MENIMV: A MULTI-VIEW BENCHMARK FOR MENISCUS INJURY SEVERITY GRADING

Shurui Xu[1], Siqi Yang[2], Jiapin Ren[3], Zhong Cao[4], Hongwei Yang[5], Mengzhen Fan[6], Yuyu Sun[5], Shuyan Li[1]

[1] Electronics, Electrical Engineering and Computer Science, Queen's University Belfast
[2] Department of Radiology, Affiliated Hospital 2 of Nantong University
[3] Xuhai College, China University of Mining and Technology
[4] Heidelberg Institute of Global Health, Faculty of Medicine and University Hospital, Heidelberg University
[5] Department of Orthopaedics, Affiliated Nantong Hospital 3 of Nantong University
[6] HSBC Business School, Peking University

ABSTRACT

Precise grading of meniscal horn tears is critical in knee injury diagnosis but remains underexplored in automated MRI analysis. Existing methods often rely on coarse study-level labels or binary classification, lacking localization and severity information. In this paper, we introduce MeniMV, a multi-view benchmark dataset specifically designed for horn-specific meniscus injury grading. MeniMV comprises 3,000 annotated knee MRI exams from 750 patients across three medical centers, providing 6,000 co-registered sagittal and coronal images. Each exam is meticulously annotated with four-tier (grade 0–3) severity labels for both anterior and posterior meniscal horns, verified by chief orthopedic physicians. Notably, MeniMV offers more than double the pathology-labeled data volume of prior datasets while uniquely capturing the dual-view diagnostic context essential in clinical practice. To demonstrate the utility of MeniMV, we benchmark multiple state-of-the-art CNN and Transformer-based models. Our extensive experiments establish strong baselines and highlight challenges in severity grading, providing a valuable foundation for future research in automated musculoskeletal imaging.

Index Terms— Meniscus Injury Grading, Knee MRI, Deep Learning, Medical Image Analysis, Dataset.

1. INTRODUCTION

The knee meniscus plays a vital role in knee function by redistributing force, absorbing impact and stabilising movement between the tibia and femur, thus helping to protect the articular cartilage from early degeneration. Tears of the anterior or posterior horn of the meniscus are one of the most common reasons for sports medicine referrals, accounting for one-third of knee arthroscopy procedures each year. Magnetic Resonance Imaging (MRI) is currently considered the gold standard for diagnosing knee injuries due to its excellent soft tissue contrast and ability to capture images from multiple angles. However, interpreting a full set of MRIs requires
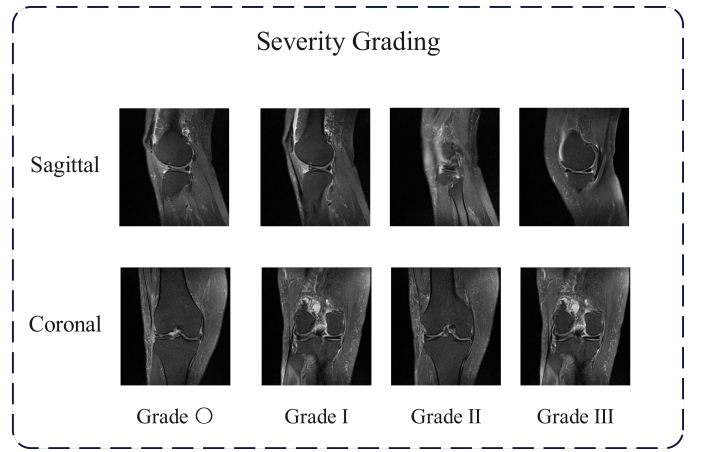

Fig. 1. Example from the MeniMV dataset. Each case includes two views—sagittal and coronal—with anterior and posterior meniscal horns annotated across four injury grades.

significant expertise and time, which can lead to inconsistencies between different interpreters, increased costs, and delays in treatment decisions. Automated systems powered by deep learning have the potential to provide more consistent and objective assessments, allowing for more accurate treatment.

Deep learning has demonstrated significant potential in automating the diagnosis of knee injuries. Early work primarily focused on cartilage segmentation and lesion detection using convolutional neural networks (CNNs) [1, 2, 3]. A major advancement came with the introduction of MR-Net [4], which utilized multi-view MRI scans (sagittal, coronal, axial) to classify entire knee exams as either "normal" or "abnormal." While this demonstrated the feasibility of deep learning for knee assessment, MRNet was limited by its reliance on coarse, study-level labels (e.g., "abnormal," "ACL tear," "meniscal tear"), lacking finer-grained information such as tear location or severity. Subsequent efforts like the fastMRI [5] addressed data scarcity by releasing large

volumes of raw k-space data and images. This effort was later extended by fastMRI+ [6], which included pathology labels. However, annotations are restricted to the coronal plane, crucially treating meniscal tears as a binary classification problem (i.e. presence or absence), overlooking the important aspect of injury grading.

To address these limitations, we present MeniMV, a multi-view benchmark dataset specifically designed for meniscus injury grading. The dataset comprises MRI scans from 750 patients collected across three medical centers, with all images independently validated by a panel of expert clinicians and radiologists to ensure high-quality and reliable annotations. Each examination includes horn-specific, four-tier tear labels (grades 0–3), capturing the severity of damage in the anterior or posterior meniscal horn. As a result, MeniMV contains 3,000 annotated knee examinations, each with paired sagittal and coronal T2-weighted MRI series, totaling 6,000 co-registered scans. As illustrated in Figure 2, MeniMV provides annotations for both sagittal and coronal views, covering the anterior and posterior meniscal horns across four injury grades. Notably, MeniMV not only more than doubles the volume of pathology-labeled data available in MRNet and fastMRI combined, but also uniquely offers the clinically essential dual-view context routinely used by radiologists during diagnosis. Leveraging this comprehensive dataset, we formulate meniscal injury assessment as a dual-head classification task. We further perform extensive experiments using standard classification backbones to evaluate their effectiveness in fine-grained severity grading on the MeniMV benchmark.

Our main contributions are summarized as follows:

- We released MeniMV, a multi-view knee MRI dataset with sagittal and coronal images, annotated for anterior/posterior horn tears across four severity levels (0–3), offering a high-quality benchmark for the research community. As far as we know, MeniMV is the first resource to provide horn-specific severity grades on paired sagittal-coronal planes.

- We benchmarked several popular backbone architectures for automated injury severity grading on the proposed dataset, including both CNN-based and Transformer-based frameworks.

2. MENIMV DATASET

2.1. Data Collection and Labeling

We retrospectively curated 750 knee MRI studies from three hospitals (405 females, 345 males; mean age 55.6 ± 12.7 years; range 14–82) under a unified protocol. To enhance sensitivity to intrameniscal signal change, only fat-suppressed T2-weighted (T2WI-FS) sequences were included (reported sensitivity ~94% for meniscal lesions [7]). All

DICOMs were de-identified with a standardized pipeline (DicomAnonymizer) and metadata archived for reproducibility [8, 9]. A panel of six senior orthopedic clinicians (each
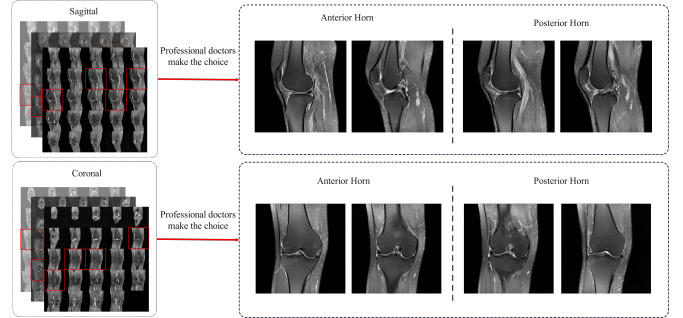

Fig. 2. The selection procedure of MeniMV. The four most diagnostically relevant slices from both sagittal and coronal MRI series are selected by chief orthopedic physicians.

$\geq$10 years' experience) independently screened each study focusing on the anterior and posterior horns. For each horn, the most informative sagittal–coronal slice pair was retained; motion-degraded images were excluded. Severity was assigned per image using the Stoller 0–III (0–3) scale with double validation. In total, MeniMV contains 3,000 expert-annotated images (1,500 sagittal–coronal pairs; two pairs per patient).

2.2. Dataset Properties and Context

MeniMV provides four images per patient (two sagittal and two coronal; one pair per horn), enabling consistent multi-view learning under a single labeling scheme. Across the cohort (age 14–82; mean 55.6 ± 12.7), Grade 0 cases are more frequent at younger ages and decline with age; Grades 1–2 are relatively uniform; Grade 3 rises notably after 50 and concentrates in the 60–70 range, consistent with age-related degeneration.

The grade distribution is: 1,331 Grade 0 (44.4%), 502 Grade 1 (16.7%), 306 Grade 2 (10.2%), and 861 Grade 3 (28.7%). The scarcity of Grades 1–2 motivates class-balanced training objectives and ordinal-aware evaluation (e.g., quadratic weighted Cohen's κ, grade-wise MAE).

Compared with prior knee MRI resources, MRNet [4] aggregates 1,370 single-center studies with mixed sequences and single-reader binary labels, and fastMRI targets reconstruction without diagnostic labels. In contrast, MeniMV is multi-center (three hospitals), standardized to T2WI-FS, and supplies.

3. MULTI-VIEW MENISCUS INJURY GRADING

3.1. Task Definition

We address meniscal injury severity classification from paired MRI views (sagittal and coronal), assigning each case to one of four grades: 0 (normal), 1 (mild), 2 (moderate), or 3 (severe), with position-specific categorization for anterior and posterior horns. Given a pair of views $\boldsymbol{x} = \{x_{\text{sagittal}}, x_{\text{coronal}}\}$ and a slice-level anatomical indicator $p \in \{0, 1\}$ (anterior = 1, posterior = 0), the objective is to predict an injury grade $y \in \mathcal{Y} = \{0, 1, 2, 3\}$ at the specified position. We formalize this as a mapping $f_\theta : (\boldsymbol{x}, p) \mapsto \mathcal{Y}$, where θ denotes the model parameters comprising a shared backbone feature extractor and two position-specific classification heads. The inclusion of p conditions the prediction on the target horn, enabling the model to treat anterior and posterior regions independently in light of their distinct clinical failure patterns.

3.2. Multi-view Encoder

We utilize a range of architectures as backbone feature extractors, including ResNet-50 [10, 11], DenseNet-121 [12, 13, 14], EfficientNet-B0 [15, 16, 17], and **Vision Transformer (ViT-B/16)** [18, 19]. We train our model from scratch using the proposed dataset. Subsequently, we extract feature embeddings from both the sagittal and coronal views, denoted as $\epsilon_{\text{sagittal}}$ and $\epsilon_{\text{coronal}}$, respectively.

While there are multiple ways to fuse these features, we empirically found that concatenation yields the best performance. Then we get the fusion feature as below:

$$\mathbf{f}_{\text{fusion}} = [\epsilon_{\text{sagittal}} \parallel \epsilon_{\text{coronal}}], \qquad (1)$$

where $\parallel$ represents concatenation along the feature dimension. This fusion enables the model to leverage complementary anatomical details from both the sagittal and coronal planes, ensuring a more comprehensive understanding of the meniscal injury.

3.3. Classification Head

After feature fusion, the concatenated feature vector $\mathbf{f}_{\text{fusion}}$ is forwarded through a classification head to predict the injury severity. We introduce a dual-classification head for grading in the anterior and posterior regions of the meniscus separately. Specifically, the model uses two separate fully connected layers:

$$P = \begin{cases} \text{softmax}\,(W_a\,\mathbf{f}_{\text{fusion}} + b_a), & \text{if } p = 1, \\ \text{softmax}\,(W_p\,\mathbf{f}_{\text{fusion}} + b_p), & \text{if } p = 0, \end{cases} \qquad (2)$$

where W_a, W_p and b_a, b_p are the parameters for the anterior and posterior classifiers, respectively. This ensures that the model learns region-specific decision boundaries.

3.4. Loss Function

We optimize a hybrid objective that couples class-imbalance–aware Focal Loss with a cross-view feature-alignment term. Given a minibatch of N labeled samples, the multi-class Focal Loss is

$$L_{\text{focal}} = -\sum_{i=1}^{N} \lambda_{y_i} \left(1 - P_{i,y_i}\right)^\gamma \log\!\left(P_{i,y_i}\right), \qquad (3)$$

where $P_{i,y} = \text{softmax}(\mathbf{z}_i)_y$ is the predicted probability for class y from logits $\mathbf{z}_i$, $\lambda_y \geq 0$ are class-specific weights to counter class imbalance, and $\gamma \geq 0$ controls the down-weighting of well-classified examples (larger γ emphasizes harder cases) . To promote coherence between sagittal and coronal representations, we further penalize discrepancies between their latent features via an MSE term:

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{f}_{\text{sagittal}}^{(i)} - \mathbf{f}_{\text{coronal}}^{(i)} \right\|_2^2. \qquad (4)$$

The overall training objective is a weighted sum of the two components,

$$L_{\text{total}} = \alpha\, L_{\text{focal}} + \beta\, L_{\text{MSE}}, \qquad (5)$$

with $\alpha, \beta \geq 0$ controlling the trade-off between label-distribution sensitivity and cross-view alignment.

4. EXPERIMENTAL ANALYSIS

4.1. Backbone and Method Analysis

Table 1 contrasts generic CNNs, domain-specific methods, and modern Transformers. Among generic CNNs trained from scratch, DenseNet-121 yields the strongest results (67.83% Acc). However, domain-specific methods (MRNet [4], 2.5D ResNet Fusion [11], and DeepKnee [3]) significantly outperform these generic baselines; specifically, DeepKnee achieves 74.22% Acc, surpassing DenseNet-121 by a notable margin (+6.39%) due to its specialized ROI localization and fusion design. Nevertheless, pretrained Transformers demonstrate the superior discrimination: the Swin-UNETR encoder attains the state-of-the-art performance (76.92% Acc, 0.6790 Macro-F1), outperforming even the specialized DeepKnee method. This progression suggests that while domain-specific engineering is effective, the large-scale representation learning capabilities of modern Transformers provide the most robust foundation for meniscal injury grading.

4.2. Fusion Strategy Analysis

To assess the effect of feature fusion in our dual-view framework, we evaluate three canonical strategies applied after independently encoding the sagittal and coronal views with a

Table 1. Performance comparison of generic backbones, domain-specific methods, and modern pretrained architectures.

Method / Backbone	Acc (%)	Macro-F1	MAE
Generic CNNs (Scratch)			
ResNet-50	57.67	0.4667	0.91
ResNeXt-50 (32×4d)	59.92	0.4874	0.88
EfficientNet-B0	55.00	0.4362	0.95
ShuffleNet-v2	54.83	0.3949	0.99
MobileNet-v2	53.89	0.3825	1.01
ConvNeXt-T	63.78	0.5311	0.81
DenseNet-121	67.83	0.5730	0.74
ViT-B/16 (scratch)	50.33	0.2907	1.10
Domain-Specific Methods			
MRNet (Adapted)	69.45	0.5912	0.68
2.5D ResNet Fusion	71.80	0.6150	0.63
DeepKnee (Adapted)	74.22	0.6385	0.58
Modern Pretrained Architectures			
ConvNeXt-V2-B (pt)	74.85	0.6550	0.55
Swin-T (pt)	75.42	0.6610	0.54
Swin-B (pt)	76.38	0.6730	0.52
ViT-B/16 (MAE/DINO pt)	73.11	0.6400	0.57
Swin-UNETR encoder (pt)	**76.92**	**0.6790**	**0.51**

shared DenseNet-121 backbone: additive fusion, which aggregates the two feature vectors via element-wise summation; concatenation, which joins the representations channel-wise; and attention-based fusion, which introduces a gated self-attention module to adaptively reweight and combine the sagittal and coronal features.

As shown in Table 2, the concatenation strategy outperforms both additive and attention-based fusion. Additive fusion offers simplicity but assumes equal importance and alignment between the sagittal and coronal features, which may not hold given their distinct anatomical information. Attention-based fusion improves over simple addition by learning dynamic weighting across modalities, but introduces extra parameters and computation, with only modest performance gains.

Table 2. Comparison of feature fusion strategies.

Fusion Strategy	Accuracy (%)	Macro-F1
Additive Fusion	70.02	0.6021
Attention-based Fusion	72.46	0.6228
Concatenation	**73.63**	**0.6347**

4.3. Cross-center Robustness (LOCO)

To assess out-of-distribution robustness, we adopted a leave-one-center-out (LOCO) protocol. As shown in Table 3, domain-specific methods generalize better than standard CNNs; specifically, DeepKnee achieves an average Macro-F1 of 0.625 and Acc of 72.5%, demonstrating that incorporating anatomical priors helps mitigate scanner variations. However, modern pretrained Transformers generalize substantially better than both generic and domain-specific baselines. Swin-UNETR achieves the highest average Macro-F1 of 0.641 and reduces the MAE to 0.56, with lower variance across centers. This indicates that while domain-specific design improves robustness, the self-attention mechanism combined with large-scale pretraining offers the most effective defense against cross-center protocol shifts.

Table 3. LOCO performance comparison including domain-specific baselines. Macro-F1 per target center; Acc/MAE averaged across centers.

Model	MF1-A	MF1-B	MF1-C	MF1-Avg	Acc-Avg (%)	MAE-Avg
DenseNet-121	0.551	0.563	0.532	0.549	65.4	0.79
ConvNeXt-T	0.574	0.586	0.557	0.572	61.9	0.83
MRNet (Adapted)	0.578	0.589	0.558	0.575	68.5	0.69
2.5D ResNet Fusion	0.601	0.612	0.581	0.598	70.2	0.65
DeepKnee (Adapted)	0.628	0.639	0.608	0.625	72.5	0.60
Swin-T (pt)	0.620	0.631	0.602	0.618	72.1	0.59
Swin-B (pt)	0.636	0.644	0.615	0.632	73.0	0.57
Swin-UNETR encoder (pt)	**0.645**	**0.653**	**0.624**	**0.641**	**73.6**	**0.56**

4.4. Demographic Robustness

Finally, we examine sex- and age-stratified results for the best model (Table 4). Gaps are small but non-negligible: male vs. female Macro-F1 differs by 0.018 with a 0.04 MAE gap; across age bands, the Macro-F1 gap is 0.025 and MAE gap is 0.06. These patterns are consistent with an age-dependent prevalence of severe degeneration and suggest that mild post-hoc calibration or reweighting could further narrow disparities without sacrificing overall accuracy.

Table 4. Demographic robustness for the best model (Swin-UNETR encoder).

Subgroup	Macro-F1	MAE	ΔF1	ΔMAE
Male	0.648	0.53	0.018	0.04
Female	0.666	0.49		
Age < 40	0.642	0.56		
Age 40–59	0.667	0.50	0.025	0.06
Age ≥ 60	0.655	0.52		

5. CONCLUSION

Our work introduces MeniMV, a large-scale, dual-view knee MRI dataset featuring horn-specific, four-level severity annotations, curated by experienced musculoskeletal radiologists. We benchmark several widely used backbone architectures—including both CNN-based and Transformer-based models—for automated injury grading on this dataset. Looking ahead, we plan to explore cross-modal integration with clinical metadata and extend our framework to other joint structures.

6. REFERENCES

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 246–253.

[3] Fang Liu, Zhaoye Zhou, Alexey Samsonov, Donna Blankenbaker, Will Larison, Andrew Kanarek, Kevin Lian, Shivkumar Kambhampati, and Richard Kijowski, "Deep learning approach for evaluating knee mr images: achieving high diagnostic performance for cartilage lesion detection," *Radiology*, vol. 289, 2018.

[4] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet," *PLoS medicine*, vol. 15, no. 11, pp. e1002699, 2018.

[5] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al., "fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning," *Radiology: Artificial Intelligence*, vol. 2, no. 1, pp. e190007, 2020.

[6] Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W Lui, Michael S Hansen, and Matthew P Lungren, "fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data," *Scientific Data*, vol. 9, no. 1, pp. 152, 2022.

[7] A HAWANA MAGED and H SOLIMAN HAZEM, "Comparison of volumetric fat suppressed proton density and standard routine mr sequences in the evaluation of meniscal tears," *The Medical Journal of Cairo University*, vol. 88, no. March, pp. 755–760, 2020.

[8] Fred Prior, Kirk Smith, Ashish Sharma, Justin Kirby, Lawrence Tarbox, Ken Clark, William Bennett, Tracy Nolan, and John Freymann, "The public cancer radiology imaging collections of the cancer imaging archive," *Scientific data*, vol. 4, no. 1, pp. 1–7, 2017.

[9] David A Clunie, "Dicom structured reporting and cancer clinical trials results," *Cancer informatics*, vol. 4, pp. CIN–S37032, 2007.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Joseph Antony, Kevin McGuinness, Noel E O'Connor, and Kieran Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 1195–1200.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] Ke Zhang, Yurong Guo, Xinsheng Wang, Jinsha Yuan, and Qiaolin Ding, "Multiple feature reweight densenet for image classification," *IEEE access*, vol. 7, pp. 9872–9880, 2019.

[14] Zhicheng Zhang, Xiaokun Liang, Xu Dong, Yaoqin Xie, and Guohua Cao, "A sparse-view ct reconstruction method based on combination of densenet and deconvolution," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1407–1417, 2018.

[15] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[16] Gonçalo Marques, Deevyankar Agarwal, and Isabel De la Torre Díez, "Automated medical diagnosis of covid-19 through efficientnet convolutional neural network," *Applied soft computing*, vol. 96, pp. 106691, 2020.

[17] Daniel L Franklin, Tara Pattilachan, and Anthony Magliocco, "Imaging based egfr mutation subtype classification using efficientnet," *Cancer Research*, vol. 82, no. 12_Supplement, pp. 5048–5048, 2022.

[18] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.