Clinically Prioritized Attention-Based Fusion of Multi-Plane Knee MRI for Robust Injury Detection

Taira Kunitomi
Graduate School of FSE
Waseda University
Tokyo, Japan
k.taira634@fuji.waseda.jp

Taiga Hayami
Graduate School of FSE
Waseda University
Tokyo, Japan
hayatai17@fuji.waseda.jp

Hiroshi Watanabe
Graduate School of FSE
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Accurate diagnosis of anterior cruciate ligament (ACL) and meniscal tears using magnetic resonance imaging (MRI) is essential for timely and effective treatment. However, existing deep learning approaches often aggregate features uniformly across MRI slices and sequences, failing to account for their varying diagnostic relevance. To address these limitations, we propose a clinically informed attention-based fusion model that integrates coronal, sagittal, and axial MRI sequences using learnable fusion weights and slice-level attention pooling. This architecture better reflects the diagnostic workflow of radiologists, who interpret multi-plane images in context, thereby enhancing both classification accuracy and interpretability. Experimental results on the MRNet dataset demonstrate consistent performance improvements across all diagnostic tasks.**

***Index Terms*—Convolutional neural networks, deep neural networks, medical diagnostic imaging, magnetic resonance imaging.**

I. Introduction

Knee injuries such as ACL and meniscal tears are prevalent in sports and often require MRI for accurate diagnosis. Timely and reliable assessment is critical for proper treatment and return-to-play decisions. Deep learning has recently shown significant promise in medical imaging, particularly for automating the detection of musculoskeletal injuries from MRI scans [1]. One widely recognized baseline model in this domain is MRNet [2], developed at Stanford University. MRNet performs binary classification of knee injuries by analyzing three orthogonal MRI sequences: sagittal T2-weighted, coronal T1-weighted, and axial proton density (PD)-weighted images. Features are extracted from each sequence using separate convolutional neural networks (CNNs), and the final classification is computed by averaging these features. However, this uniform averaging approach overlooks the fact that different MRI sequences provide varying diagnostic insights. For example, T1- and T2-weighted images are particularly informative for detecting ACL and meniscal tears, whereas other sequences may be less informative. Additionally, not all slices within a sequence contribute equally to diagnosis — some slices may contain critical pathological features, while others may be irrelevant.

To address these limitations, we propose a multi-view MRI classification model that incorporates clinical knowledge into both spatial and sequence-level feature fusion. We introduce a slice-level attention mechanism that weights diagnostically salient slices, enabling the model to focus on relevant regions. A learnable sequence fusion module is also used, with weights initialized based on clinical priors to reflect the diagnostic value of each MRI view. This dual-attention strategy allows selective emphasis of meaningful features across both dimensions. Experiments on the MRNet dataset demonstrate consistent performance gains over the MRNet baseline across all diagnostic tasks.

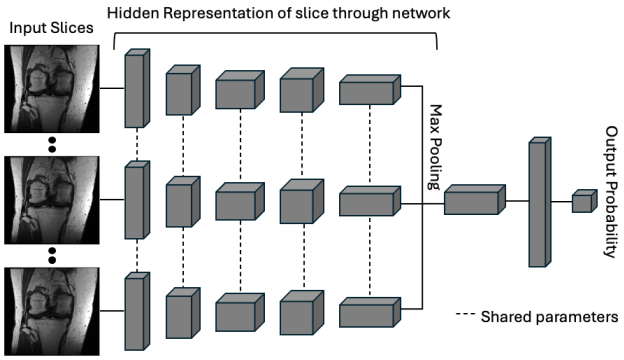


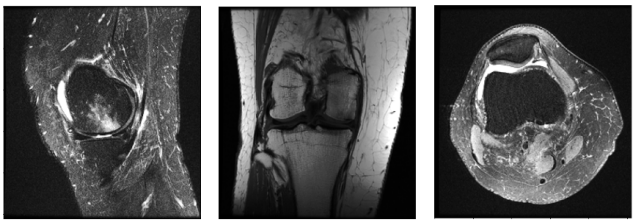
Fig. 1. The model architecture of MRNet [2], which processes sagittal, coronal, and axial MRI sequences independently using 2D CNNs. Features from each plane are pooled and averaged for classification, without modeling slice-wise or sequence-level diagnostic variation.

A. Deep Learning Approaches for MRI Diagnosis

Deep learning–based methods are increasingly effective for automating MRI interpretation. U-Net [3], an encoder–decoder segmentation model, has shown strong performance in tasks like brain tumor and musculoskeletal lesion detection. For classification, MRNet [2], developed at Stanford, is a key model for diagnosing knee injuries. It analyzes sagittal, coronal, and axial MRI views separately and fuses their outputs to predict conditions such as ACL and meniscal tears. It serves as a strong baseline for multi-view MRI classification tasks.

Figure 1 shows the MRNet architecture. Each view is processed by a 2D CNN to extract slice-level features, which are pooled and passed through fully connected layers. The final prediction is obtained by combining outputs from all views.

979-8-3315-2415-9/25/$31.00 ©2025 IEEE



(a) Sagittal T2 (b) Coronal T1 (b) Axial PD

Fig. 2. Three MRI sequences from the MRNet dataset: (a) sagittal T2-weighted, (b) coronal T1-weighted, and (c) axial PD-weighted views. These orthogonal planes are used as input for classification models in multi-view knee MRI analysis.

B. Public Knee MRI Datasets

Some publicly available knee MRI datasets, such as those based on functional MRI (fMRI), have been primarily developed for tasks in image reconstruction or signal modeling [4]. These datasets typically lack clinical annotations regarding injury presence or type, limiting their utility for supervised learning in diagnostic applications.

In contrast, the MRNet dataset [2] provides a large-scale, clinically annotated collection of knee MRI exams labeled for anterior cruciate ligament (ACL) tears, meniscal tears, and general abnormalities. Each exam includes three orthogonal MRI sequences—sagittal T2-weighted, coronal T1-weighted, and axial proton density (PD)-weighted—capturing complementary diagnostic information across anatomical planes.

The dataset encompasses a diverse patient population and imaging variations across sequences. Detailed statistics, including the number of exams, patient demographics, and sequence distribution, are provided in Table I. These characteristics make MRNet a suitable and widely adopted benchmark for multi-view knee MRI classification.

II. Proposed Method

We propose an enhanced model of MRNet by incorporating two key components: slice-level attention pooling and clinically guided sequence-level fusion. This design reflects the way radiologists focus on diagnostically important slices and MRI sequences depending on the task. The overall structure is illustrated in Figure 3 .

TABLE I
THE DEMOGRAPHIC INFORMATION OF THE DATASET

Statistics	Training	Validation
Number of exams	1130	120
Number of patients	1088	111
Number of female patients (%)	480 (42.5%)	50 (41.7%)
Age, mean (SD)	38.3 (16.9)	36.3 (16.9)

A. Slice-Level Attention Pooling

In conventional 3D medical image classification frameworks such as MRNet, a fixed aggregation method—typically max or average pooling—is employed to compress slice-wise features into a single sequence representation. This approach implicitly assumes that all slices contribute equally to the final prediction. However, in clinical practice, radiologists often pay particular attention to diagnostically informative slices—for instance, those capturing ligament tears or abnormal joint morphology—while ignoring irrelevant or redundant views.

To better reflect this clinical reasoning process, we introduce a slice-level attention pooling mechanism that adaptively weights each slice based on its latent diagnostic relevance. Formally, given a sequence of slice features $\{f_1, f_2, \ldots, f_N\} \in \mathbb{R}^{N \times d}$, where N is the number of slices and d is the feature dimension, we compute scalar attention scores $\{a_1, a_2, \ldots, a_N\}$ through a shallow feed-forward network with `tanh` activation:

$$a_i = \text{MLP}(f_i) = W_2 \cdot \tanh(W_1 f_i + b_1) + b_2. \tag{1}$$

These scores are then normalized across the slice dimension via the softmax function:

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^{N} \exp(a_j)}. \tag{2}$$

The final feature representation for the entire sequence is obtained as the weighted sum of slice features:

$$F = \sum_{i=1}^{N} \alpha_i f_i. \tag{3}$$

This attention mechanism allows the model to selectively emphasize slices containing pathology, such as torn ligaments or displaced menisci, while suppressing irrelevant frames. Unlike fixed pooling strategies, this learnable approach enhances sensitivity to subtle abnormalities and improves interpretability by explicitly revealing which slices contribute most to the model's prediction.

Empirically, we observe that attention weights often concentrate on clinically meaningful regions—for example, mid-sagittal slices showing the intercondylar notch in ACL tear detection—demonstrating that our model aligns well with diagnostic behavior in real-world radiology.

B. Softmax-weighted Sequence Fusion with Diagnostic Priors

While MRNet simply averages features from the sagittal, coronal, and axial MRI sequences, our model introduces learnable fusion weights that capture the task-specific diagnostic relevance of each sequence. These weights are initialized based on established clinical insights and are refined during training to reflect the modality-specific contributions most beneficial for each classification task.

Each MRI sequence provides complementary diagnostic information. T1-weighted images offer high-resolution visualization of bone and soft tissue anatomy, making them well

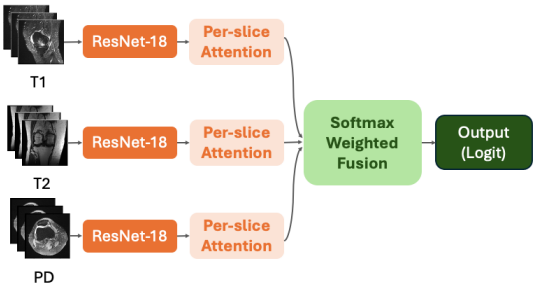


Fig. 3. Architecture of the proposed attention-based fusion model. Unlike MRNet, which uniformly averages features across slices and sequences, our model applies slice-level attention and clinically guided sequence fusion to selectively emphasize diagnostically relevant information across all views.

suited for evaluating osseous morphology and chronic structural abnormalities. T2-weighted images are sensitive to fluid accumulation and pathological changes such as inflammation, edema, or hematoma, which are often indicative of acute soft-tissue injuries including ACL or meniscal tears. PD-weighted images emphasize contrast at tissue interfaces, aiding in the detection of subtle boundary abnormalities such as meniscal fissures or cartilage delamination.

These distinct yet complementary roles are summarized in Table II. By incorporating this modality-specific understanding into the fusion process, our model selectively emphasizes diagnostically salient features across imaging planes, thereby aligning more closely with radiological interpretation strategies.

Based on these clinical characteristics, we initialize the fusion weights to prioritize sagittal and coronal sequences (T2- and T1-weighted) for tasks such as ACL and meniscal tear detection, where soft tissue integrity and joint structure are critical. For general abnormality detection—where diverse pathologies may appear across all planes—we adopt a uniform initialization to reflect the balanced diagnostic relevance of each sequence [5]–[7].

Let $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3 \in \mathbb{R}^d$ denote the feature vectors obtained from the three MRI sequences, and let $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{w}) \in \mathbb{R}^3$ be the learned fusion weights (normalized such that $\sum_i \alpha_i = 1$). The final fused feature vector $\mathbf{f}_{\text{fused}}$ is computed as:

$$\mathbf{f}_{\text{fused}} = \alpha_1 \cdot \mathbf{f}_1 + \alpha_2 \cdot \mathbf{f}_2 + \alpha_3 \cdot \mathbf{f}_3. \tag{4}$$

In practice, the initial values of $\mathbf{w}$ were set to reflect clinical preferences:

- For ACL and meniscal tear detection: The initial weights were set to $(0.45, 0.45, 0.10)$, reflecting the dominance of coronal and sagittal views.
- For abnormality detection: A uniform initialization of $(0.33, 0.33, 0.33)$ was used to reflect equal importance across all sequences.

TABLE II
DIAGNOSTIC CHARACTERISTICS OF DIFFERENT MRI SEQUENCES

MRI Sequence	Diagnostic Strengths
T1-weighted	Clearly shows anatomical structures of bones and soft tissues. Useful for identifying bone morphology and structural abnormalities.
T2-weighted	Highlights tissues with high water content (e.g., inflammation, swelling, hematoma). Effective for detecting abnormalities in soft tissues.
PD-weighted	Emphasizes contrast between bones and soft tissues. Suitable for detecting subtle boundary abnormalities.

III. EXPERIMENT

To validate the effectiveness of our proposed method, we conducted a rigorous comparative evaluation by re-implementing the original MRNet [2] under identical training conditions, including optimizer settings, batch size, data augmentation, and cross-validation splits. This ensured a fair and controlled comparison, isolating the impact of architectural differences.

Table 2 presents the AUC (Area Under the ROC Curve) scores for both models across three binary classification tasks: ACL tear, meniscal tear, and general abnormality detection. AUC is a widely accepted metric in medical imaging that reflects the model's ability to discriminate between positive and negative cases across varying thresholds.

Across all tasks, our model consistently outperformed the MRNet baseline. This performance gain is attributed to two key design components. First, the slice-level attention module explicitly weights each slice based on its task-specific importance, allowing the model to focus on diagnostically relevant regions and ignore irrelevant information. Second, the sequence-level fusion module integrates anatomical planes using learnable weights initialized from clinical priors, capturing the differing diagnostic value of each sequence in a structured and interpretable manner.

The most significant improvement was observed in ACL tear detection, where our model achieved an AUC of 0.910 compared to 0.852 with the baseline. This task requires the detection of subtle discontinuities in ligament structure, which are often visible only in a few key sagittal slices. The slice attention module effectively emphasized frames such as the intercondylar notch and ligament attachment regions, enabling robust classification even in the presence of noisy or redundant slices.

In the meniscal tear and abnormality detection tasks, our model also demonstrated notable gains in AUC. These results confirm that combining spatial attention with clinically guided fusion improves robustness and generalizability across diverse structural and pathological contexts.

Overall, our model not only improves classification accuracy over the MRNet baseline but also closely mimics the diagnostic strategies employed by radiologists. In clinical workflows, physicians systematically scroll through sequences to identify pathological patterns and dynamically prioritize specific views

TABLE III
AUC SCORES OF EACH MODEL ACROSS DIAGNOSTIC TASKS

Model	ACL Tear	Meniscus Tear	Abnormality
MRNet (Baseline)	0.852	0.856	0.847
Ours (Proposed)	**0.910**	**0.899**	**0.877**

depending on the suspected condition. By modeling both slice-level and sequence-level relevance in an end-to-end manner, our architecture bridges algorithmic prediction with domain-expert reasoning.

A. Ablation Study

To further assess the contribution of each MRI sequence to model performance, we conducted a targeted ablation study focusing on the ACL tear classification task. Specifically, we compared the proposed learnable fusion strategy with three fixed-weight configurations, each constructed by omitting one sequence through an assigned weight of zero. Table IV reports the AUC results for each configuration.

Among the fixed settings, the configuration with equal weighting of T1- and T2-weighted sequences, while excluding PD, produced the best AUC (0.878). Nevertheless, this configuration remained inferior to the proposed adaptive fusion method, which achieved an AUC of 0.910. This performance gap underscores the limitations of manually fixed fusion weights and highlights the importance of data-driven adjustment mechanisms in capturing diagnostic synergies across sequences.

Performance consistently deteriorated when any single modality was excluded, indicating that all three sequences—T1, T2, and PD—contribute complementary and non-redundant diagnostic information. T1- and T2-weighted images provide critical anatomical and pathological detail, particularly concerning ligament integrity, edema, and joint effusion, which are highly relevant for ACL tear detection. Although PD-weighted images typically receive lower clinical emphasis, their ability to enhance soft tissue–bone interface contrast appears to support the discrimination of subtle abnormalities.

The reduced performance of the (0.5, 0.5, 0.0) configuration suggests that even minimally weighted sequences can supply auxiliary cues that aid the model in decision-making. These results validate our design choice to employ an attention-based fusion strategy that dynamically learns task-specific weighting, enabling optimal integration of multi-sequence inputs. Such adaptivity is particularly beneficial for complex diagnostic tasks where subtle patterns may span across multiple anatomical planes and contrast mechanisms.

IV. CONCLUSION

We proposed an MRI classification model that integrates slice-level attention and clinically guided sequence fusion to enhance the diagnostic accuracy of knee injury detection. By incorporating clinical insights—such as prioritizing T1- and T2-weighted images for ACL and meniscal tear diagnosis—our model effectively focuses on relevant features across both spatial and anatomical dimensions. Experiments on the MRNet dataset demonstrated consistent improvements over the baseline across all classification tasks.

TABLE IV
ABLATION STUDY ON ACL TEAR DETECTION (AUC)

Fusion Weights (T1, T2, PD)	AUC	Δ vs. Ours
(0.5, 0.5, 0.0)	0.878	-0.032
(0.0, 0.5, 0.5)	0.864	-0.046
(0.5, 0.0, 0.5)	0.867	-0.043
Learnable (Ours)	**0.910**	—

The architecture reflects radiological reasoning by dynamically weighting slices based on diagnostic contribution and learning task-specific fusion weights that capture the varying clinical importance of each sequence. This design aligns more closely with expert-level interpretation, improving both performance and transparency.

For future work, we aim to enhance the interpretability of our model by generating spatially resolved attention visualizations, such as slice-wise heatmaps and modality-wise saliency maps, to provide clinicians with intuitive insights into model decisions. We also plan to address a current limitation—namely, the absence of inter-slice contextual modeling—by incorporating sequence-aware modules that capture anatomical continuity across adjacent slices, which is important for identifying spatially distributed pathologies.

In addition, we intend to validate the robustness of our approach using external datasets acquired under diverse imaging protocols, and to extend its applicability to other musculoskeletal regions, such as the shoulder or spine. Ultimately, we aim to build a clinically deployable and interpretable diagnostic support system adaptable to real-world medical imaging settings.

REFERENCES

[1] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[2] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLOS Medicine*, vol. 15, no. 11, pp. e1002699, 2018.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[4] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, "fastMRI: An open dataset and benchmarks for accelerated MRI," arXiv preprint arXiv:1811.08839, 2018.

[5] M. P. Recht, R. E. Kramer, P. J. Marcelis, and R. A. Daffner, "MR imaging of the knee: current status and future directions," *American Journal of Roentgenology*, vol. 167, no. 3, pp. 593–600, Sep. 1996.

[6] Haaga, J. R., Lanzieri, C. F., Gilkeson, R. C. (2008). *CT and MRI of the Whole Body* (Vol. 1–2). Elsevier Health Sciences.

[7] McRobbie, D. W., Moore, E. A., Graves, M. J., Prince, M. R. (2017). *MRI from Picture to Proton* (3rd ed.). Cambridge University Press.