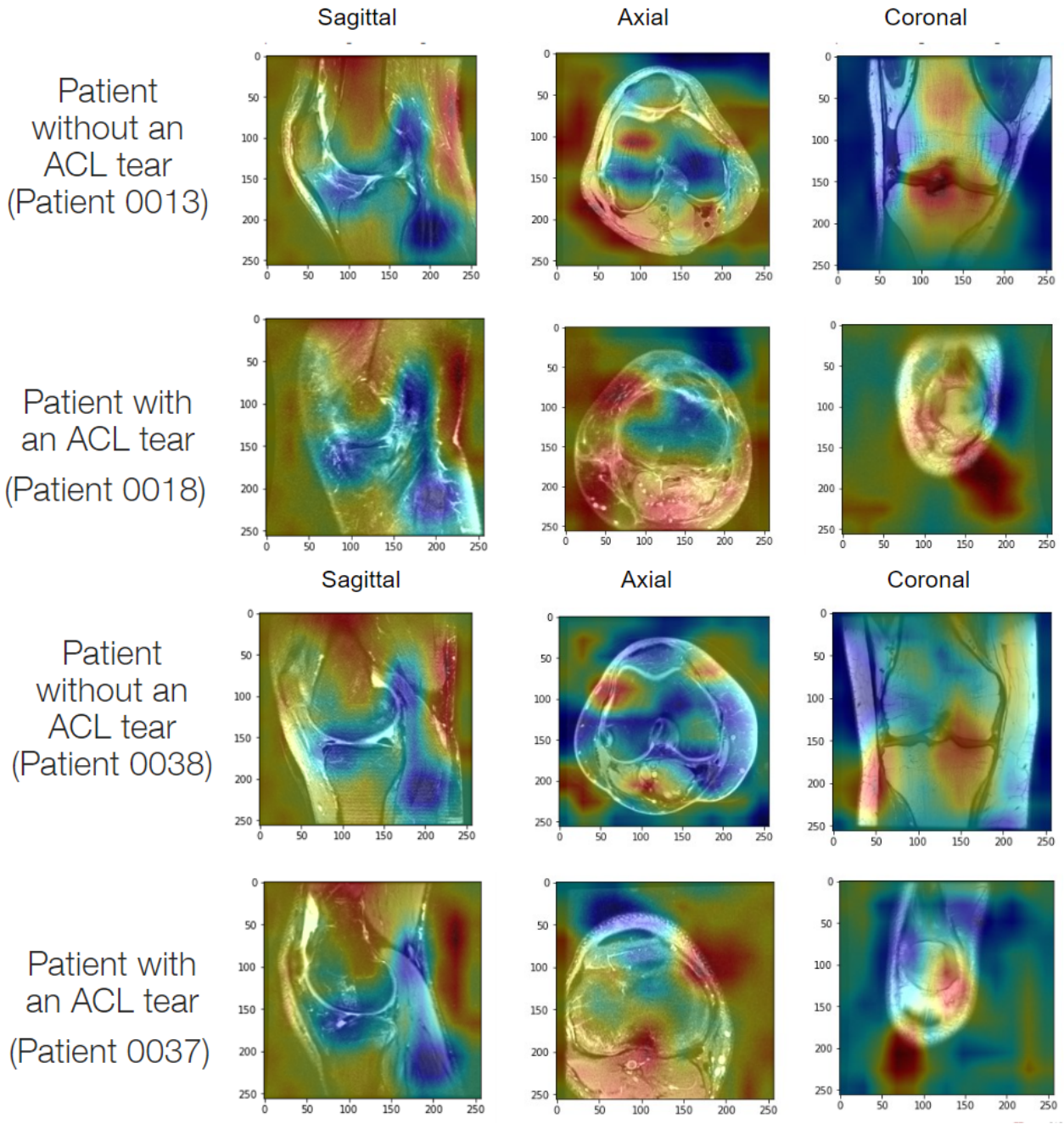



Figure 3. The above images show 8 patients, 4 with an ACL tear and 4 without. The sagittal, axial, and coronal views of the 18th slice of each patient's 36 slice volume is shown. To create this heatmap, the features of the output layer of the CNN used are projected back onto the last convolution layer map. The addition of a colormap ('heatmap') reveals areas that were weighted highly during classification, shown in blue, while regions with lower weights are red, yellow, or green. Of note is that some of the coronal slices for the patients with an ACL tear do not completely show the knee area.

**Discussion**

In this study, we expanded upon the highest-performing computational model for ACL tear diagnosis to our knowledge, MRNet. While MRNet has shown impressive performance in detecting ACL

tears from knee MRIs, it requires significant amounts of data and is relatively opaque, both of which limit clinical use. To address these limitations, we explored the feasibility of using one MRI series type for ACL tear prediction, rather than the three that are conventionally used (i.e.sagittal, coronal, and axial). We trained and evaluated three CNNs: a sagittal CNN, coronal CNN, and axial CNN. Of the three models, the axial CNN produced the best combination of sensitivity and specificity, captured in the AUROC. Encouragingly, its performance almost matched that of the original MRNet, despite its nearly sixfold decrease in training time. It is intuitive that the axial CNN outperformed both the coronal and sagittal CNNs when considering the location of the ACL in the middle of the knee.

To delineate what each CNN was learning, we employed CAMs, which reveal the image regions most associated with the classification prediction. These mappings indicate that predictions made by the axial and sagittal CNNs use the ACL region most strongly in classifications of ACL tears, while coronal images do not. Additionally, issues with the ground truth image in coronal projections may have contributed to the lower validation accuracy of the coronal CNN. This is in line with the better performance of the axial and sagittal CNNs, as compared to the coronal CNNs, and supports the hypothesis that the axial and sagittal models are using information about the ACL when making their classifications.

In the future, measuring additional model performance metrics that better capture class imbalance (i.e. precision, recall, and F1) would be valuable for comprehensively evaluating model performance. This, in combination with a clinical utility cost-benefit analysis, would elucidate how the model could be implemented in clinical practice. Additionally, further exploration of the usefulness of CAMs in assisting radiologists is warranted.

**Conclusion**

Our findings indicate that the high performance of MRNet may be possible with only a fraction of the original data. This significantly reduces the burden of model maintenance after deployment, especially for consistent retraining efforts.[8] Ultimately, this offers hope for a more tractable diagnostic model that can be included in the radiologist workflow to improve care for patients with knee injuries.

**References**


1. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, et al. (2018) Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLOS Medicine 15(11): e1002699. https://doi.org/10.1371/journal.pmed.1002699
2. Azcona D, McGuinnes K, Smeaton AF. (2020) A Comparative Study of Existing and New Deep Learning Methods for Detecting Knee Injuries using the MRNet Dataset. Arxiv. https://doi.org/10.48550/arXiv.2010.01947
3. Lao Y, Jia B, Yan P, Pan M, Hui X, Li J, Luo W, Li X, Han J, Yan P, Yao L. Diagnostic accuracy of machine-learning-assisted detection for anterior cruciate ligament injury based on magnetic resonance imaging: Protocol for a systematic review and meta-analysis. Medicine (Baltimore). 2019 Dec;98(50):e18324. doi: 10.1097/MD.0000000000018324. PMID: 31852123; PMCID: PMC6922500.
4. Awan MJ, Rahim MSM, Salim N, Mohammed MA, Garcia-Zapirain B, Abdulkareem KH. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. Diagnostics (Basel). 2021 Jan 11;11(1):105. doi: 10.3390/diagnostics11010105. PMID: 33440798; PMCID: PMC7826961.
5. Siouras A, Moustakidis S, Giannakidis A, Chalatsis G, Liampas I, Vlychou M, Hantes M, Tasoulis S, Tsaopoulos D. Knee Injury Detection Using Deep Learning on MRI Studies: A Systematic Review.

Diagnostics (Basel). 2022 Feb 19;12(2):537. doi: 10.3390/diagnostics12020537. PMID: 35204625; PMCID: PMC8871256.
6. Zhao M, Zhou Y, Chang J, et al. The accuracy of MRI in the diagnosis of anterior cruciate ligament injury. Ann Transl Med. 2020;8(24):1657. doi:10.21037/atm-20-7391
7. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221-248. doi:10.1146/annurev-bioeng-071516-044442
8. MLOps: Continuous delivery and automation pipelines in machine learning. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning. Accessed 3 June 2022.
9. Arif. Implementation of Class Activation Map (CAM) with PyTorch. 10 June 2020. Retrieved from https://medium.com/intelligentmachines/implementation-of-class-activation-map-cam-with-pytorch-c32f7e414923