MRNet MRI Classification

Cameron Rader
Min H. Kao Department of EECS
University of Tennessee at Knoxville
Knoxville, United States
crader6@vols.utk.edu

Suneil Patel
Min H. Kao Department of EECS
University of Tennessee at Knoxville
Knoxville, United States
spate200@vols.utk.edu

Babita Babita
Min H. Kao Department of EECS
University of Tennessee at Knoxville
Knoxville, United States
bbabita@vols.utk.edu

Ali Behbani
Mechanical, Aerospace and Biomedical Engineering Dept.
University of Tennessee at Knoxville
Knoxville, United States
abehbaha@vols.utk.edu

Pablo Castrejon
Mechanical, Aerospace and Biomedical Engineering Dept.
University of Tennessee at Knoxville
Knoxville, United States
pcastrej@vols.utk.edu

***Abstract*—Each year, approximately 400,000 knee reconstructions are performed in the United States due to injuries to the Anterior Cruciate Ligament (ACL). This type of injury typically occurs when strain stress is exerted on the knee ligaments. Most often, damage is caused by sudden stops while running, improper landings from jumps, or rotational forces applied to the knee.**

ACL injuries are often identified by the presence of microcracks or tears in the ligaments. Currently, one of the most accurate and reliable diagnostic tools for detecting ACL damage is Magnetic Resonance Imaging (MRI). While MRI provides sufficient image quality to detect ligament cracks, accurate diagnosis still depends heavily on the expertise of trained radiologists. The process of analyzing scans can be tiresome, time-consuming, and prone to human error. Moreover, accurately determining the extent of the damage is even more challenging, despite its importance—moderate lesions, for instance, may be treated conservatively without the need for surgical intervention.

In this work, we propose the study, analysis, and implementation of Convolutional Neural Networks (CNNs) to develop a decision-support model aimed at detecting ACL-related knee injuries and classifying cases based on severity. Specifically, we explore a baseline model based on ResNet18, with the goal of optimizing it for the categorization of ACL ligament damage.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The interpretation of knee MRI presents significant clinical challenges due to diagnostic variability caused by the subjective nature of image analysis [1] and the complexity introduced by various pathologies [2], [3] . Radiologists must handle a large volume of data, as knee MRI scans typically include multiple orthogonal and oblique planes acquired using different pulse sequences at high resolution [4], [5] . This complexity makes the analysis of large image sets both time-consuming [5], [6], [7] and error-prone [1], [6], [7], [8], [9] , further complicated by cognitive fatigue and distractions [1], [3], [10] . In addition, inconsistent intensity readings, even under standardized protocols, require frequent manual adjustments, complicating tasks such as segmentation and quantitative assessment [9].

Currently, workflows rely heavily on manual interpretation, creating bottlenecks when detecting subtle injuries and providing precise measurements [7], [8], [9], [11], [12] . These workflow limitations, which also include time constraints, inter and intra-observer variability [1], [13] , as well as the risk of overlooking subtle findings [3], [14] , highlight the need for automated solutions [6], [9], [15] .

By applying deep learning techniques and AI-based methods, automated knee MRI interpretation aims to improve clinical efficiency [7], [15], [16]. These approaches seek to reduce the time radiologists spend on routine evaluations [7], [17] , decrease diagnostic variability through objective, standardized analysis [1], [11], and improve detection accuracy across various knee abnormalities [15], [17] . Ultimately, this integration is expected to streamline clinical workflows, support personalized diagnostic and treatment strategies, and lastly, enhance resource utilization in healthcare [17], [18], [19], [20] .

The primary objective of advancing deep learning for knee MRI interpretation is to develop more accurate, robust, and clinically applicable models that enhance radiologists' diagnostic capabilities [2], [26]. Mostly building upon MRNet, which provided a benchmark dataset for common knee injuries [20], future studies aim to address its limitations by expanding pathology coverage, improving spatial contextualization using 3D MRI data, and enhancing model generalization across different MRI protocols and scanner vendors [20]. The leverage of 3D CNN architectures as well as integrating multi-planar fusion techniques could improve diagnostic accuracy, particularly for complex injuries or subtle cartilage lesions. Domain adaptation and data augmentation strategies are crucial for mitigating cross-domain variability, ensuring model robustness across diverse imaging conditions. Task-specific models trained for detecting individual pathologies [20] , as well as segmentation-assisted classification approaches, may offer superior accuracy by incorporating anatomical priors [13], [29] . These innovations collectively contribute to the development of more reliable and clinically integrated AI-driven knee MRI

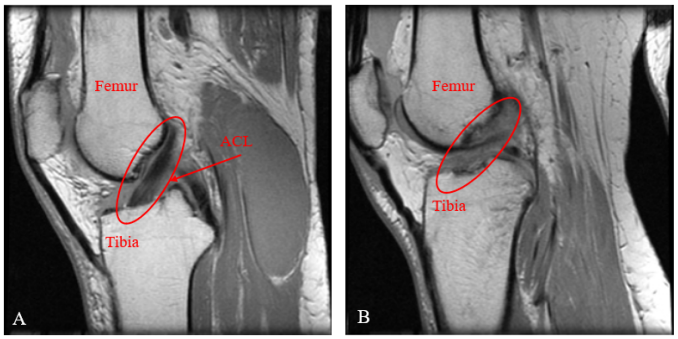

Fig. 1. From the MRI images above published by Weber State University it is possible for a trained radiologist to provide an accurate diagnosis. Figure A shows how an ACL should look like under normal conditions, the ACL ligament intersecting the femur and tibia can be observed in a darker tone, while in Figure B we can observe an apparent absence of the ligament ACL,

analysis systems, ultimately enhancing diagnostic precision and efficiency in musculoskeletal imaging [13], [20], [28] .

II. Literature Review

Review – The Anterior Cruciate Ligament is a fibrous band connecting the femur to the tibia, it has to be strong enough to support the stresses applied such as the individual's own weight, however this ligament can be torn or partially torn due to high intensity physical training or due to non-natural sudden motions applied to the knee. An injured ACL is usually followed by a "popping sound" and manifests intense pain followed by excess fluid accumulation in the tissues around the knee, leading to instability and limitations on the range of motion [31].

Early detection of ACL injury is crucial to prevent irreversible damage, in many cases ACL injury can cause damage to the meniscus and articular cartilage and in some cases may even lead to osteoarthritis, current detection methods include the use of a magnetic resonance imaging (MRI), this type of examinations offer a high accuracy diagnosis of ACL injury, and has shown to be reliable when compared to other used methods such as arthroscopy analysis. A physician is needed to interpret MRI and arthroscopy analysis, this analysis is considerably time-consuming and requires experienced experts to correctly diagnose ACL injury [32]. By taking advantage of modern, high capacity GPU's it is possible to analyze and process enormous quantities of data.

Deep learning, particularly convolutional neural networks (CNNs), has become a leading approach in medical image analysis [21] , demonstrating expert-level accuracy in detecting subtle lesions across various imaging domains [22], [23], [24] . CNNs excel at learning hierarchical data representations [16], [25] , making them well-suited for knee MRI interpretation, which involves complex anatomical structures and diverse pathologies. They have been successfully applied to knee cartilage segmentation using multi-stream architectures and have shown high diagnostic performance in detecting cartilage lesions, anterior cruciate ligament (ACL) injuries, osteoarthritis, and meniscal abnormalities. Transfer learning further enhances deep learning applications in medical imaging by leveraging knowledge from large-scale datasets, such as ImageNet, to improve performance on smaller datasets like knee MRI pathology detection [9], [13], [20] . Pretrained models, such as VGG16 [13] , provide a strong initialization that reduces training time, prevents overfitting, and enhances feature extraction for tasks like tissue segmentation and image contrast conversion. To address dataset variability caused by different MRI protocols and scanner settings, data augmentation techniques—such as rotations, flips, and scaling—expand training diversity, improving model generalization and robustness across unseen datasets. Additionally, interpretability techniques, such as Grad-CAM and saliency maps, enhance model transparency by highlighting key image regions that influence predictions, fostering clinician trust, and ensuring that the model focuses on clinically relevant features rather than spurious correlations. These advancements collectively contribute to the feasibility of deep learning-driven knee MRI analysis, offering improved diagnostic accuracy, efficiency, and reliability in clinical practice [21], [24] By enhancing interpretability through methods like Grad-CAM or saliency maps it is possible to improve clinician trust and adoption by providing insights into model decision-making [27]. To assess model performance, specially against human experts, Researchers from multi-reader studies would involve radiologists interpreting knee MRI cases with and without AI assistance, with evaluations based on key metrics such as AUC-ROC, sensitivity, specificity, accuracy, and inter-rater reliability [20] . These studies should also assess workflow efficiency improvements, including interpretation time reduction [28] . Technical advancements such as multi-modal learning (leveraging multiple MRI sequences) [13], [20] , attention mechanisms (focusing on relevant anatomical structures) [9] , and hybrid models (combining CNNs with traditional machine learning techniques) [9] can further optimize performance.

III. Technical Approach

A. MRNet Dataset Overview

The MRNet dataset, developed by the Stanford ML Group, consists of knee MRI exams from approximately 1,370 patients. Each examination includes three standard diagnostic views: axial, coronal, and sagittal. The dataset is structured to support classification across three distinct tasks:

- Abnormality detection
- Anterior cruciate ligament (ACL) tears
- Meniscal tears

Each MRI examination is stored as multiple `.npy` files—one per anatomical view. These files contain 3D volumes represented as sequences of 2D slices captured along the respective anatomical plane. The number of slices varies significantly between patients and views, ranging from approximately 15 to over 100 slices per view, presenting a challenge for standardized processing.

B. Data Pre-processing Pipeline

In the development of this pipeline, a great deal of attention was placed in the pre-processing of our data samples. MRI data, unlike natural images that traditional CNN's are trained on, can be extremely variable in size, intensity, and orientation. Because of this, careful consideration was taken to ensure that our images were normalized and prepared for training on backbone CNN's that would be less than optimally configured for MRI data and medical imaging in general. The techniques below attempt to mitigate these differences and give the model the best chance for success in classifying these types of images:

1. Volume Standardization: Each MRI exam is naturally stored as a 3-D .npy volume of 2-D slics. Across the samples these slice counts per exam varied significantly, to create a consistent input format the following sequence of transformations were applied to the data samples:

2. Volume Orientation Normalization: Volume Orientation Normalization: Each 3-D volume was checked and converted to a (Slice, Height, Width) format to match the expected shape needed for input into a pretrained backbone feature extractor.

3. Slice Count Normalization: To address the varying slice counts across patient samples a slice cropping technique was employed. This slice cropping sets a max slice allowance and normalizes all samples to either a max slice count of 32 or 64. Early experiments with slice cropping degradated the performance of the model, however by cropping the outer spaces of the MRI, focusing on the center of the image, we were able to capture the most central and informative slices.

4. Per-Slice Min-Max normalization: Each 2-D slice is normalized to values in the [0,1] range. A standard practice in an attempt to avoid differences in contrast between the MRI scans in our dataset

5 . Spatial resizing: After all of this normalization, slices are then resized to a consistent resolution of 224x224 pixels

6 . Channel multiplication for CNN Compatability: Because the MRI samples are created as single-channel greyscale images and our backbone feature extractors were trained on 3-channel RGB images, the pipeline expands our samples replicating the single greyscale channel into 3 matching the expected channel input of the pre-trained backbone.

7 . ImageNet Normalization: To get the most out of the transfer learning process, slices are normalized using common ImageNet mean and standard deviation values ensuring that the CNN's can train within their expected ranges.

All of the normalization and standardization techniques above are implemented in an attempt to decrease the stark differences between medical imaging data and the RGB images that most of the pre-trained feature extractors have been trained on. Without these techniques, all of the knowledge that the backbone feature extractor has learned will become obsolete during our transfer learning process. By attempting to mirror this pretrained-CNN training data for our transfer learning task the input data can be properly evaluated and learned from.

C. Data Augmentation Strategy

Throughout the implementation process of this MRI classification pipeline, a variety of different augmentation strategies were considered in an attempt to provide the model with wide generalization ability. While the target of these strategies is to force better generalization, the rational for a strong and innovative augmentation strategy is more complex. When desigining the augmentation strategy a core challenge stood out:

- Unlike natural image datasets (e.g., ImageNet) that offer millions of diverse images, the MRNet dataset provided a relatively small and specialized sample of knee MRIs across three diagnostic planes.
- Because of this scarcity, every single MRI scan carried a disproportionate weight in the model's learning process. It wasn't just about avoiding overfitting — it was about making sure the model could form strong, clean, foundational representations of what real MRI features look like without the representations being corrupted too early by aggressive transformations.

1. Why Simple Augmentation could be efficient:

- Too little augmentation, or simpler standard augmentation, meant that the model would quickly overfit even after a very small number of epochs, memorizing the extremely small dataset
- Too much augmentation applied from the start risked corrupting the critical low-level features that define MRI images — fine tissue boundaries, meniscus lines, ligament tears — which are much subtler than the features models learn on natural image datasets.
- Unlike classifying cats and dogs found in the ImageNet dataset, every small pixel gradient in an MRI matters. Aggressive or random augmentations early in training could degrade the already fragile signals the model needed to detect.

2. Why a Progressive Augmentation Strategy was considered:

- A data augmentation scheduler, similar to techniques like curriculum learning, was considered to mirror a staged learning process much like how a human would learn
- First, the model should focus on understanding clean, high-quality representations of MRIs without distortions.
- Then, after forming some "mental models" of what normal MRI structures look like, it could be challenged or "tested" progressively with distorted or augmented versions, encouraging it to develop more robust, invariant feature detectors.
- Finally, after those robustness skills were built, it made sense to return to clean data late in training to "polish" the learned representations on real-looking images — ensuring the final features remained clinically meaningful and undistorted.

These considerations led to the approach evaluated below and then compared to the simple, standard data augmentation.

3. Three-phase curriculum (DataAugmentationScheduler):

- **Warm-up** – first 10 percent of epochs, no augmentation.
- **Phase 2: Exploration** – next 40 percent of epochs. During phase 2, augmentation is linearly applied. Augmentation probability is increased from 0.4 to 0.9. Rotation increases from +3 degrees to +15 degrees. Brightness increases from +2 percent to +15 percent.
- **A**dditionally, Gaussian Blur is applied with a 3 percent probablity. As well as Random Cutout which is applied with a 4 percent probablity, zeroing out approximately 25 percent of each slice's area.
- **Fine-tune** – final 50 percent of epochs, augmentation is significantly reduced to finalize training on cleaner representations. This allows the model to consolidate robust features learned during the harder exploration phase, realigning final feature maps to match clean clinical MRI distributions.

This 3 phase augmentation strategy was then compared to the Simple Augmentation approach outlined below and the results were evaluated

4. Static transform (SimpleMRIAugmentation):

- Random rotation up to $\pm 5°$.
- Random brightness change up to $\pm 5\%$.
- Each applied with probability 0.5 on every slice.

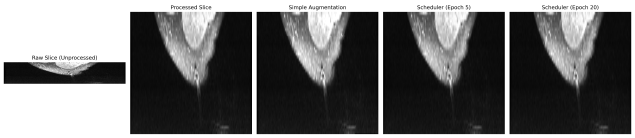

Figure 2. Displays the progression MRI images take as they are augmented through the model using the 3-phase scheduler.

D. Dataset Challenges and Insights

The MRNet dataset presents several unique challenges:

1) **Class Imbalance:** Positive label distribution varies across tasks—25% (abnormality), 17% (ACL tears), 35% (meniscal tears)—requiring robust evaluation strategies.
2) **Anatomical View Significance:** Some pathologies are more evident in certain views (e.g., ACL tears in sagittal, meniscal tears in coronal). This anatomical insight guided our modeling approach.
3) **Slice Relevance Variability:** Many slices contain irrelevant or noisy information. Intelligent slice selection or attention weighting could improve performance.
4) **Memory Constraints:** Processing high-resolution, multi-slice volumes is GPU-intensive. Our pipeline balances informativeness with hardware limitations.

IV. Baseline Selection

A. Motivation

Before proposing customized architectures we first establish a strong, transparent baseline that every subsequent variant can be fairly compared against. Three main pre-trained CNN feature extractors were evaluated and compared: ResNet-18, ResNet-34, and DenseNet-121.**1. Historical Precedent.** ResNet-18 was used in the original Stanford MRNet paper, and nearly every strong attempt at the competition around 2018–2019 had adopted it as the default backbone. As such, it represented a reliable and well-understood starting point. **2. Model Simplicity.** ResNet-18 strikes a balance between sufficient feature extraction capacity and relatively low model complexity, making it a reasonable choice for relatively small and specialized medical datasets like MRNet. **3. Transfer Learning Compatibility.** Pre-trained ResNet-18 weights (trained on ImageNet) are widely available and have been proven effective for fine-tuning tasks with limited medical imaging data.

Thus, ResNet-18 was selected as the baseline backbone to anchor our experiments.

B. Exploring Deeper Networks: Evaluating ResNet-34 and DenseNet-121

Given that the MRNet competition this project is built around is approximately six years old, it should be noted CNN architectures have evolved considerably and exploration of newer, deeper, more powerful networks could possibly provide improved classification. Newer, denser backbones such as ResNet-34 and DenseNet-121 offered:

- Increased depth and representational capacity
- Theoretical ability to extract more complex and subtle features
- Improvements in natural image classification benchmarks over ResNet-18

Based on this, our hypothesis was that upgrading to a deeper, more powerful backbone might improve performance on MRNet, especially given the subtle, high-dimensional features that characterize MRI scans.

C. The Impact of Dataset Size

Through extensive experimentation, we discovered that deeper and denser networks like ResNet-34 and DenseNet-121 consistently overfit the MRNet dataset, leading to:

- Higher training AUC, but poorer validation performance
- Large generalization gaps
- Clear signs of overfitting on the limited training samples

It became clear that the choice of backbone is heavily dependent on the number of available training samples.

On a larger dataset (e.g., 10,000 MRI samples) it is likely that ResNet-34 or DenseNet-121 could outperform ResNet-18 by learning finer, more complex patterns. However, on our dataset (˜1,370 studies), the limited sample size meant that deeper models had too much capacity relative to the amount of available training data, causing them to memorize noise rather than generalize to meaningful anatomical features. This realization was pivotal to shaping our modeling strategy, providing a pivotal understanding that higher capacity models do not always indicate better performance across datasets of differing sizes.

V. Baseline Implementation

Our baseline implementation was designed to not only replicate strong results from the original Stanford MRNet project, but also push the pipeline forward with modern deep learning practices tailored for small-scale medical imaging datasets. Following the careful selection of data preparation techniques and the baseline backbone architecture, we designed and implemented a complete MRI classification pipeline. This section outlines the flow of our pipeline — from model construction, through the training process, to evaluation — while highlighting key design decisions that were intended to improve model performance and generalization.

A. Model Construction

Building on the selected ResNet-18 backbone, we designed a modular architecture capable of handling the noisy and variable nature of MRI data:

1) Feature Extractor (CNN Backbone): ResNet-18 pretrained on ImageNet was used as the backbone for feature extraction, with its final fully connected layer removed.

2) Slice Attention Mechanism: Early experimentation provided an interesting insight into how the model viewed each slice as it was passed through. Given that not all MRI slices are equally informative, we introduced a lightweight slice attention module. This attention module was included in an attempt to correct early poor validation and testing accuracies. Each slice embedding was passed through a small neural network (Linear $\rightarrow$ Tanh $\rightarrow$ Linear) to compute attention scores, which were then softmaxed across slices to create slice weights. This mechanism allowed the model to dynamically prioritize slices likely to contain pathology, rather than treating all slices equally.

3) Mean-Max Feature Pooling: To capture both average trends and peak activations across the slices, we concatenated the mean and max pooled features before passing them to the classification head. This strategy provided a more informative feature representation without significantly increasing model complexity.

4) Classification Head: The pooled features were passed through a two-layer fully connected classifier with batch normalization, dropout regularization, and ReLU activations to predict the probability of abnormality, ACL tear, or meniscus tear.

B. Training Strategy

Training the model effectively on a small and specialized dataset required careful handling to avoid overfitting and instability. Several strategies were implemented:

1) Backbone Freezing and Progressive Unfreezing: Initial experiments with a fully frozen backbone led to rapid overfitting and poor validation performance. To address this, we implemented a progressive backbone unfreezing schedule:

- The feature extractor remained fully frozen for the first 50 percent of training epochs, allowing the classification head to stabilize.
- At the midpoint of training, the top 40 percent of backbone layers were unfrozen, gradually allowing the model to specialize its features.
- Near the end of training (80 percent of epochs), the full backbone was unfrozen, with an increased learning rate for the backbone parameters to enable meaningful fine-tuning.

This approach provided the backbone time to adapt to the MRI domain without destabilizing early training dynamics.

2) Data Augmentation Strategy in Practice: While the data augmentation methodology was described earlier, it is important to note how it was practically integrated:

- A `DataAugmentationScheduler` dynamically adjusts augmentation strength over the course of training.
- Augmentation strategies included random rotations, brightness shifts, Gaussian blurring, and random cutout, with augmentation probability increasing during the "exploration" phase and decreasing during the final "fine-tuning" phase.

This curriculum-based augmentation scheduling helped prevent the model from overfitting early while allowing it to consolidate clean feature representations before final convergence.

3) Optimization and Loss:

- **Loss function:** Binary cross-entropy loss with logits was used for all tasks. Positive class weighting was dynamically calculated based on the training set label distribution to mitigate class imbalance effects.
- **Optimizer:** We used AdamW, an optimizer well-suited for deep learning tasks with weight decay regularization, which helped improve generalization.
- **Differential Learning Rates:** To coincide with our progressive unfreezing approach, a progressive learning rate structure was applied. To prevent destabilizing pretrained weights, the backbone was trained with a learning rate three times lower than the classification head. A OneCycleLR scheduler was used to warm up learning rates early in training and decay them steadily toward zero, promoting better convergence.

Baseline Performance Metrics

To evaluate the effectiveness of our transfer-learned ResNet18 model, we report several performance indicators across the three classification tasks in the MRNet dataset: abnormality detection, ACL tear detection, and meniscal tear detection. All

experiments were variable controlled examining the effect of a single change for each experiment:

1. Training and Validation Curves

Figure 3 displays the training and validation loss curves over training epochs

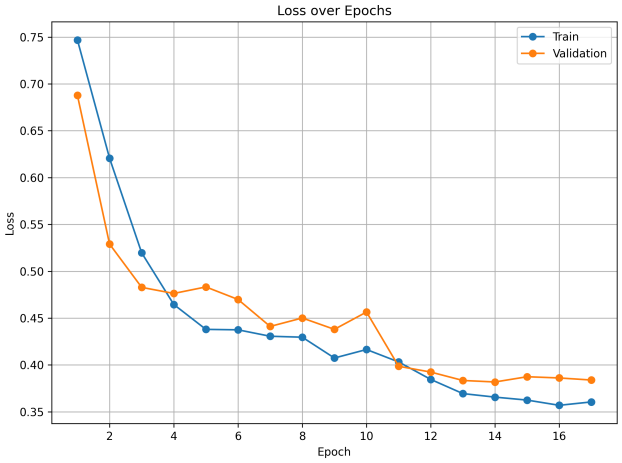

Figure 3. Training progress showing training/validation loss across epochs

2. Final Test Set Evaluation

Abnormality Detection:

- **Test AUC:** 0.87
- **Test Accuracy:** 82.3

ACL tear Detection:

- **Test AUC:** 0.837
- **Test Accuracy:** 84.1

Meniscal tear Detection:

- **Test AUC:** PLACEHOLDER
- **Test Accuracy:** PLACEHOLDER

A confusion matrix for one of the tasks (e.g., abnormality detection) is shown below. The visualization displays the model's ability to classify true positive and true negative data samples. As well as it's shortcomings with it's incorrect predictions

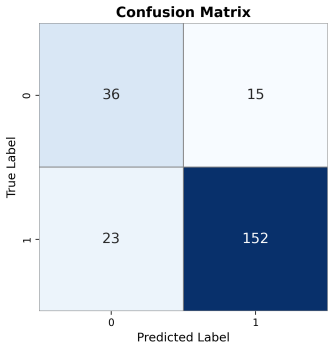

Figure 2. Confusion matrix for abnormality detection

C. 3. Ablation Studies and Experiment Results

1) 1. Effect of Attention Mechanism:

a) Purpose.: We aimed to investigate whether the inclusion of a slice attention mechanism improved model performance compared to uniform slice pooling across all MRI slices.

b) Experimental Setup.: Two models were trained under identical conditions: one with the slice attention module enabled, and one without. All other hyperparameters, including the backbone architecture, augmentation scheduler, and training epochs, were kept the same.

Model	Validation AUC	Test AUC
With Attention	**XX.XX**	**XX.XX**
Without Attention	**XX.XX**	**XX.XX**

TABLE I
COMPARISON OF MODELS WITH AND WITHOUT SLICE ATTENTION MECHANISM.

c) Results.:

d) Analysis.: The model utilizing slice attention achieved higher validation and test AUC scores compared to the model without attention. This suggests that learning to dynamically prioritize informative slices helps the model develop stronger feature representations, leading to better generalization on unseen MRI scans.

2) 2. Impact of Data Augmentation Strategy:

a) Purpose.: We investigated whether using a progressive three-phase augmentation scheduler could improve model performance compared to applying simple static augmentation throughout training.

b) Experimental Setup.: Two models were trained: one with the static `SimpleMRIAugmentation` applied uniformly across all epochs, and one with the three-phase `DataAugmentationScheduler` that dynamically increased and decreased augmentation strength during training.

Augmentation Strategy	Validation AUC	Test AUC
Augmentation Scheduler	**XX.XX**	**XX.XX**
Simple Augmentation	**XX.XX**	**XX.XX**

TABLE II
COMPARISON OF SIMPLE AUGMENTATION VS. AUGMENTATION SCHEDULER.

c) Results.:

d) Analysis.: The progressive augmentation scheduler achieved improved validation and test AUC scores relative to static augmentation. This suggests that dynamically increasing augmentation difficulty mid-training helped the model learn more robust features without overwhelming it early on, aligning well with curriculum learning principles.

3) 3. Backbone Freezing Strategies:

a) Purpose.: We aimed to evaluate how different backbone training strategies impacted model generalization, comparing fully frozen, partial unfreezing, and fully unfrozen approaches.

b) Experimental Setup.: Three models were trained:

- **Frozen Backbone**: The feature extractor remained frozen throughout training.
- **Partial Unfreezing**: The backbone was gradually unfrozen, first partially, then fully during training.
- **Fully Unfrozen**: All backbone layers were trainable from the beginning of training.

All other hyperparameters were kept constant across runs.

Freezing Strategy	Validation AUC	Test AUC
Partial Unfreezing	**XX.XX**	**XX.XX**
Frozen Backbone	**XX.XX**	**XX.XX**
Fully Unfrozen	**XX.XX**	**XX.XX**

TABLE III
COMPARISON OF FREEZING STRATEGIES ON MODEL PERFORMANCE.

c) Results.:

d) Analysis.: Partial unfreezing yielded the best balance between training stability and generalization. Models with a fully frozen backbone struggled to adapt to the MRI domain, while fully unfrozen backbones tended to overfit quickly. Gradually introducing trainable backbone layers over the course of training allowed the model to stabilize initially and adapt meaningfully later.

D. 4. Summary of Results and Insights

The transfer-learned ResNet18-based model provides a strong foundation for our ongoing research. Our implementation, built upon the original MRNet architecture proposed by the Stanford ML Group, introduces several methodological innovations aimed at enhancing model generalization and interpretability. Notably, we incorporated a three-phase data augmentation scheduling approach, attention-weighted slice pooling, and experimented with backbone layer freezing strategies for optimized transfer learning. These adjustments provided a structured approach to mitigating overfitting and allowed for more nuanced learning of clinically relevant features. Preliminary results from our pipeline indicate promising performance, demonstrating that these methodological variations can effectively leverage the strengths of the MRNet dataset while offering valuable insights into the potential improvements over the original implementation.

VI. MODEL IMPROVEMENTS AND FUTURE WORK

A. Completing the Multi-View Ensemble

Although our pipeline supports single-view (axial, coronal, or sagittal) training, we only partially implemented the full three-view ensemble architecture. Combining features extracted independently from axial, coronal, and sagittal planes before classification could provide several benefits:

- **C**apture complementary anatomical information present across different planes.
- **I**mprove robustness, as certain pathologies may be more visible in one plane than another.
- **P**otentially achieve a significant boost in AUC, especially for ACL and meniscus tear tasks that proved to be more challenging to classify than abnormality detection

Future work: Finalizing and training the three-view ensemble would likely lead to stronger classification performance across all tasks.

B. Capturing 3D Spatial Context

Currently, the model processes individual 2D slices independently and uses attention mechanisms to aggregate across them. However, this approach does not allow the model to directly model 3D spatial relationships between adjacent slices.

Potential Improvements:

- **E**xplore using 3D convolutional networks (3D CNNs) to learn volumetric features that might not be present looking soley at the 2-D slices of the 3-D volume
- **A**lternatively, treat MRI volumes as sequences and apply transformer models or recurrent networks (like GRUs) across slices to capture inter-slice relationships.

Impact: Capturing the full 3D structure could allow the model to learn richer, more consistent representations of anatomical features and abnormalities.

C. Alternative Backbone Architectures

While ResNet-18 proved effective given our dataset size, future experiments could explore more advanced or specialized backbone networks, such as:

- **M**edical imaging-specific models, e.g., DenseNet variants pretrained on RadImageNet.
- **L**ightweight modern CNNs like EfficientNet or ConvNext for better performance without heavy computational cost.
- **H**ybrid models combining convolution and self-attention (e.g., ConvNeXt or MobileViT) that might be better suited to capturing complex MRI structures.

Impact: Newer architectures could improve feature extraction without requiring substantially larger datasets.

D. Data Expansion and Semi-Supervised Learning

A major limitation throughout the project was the relatively small size of the MRNet dataset (1,370 examples).

Potential strategies to overcome this:

- **D**ata Augmentation: Explore stronger augmentation methods like CutMix, Mixup, or synthetic data generation
- **S**emi-Supervised Learning: Use models trained on labeled data to generate pseudo-labels for large amounts of unlabeled MRI scans.
- **T**ransfer learning from related tasks: Pretraining on similar medical imaging datasets could provide better initialization than ImageNet.

Impact: Expanding the training data or leveraging unlabeled data could substantially improve generalization and model robustness.

E. Improved Attention Mechanisms

Our current slice attention mechanism is a lightweight two-layer MLP, which worked well, but future directions could include:

- **S**elf-attention or Transformer-style attention across slices to better model long-range dependencies.
- **H**ierarchical attention, where slice-level attention is combined with regional attention within slices.
- **A**ttention Regularization, such as entropy penalties, to encourage sharper or more meaningful attention distributions.

Impact: Stronger attention mechanisms could lead to more interpretable and more accurate models.

F. Further Exploration of Freezing Strategies

While the progressive unfreezing strategy introduced in our pipeline improved model generalization, it represents only one possible configuration. Future experiments could explore:

- **V**arying the percentage of the backbone that is unfrozen at different stages of training.
- **A**ltering the order in which layers are unfrozen — for instance, unfreezing earlier convolutional blocks first versus later, higher-level blocks first.
- **D**ynamically adjusting the learning rates of different layers based on their depth within the network during unfreezing.

Impact: Fine-tuning the freezing and unfreezing schedule could unlock further performance gains by better matching the model's adaptation curve to the learning needs of the MRI data.

G. Further Exploration of Augmentation Scheduling

Our implementation of a three-phase augmentation scheduler (clean $\rightarrow$ augmented $\rightarrow$ clean) showed promising results, but several alternative augmentation curricula could be investigated:

- **A**ugmented $\rightarrow$ Clean: Start training with strong augmentation and gradually phase it out.
- **C**lean $\rightarrow$ Augmented: Start clean, then introduce augmentation progressively without returning to clean data.
- **A**ugmented $\rightarrow$ Clean $\rightarrow$ Augmented: Start with augmentation, fine-tune on clean data, then reintroduce harder augmentations late in training.

Impact: By systematically exploring these different augmentation progression schemes, it would be possible to better understand how curriculum timing affects feature learning and generalization, leading to even more robust training pipelines.

H. Summary

In summary, while the pipeline presented in this work achieved strong baseline performance, several promising avenues for improvement remain. Completing the three-view ensemble, incorporating full 3D spatial reasoning, exploring more powerful backbones, expanding data availability, further investigating our freeze and augmentation scheduler strategies, and strengthening attention mechanisms represent key future directions to further advance model performance and clinical applicability.

PROJECT REPOSITORY

The complete implementation, including preprocessing scripts, model architecture, training logs, and evaluation code, is available on GitHub:

`https://github.com/Cameronr11/DL_Project_Team15`

REFERENCES

[1] L. G. Nyúl and J. K. Udupa, "On Standardizing the MR Image Intensity Scale," *Magn. Reson. Med.*, vol. 42, pp. 1072–1081, 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6.

[2] I. Štajduhar *et al.*, "Semi-automated detection of anterior cruciate ligament injury from MRI," *Comput. Methods Programs Biomed.*, vol. 140, pp. 151–164, 2017.

[3] R. Mackenzie, C. R. Palmer, D. J. Lomas, and A. K. Dixon, "Magnetic resonance imaging of the knee: diagnostic performance studies," *Clin. Radiol.*, vol. 51, no. 4, pp. 251–257, 1996, doi: 10.1016/S0009-9260(96)80341-2.

[4] L. P. Cheung, K. C. P. Li, M. D. Hollett, A. G. Bergman, and R. J. Herfkens, "Meniscal tears of the knee: accuracy of detection with fast spin-echo MR imaging and arthroscopic correlation in 293 patients," *Radiology*, vol. 203, no. 2, pp. 508–512, 1997, doi: 10.1148/RADIOLOGY.203.2.9114113.

[5] K. Doi, "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential," *Comput. Med. Imaging Graph.*, vol. 31, no. 4–5, p. 198, 2007, doi: 10.1016/J.COMPMEDIMAG.2007.02.002.

[6] J. Elvenes, C. P. Jerome, O. Reikerås, and O. Johansen, "Magnetic resonance imaging as a screening procedure to avoid arthroscopy for meniscal tears," *Arch. Orthop. Trauma Surg.*, vol. 120, no. 1–2, pp. 14–16, 2000, doi: 10.1007/PL00021235.

[7] J. A. Feller and K. E. Webster, "Clinical value of magnetic resonance imaging of the knee," *ANZ J. Surg.*, vol. 71, no. 9, pp. 534–537, Sep. 2001, doi: 10.1046/J.1440-1622.2001.02183.X.

[8] S. H. Hong *et al.*, "Grading of anterior cruciate ligament injury. Diagnostic efficacy of oblique coronal magnetic resonance imaging of the knee," *J. Comput. Assist. Tomogr.*, vol. 27, no. 5, pp. 814–819, 2003, doi: 10.1097/00004728-200309000-00022.

[9] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[10] B. Rizk *et al.*, "Meniscal lesion detection and characterization in adult knee MRI: A deep learning model approach with external validation," *Phys. Med.*, vol. 83, pp. 64–71, Mar. 2021, doi: 10.1016/J.EJMP.2021.02.010.

[11] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/JAMA.2016.17216.

[12] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in CT images using deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Oct. 2016, vol. 2016-October, pp. 243–250, doi: 10.1109/IJCNN.2016.7727205.

[13] F. Liu *et al.*, "Deep Learning Approach for Evaluating Knee MR Images: Achieving High Diagnostic Performance for Cartilage Lesion Detection," *Radiology*, vol. 289, no. 1, pp. 160–169, Oct. 2018, doi: 10.1148/RADIOL.2018172986.

[14] E. G. McNally, K. N. Nasser, S. Dawson, and L. A. Goh, "Role of magnetic resonance imaging in the clinical management of the acutely locked knee," *Skeletal Radiol.*, vol. 31, no. 10, pp. 570–573, 2002, doi: 10.1007/S00256-002-0557-1.

[15] A. M. Naraghi and L. M. White, "Imaging of Athletic Injuries of Knee Ligaments and Menisci: Sports Imaging Series," *Radiology*, vol. 281, no. 1, pp. 23–40, Oct. 2016, doi: 10.1148/RADIOL.2016152320.

[16] L. Oakden-Rayner *et al.*, "Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, May 2017, doi: 10.1038/s41598-017-01931-w.

[17] E. H. G. Oei *et al.*, "MR imaging of the menisci and cruciate ligaments: a systematic review," *Radiology*, vol. 226, no. 3, pp. 837–848, Mar. 2003, doi: 10.1148/RADIOL.2263011892.

[18] A. Prasoon *et al.*, "Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network," in *Lect. Notes Comput. Sci.*, vol. 8150, no. Part 2, pp. 246–253, 2013, doi: 10.1007/978-3-642-40763-5-31.

[19] C. Rangger *et al.*, "Influence of magnetic resonance imaging on indications for arthroscopy of the knee," *Clin. Orthop. Relat. Res.*, vol. 330, pp. 133–142, 1996, doi: 10.1097/00003086-199609000-00016.

[20] K. Mead *et al.*, "MRI deep learning models for assisted diagnosis of knee pathologies: a systematic review," *Eur. Radiol.*, pp. 1–13, Oct. 2024, doi: 10.1007/S00330-024-11105-8/TABLES/4.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[22] D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., "Computer Vision – ECCV 2014," vol. 8689, 2014, doi: 10.1007/978-3-319-10590-1.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

[24] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end Learning of Deep Visual Representations for Image Retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, Oct. 2016, doi: 10.1007/s11263-017-1016-8.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1717–1724, Sep. 2014, doi: 10.1109/CVPR.2014.222.

[26] N. C. Nacey *et al.*, "Magnetic resonance imaging of the knee: An overview and update of conventional and state of the art imaging," *J. Magn. Reson. Imaging*, vol. 45, no. 5, pp. 1257–1275, May 2017, doi: 10.1002/JMRI.25620.

[27] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/NATURE21056.

[28] F. Yang *et al.*, "Automated deep learning-assisted early detection of radiation-induced temporal lobe injury on MRI: a multicenter retrospective analysis," *Eur. Radiol.*, pp. 1–13, Mar. 2025, doi: 10.1007/S00330-025-11470-Y/TABLES/4.

[29] B. P. Boden *et al.*, "Mechanisms of anterior cruciate ligament injury," *Orthopedics*, vol. 23, no. 6, pp. 573–578, 2000.

[30] M. Zhao *et al.*, "The accuracy of MRI in the diagnosis of anterior cruciate ligament injury," *Ann. Transl. Med.*, vol. 8, no. 24, p. 1657, 2020.

[31] B. Yu and W. E. Garrett, "Mechanisms of non-contact ACL injuries," *Br. J. Sports Med.*, vol. 41, suppl. 1, pp. i47–i51, 2007.

[32] Y. Lao *et al.*, "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*, vol. 98, no. 50, p. e18324, 2019.