PMMG: CLINICAL PRIOR-GUIDED MULTIMODAL MULTI-TASK CARTILAGE DEFECT GRADING MODEL FOR PATELLOFEMORAL OSTEOARTHRITIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Patellofemoral osteoarthritis (PFOA) is a prevalent degenerative joint disease and a major subtype of knee osteoarthritis, characterized by irreversible cartilage defects that often present as irregular, multi-focal lesions across the patella and femur. This complexity demands early, fine-grained assessment to guide personalized, stepwise interventions. Multi-sequence Magnetic Resonance Imaging (MRI) offers superior soft-tissue visualization, yet interpreting the complex cartilage damage requires extensive radiological expertise, which makes reads subjective and time-consuming. This often leads to diagnostic delays and missed opportunities for early treatment. Existing methods are limited in that cartilage grading remains binary, without providing separate multi-level assessments for patellar and femoral cartilages. To address this, we propose **PMMG**, a clinical **P**rior-guided **M**ultimodal **M**ulti-task cartilage defect **G**rading model. Guided by clinical priors, it integrates multimodal information via multi-sequence collaborative learning with a level-wise attention. Additionally, a Cartilage-Focused Diffusion data augmentation strategy mitigates inherent class imbalance through synthetic data generation. The framework jointly performs fine-grained grading for both patellar and femoral cartilage. We construct the first large-scale multi-center multimodal PFOA dataset comprising 1,173 individuals. Extensive experiments demonstrate that PMMG (Accuracy: 0.739; AUC: 0.891) significantly outperforms state-of-the-art methods, achieves grading accuracy approaching that of senior radiologists (Accuracy: 0.767) while substantially improving the diagnostic performance of junior radiologists when used as an assistive tool, and achieves promising performance on an independent external test cohort (Accuracy: 0.705). This unified framework provides a comprehensive clinical assessment, enabling earlier and more precise stepwise interventions. Code is available at `Here`.

1 INTRODUCTION

Patellofemoral osteoarthritis (PFOA) is a prevalent degenerative joint disease and a major subtype of knee osteoarthritis (KOA), primarily characterized by irreversible cartilage defects in its early stages Hill et al. (2025). Globally, approximately 64–69% of patients with chronic knee pain exhibit radiologically visible PFOA, which is associated with a high rate of disability and has become one of the most common chronic conditions in older adults Collaborators et al. (2017). Since the patellofemoral joint is typically regarded as the initial site of knee degeneration Lankhorst et al. (2017); Stefanik et al. (2016), accurate identification of PFOA cartilage defect progression is critical for early intervention for knee pain Macri (2017).

In clinical practice, multi-sequence Magnetic Resonance Imaging (MRI) is the primary modality used to assess the PFOA cartilage defect progression. As shown in the left part of **Figure 1**, cartilage defects may be present at different grades in the patella and femur in the same patient, with lesions often being irregular and locally distributed. This requires substantial radiological expertise to interpret, rendering the process subjective and inefficient Mead et al. (2025). Moreover, the pathophysiological mechanisms underlying PFOA remain incompletely understood, and current diagnostic and prognostic methods often fail to capture the early and complex characteristics of disease progression Fu et al. (2025). Consequently, PFOA is frequently misdiagnosed or over-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

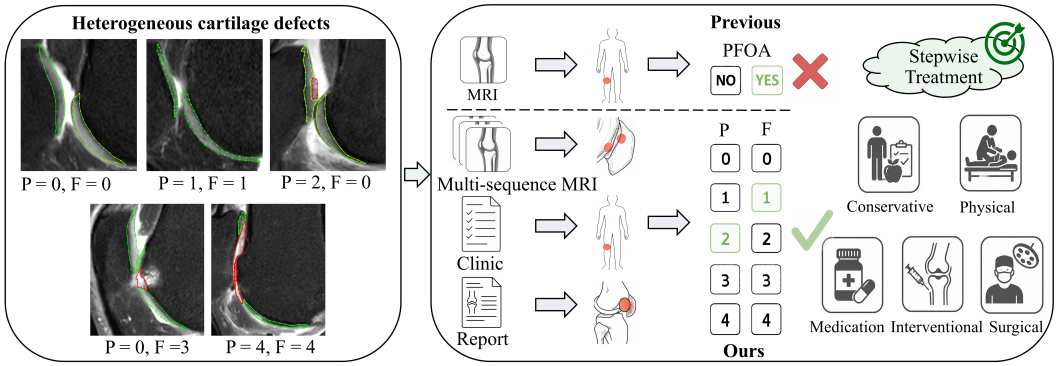

Figure 1: Workflow comparison of previous methods vs. the proposed PMMG framework.

looked, resulting in delayed interventions and potential joint necrosis, which may ultimately require patellofemoral joint replacement surgery, an invasive procedure associated with long-term complications Yeoh et al. (2021); Lv et al. (2025). It is essential to recognize that different stages of PFOA cartilage defects require distinct early intervention strategies Liu et al. (2018); Guo et al. (2024). Therefore, the development of a method to automatically identify the stages of PFOA cartilage defects progression is crucial for personalized, stepwise treatment and the prevention of invasive surgery. Such a method can enhance diagnostic efficiency, reduce subjective bias, and enable early intervention, ultimately improving disease trajectories and optimizing the clinical process from diagnosis to treatment Fuggle et al. (2025); Wang et al. (2022); Veronese et al. (2022).

Deep learning has garnered significant attention for its ability to autonomously learn representative features from medical images. Its superior performance in knee lesion analysis has been acknowledged Vuskov et al. (2026). However, existing studies face several limitations. On one hand, many studies focus on the overall progression of KOA, which encompasses both PFOA and femorotibial osteoarthritis, only implicitly including PFOA, not fine-grained distinguishing PFOA and femorotibial osteoarthritis, which restricts the ability to identify the early stages of PFOA progression Singh et al. (2019); Zhang et al. (2023); Roy et al. (2025); Huo et al. (2022). On the other hand, other studies are limited to coarse-grained classification that merely determines the presence or absence of PFOA based on single modality, without providing fine-grained assessments of cartilage damage within joints, which directly prevents accurate stepwise intervention for patients. Bayramoglu et al. (2021; 2022); Zhang et al. (2022); Wang et al. (2023).

In this paper, we present the Clinical Prior-Guided Multimodal Multi-task Cartilage Defect Grading Model for PFOA (PMMG). As shown in **Figure 1**, unlike previous studies limited to binary PFOA diagnosis using single-modality data, our approach leverages multimodal data from three levels to perform fine-grained grading of cartilage defects on both the patellar cartilage (P) and femoral cartilage (F) simultaneously. This advancement provides valuable guidance for stepwise treatment in clinical practice. To the best of our knowledge, this is the first automated grading method to characterize the complex disease progression of cartilage defects in PFOA. The contributions of this paper are three-fold: **(1)** We introduce PMMG that integrates multimodal information via multi-sequence collaborative learning with a level-wise attention mechanism, and we construct the first multimodal PFOA dataset, comprising 1,173 individuals from two medical centers. **(2)** Additionally, to mitigate the class imbalance inherent in real-world cohorts, we propose Cartilage-Focused Diffusion data augmentation, which addresses class imbalance via synthetic data generation. **(3)** Extensive experiments demonstrate that PMMG not only outperforms state-of-the-art methods and generalizes well on an independent external test cohort, but also achieves grading accuracy superior to that of junior radiologists, and when used as an assistive tool, significantly boosts their diagnostic performance, underscoring its clinical practicality.

2 Related Works

Deep learning for knee disease analysis. Deep learning has been widely applied to knee MRI analysis. For comprehensive classification, the CoPAS model with co-plane attention Qiu et al. (2024)

addresses twelve abnormality types, the KAMRNet system Xie et al. (2025) performs multi-center detection of nine conditions, the 3D slice transformer network Vuskov et al. (2026) covers 23 knee pathologies, and the attention-enhanced 3D residual network Ashames et al. (2025) is employed for classification tasks. On the modeling side, an adversarial data augmentation method Liu et al. (2025) and a Memory-Guided Transformer Huang et al. (2025) improve performance through data expansion and multi-level feature extraction, respectively. For KOA progression prediction, a multimodal Transformer fusion approach Panfilov et al. (2025) integrates imaging data to predict radiographic progression across multiple time windows. Furthermore, a model based on 3D bone morphology point clouds Wang et al. (2026) predicts anterior cruciate ligament injury risk, highlighting the role of structural factors.

AI for knee cartilage assessment. AI in knee cartilage assessment has formed a complete pipeline. For segmentation, a 2.5D U-Net with entropy maps provides explainable segmentation Sanogo et al. (2026), while a 3D-UNet framework generates thickness maps and a continuous score, achieving high sensitivity and reproducibility Margain et al. (2026). In shape modeling, a hybrid explicit-implicit neural shape model outperforms traditional methods in reconstruction accuracy Gatti et al. (2024); implicit surface reconstruction achieves annotation-free thickness mapping Wang et al. (2025a); a multi-view graph network enables precise cartilage defect grading Zhuang et al. (2026). At the system level, qMRI biomarkers combined with principal component analysis form a digital twin foundation Hoyer et al. (2025), while statistical shape models validate generalizability to pathological variations Gounot et al. (2026).

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

As depicted in **Figure 2**, PMMG comprises a multi-view branch (sagittal PDW, coronal PDW, axial T2) and a multi-modal branch (sagittal PDW, sagittal T1, sagittal PD), both sharing an identical Multi-sequence Collaborative Learning with Level-wise Attention architecture that constructs slice-, sequence-, and modality-level attention to successively select key slices, fuse primary and auxiliary sequences, and integrate imaging with textual reports and scale scores. To combat class imbalance, Cartilage-Focused Diffusion augmentation, which generates samples solely on cartilage regions, and Confidence Weighted Loss, which dynamically weights hard samples, are employed, jointly optimized with branch-wise losses. The outputs of the two branches are concatenated and fed into a classification head for final grading.

3.2 MULTI-SEQUENCE COLLABORATIVE LEARNING WITH LEVEL-WISE ATTENTION

The proposed Multi-sequence Collaborative Learning with Level-wise Attention module is designed as a unified architecture that operates on each branch of PMMG, enabling hierarchical feature refinement across slices, sequences, and modalities. Three clinical priors inherent to PFOA cartilage defect grading guide the design of this module: (i) cartilage defects are spatially sparse and typically manifest in only a small subset of MRI slices, requiring the module to identify informative slices while suppressing irrelevant background; (ii) multi-sequence MRI data exhibit a primary–auxiliary hierarchy, where the fat-suppressed proton-density-weighted (PDW) sagittal sequence provides the most direct diagnostic evidence, with other sequences offering complementary structural information; and (iii) PFOA diagnosis follows an imaging-first paradigm, where radiology reports and clinical scales serve as supplementary references rather than primary decision-making sources. To systematically incorporate these priors, as illustrated in **Figure 2**, the module constructs three levels of attention—slice-level, sequence-level, and modality-level—in a collaborative manner, enabling the model to progressively distill grading-related information while maintaining end-to-end trainability.

At the slice level, motivated by the spatial sparsity of cartilage defects, we introduce a Multi-slice Attention mechanism that selectively aggregates grading-relevant features from each MRI sequence while filtering out non-informative slices. The core idea is to construct a small set of N_s learnable slices, $S_L = [S_{L_1}, S_{L_2}, \ldots, S_{L_{N_s}}]$, and append them to each input sequence, yielding $S_i = [S_{I_i}, S_L]$, where S_{I_i} denotes the i-th image sequence. Each sequence is first encoded by a shared encoder E_I to obtain its feature representation. Multi-head self-attention is then applied to force the key information from each sequence's core slices into these learnable slices. The attention

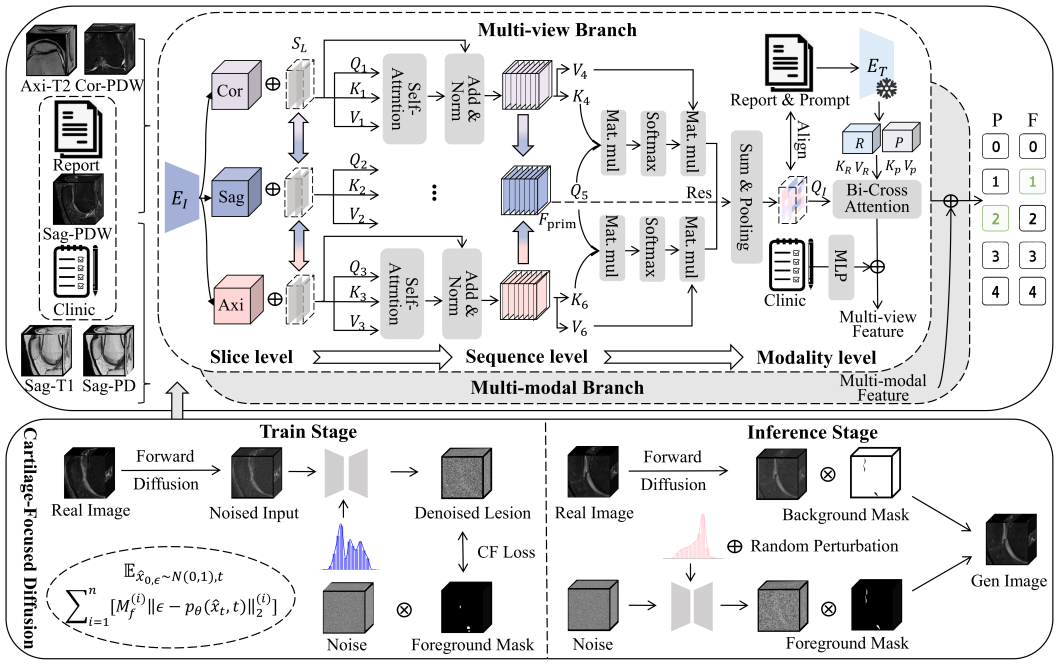

Figure 2: Overview of PMMG framework. The multi-view branch and the multi-modal branch share an identical structure: Multi-sequence Collaborative Learning with Level-wise Attention, performing slice-level, sequence-level, and modality-level fusion sequentially. Cartilage-Focused Diffusion augmentation is applied to alleviate class imbalance.

flow is formulated as:

$$\left[S_{I_i}^{l+1}\right]\left[S_{L_i}^{l+1}\right] = \text{Transformer}\left(\left[S_{I_i}^{l}\right]\left[S_{L_i}^{l}\right];\theta_i\right), \tag{1}$$

followed by aggregation and update of the learnable slices: $\hat{S}_{L_i}^{l+1} = \text{Avg}_i\left(S_{L_i}^{l+1}\right)$. This iterative process compels the learnable slices to capture and share critical information across all sequences, producing the final slice-level output $F_{\text{slice}} = [\hat{S}_{I_1}][\hat{S}_{I_2}]\cdots[\hat{S}_{I_i}]$, where all cross-channel attention must pass through the learnable slices, thereby distilling grading-relevant semantics.

At the sequence level, to address the primary–auxiliary hierarchy among MRI sequences, we design a Multi-sequence Attention mechanism that enables global cross-sequence interaction. Following clinical practice, the fat-suppressed PDW sagittal sequence is designated as the primary sequence, while other sequences serve as auxiliary sequences. Specifically, a linear transformation with weight W_Q is applied to the primary feature F_{prim} to obtain the query Q, while separate transformations with weights W_{K_i} and W_{V_i} are applied to auxiliary features to obtain keys K_i and values V_i, resulting in the preliminary attention features:

$$F_{\text{prel}} = \sum_{i=1}^{2}\left(\text{softmax}\left(QK_i^T/\sqrt{d_k}\right)V_i\right). \tag{2}$$

These features are subsequently processed by normalization and linear layers. We then add the original primary features to preliminary attention features to learn the residual between the original and attention-based features. The final output of Multi-sequence Attention is computed as:

$$F_{\text{sequence}} = \text{sum}(\text{pooling}\left(F_{\text{prim}},\ F_{\text{prel}}\right)). \tag{3}$$

In this manner, Multi-sequence Attention identifies and integrates similar features from auxiliary sequences, thereby enriching the primary features with additional spatial information and providing global guidance to Multi-slice Attention.

At the modality level, following the image-dominant diagnostic paradigm, we introduce a Multi-modal Attention mechanism to integrate unstructured textual information from radiology reports and grading-criteria prompts, further incorporating clinical scale scores as structured context. The

primary features in this module are derived from the fusion of slice-level and sequence-level outputs, $F_{\text{image}} = \text{cat}(F_{\text{slice}}, F_{\text{sequence}})$, while auxiliary features F_{text} are encoded by a text encoder E_T, which is fine-tuned on the training cohort using the pretrained M3D Bai et al. (2024) architecture. The prompts are constructed directly from textual descriptions of the evaluation criteria, as detailed in **Appendix A.1**. Through bidirectional cross-attention analogous to the sequence-level design, the modality-level attention enhances image representations with lesion-aware textual guidance, yielding the final branch output F_{modality}.

Notably, both the multi-view branch and the multi-modal branch share this identical architecture, with the only difference being their input composition. The two branch outputs are concatenated and fed into a classification head for final grading prediction.

3.3 Cartilage-Focused Diffusion

Developing models with real-world cohorts often presents challenges related to class imbalance. Specifically, suspected cases (corresponding to Grade 0 and Grade 1) constitute the majority of the cohort, which can lead the model to favor the majority class. To address this issue, we introduce a Cartilage-Focused Diffusion generative data augmentation strategy. It aims to focus exclusively on cartilage regions and generate synthetic samples with both patellar and femoral cartilage defects, balancing the class distribution.

As illustrated in **Figure 2**, the process works as follows: the original image is represented as $\hat{x}_0$, and the reverse diffusion steps from x_t to x_{t-1} are decoupled into cartilage and background components. The reverse diffusion process is described by:

$$p_\theta(o_{t-1} \mid x_t) = \mathcal{N}\big(o_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big), \tag{4}$$

and the forward diffusion process is defined as:

$$q(\hat{x}_t \mid \hat{x}_0) = \mathcal{N}\big(\hat{x}_t; \sqrt{\bar{\gamma}_t}\hat{x}_0, (1 - \bar{\gamma}_t)I\big), \tag{5}$$

with

$$\tilde{x}_{t-1} = o_{t-1} \odot M_f + x_{t-1} \odot M_b. \tag{6}$$

Here, M_f and M_b represent the cartilage foreground mask and background mask, respectively. The cartilage o_{t-1} is predicted using the reverse diffusion process p_θ from the 3D U-Net, while x_{t-1} is derived by adding noise through forward diffusion q from the original image x_0. $\bar{\gamma}_t$ is defined as $\prod_{s=1}^{t}(1 - \upsilon_s)$, where υ_t represents the variance schedule parameter, which controls the rate at which noise is progressively injected over time. This method accurately preserves the background without requiring additional predictions.

Additionally, texture control based on histograms is employed, where the cartilage texture histogram h is used as a condition through cross-attention, i.e., $o_{t-1} \sim p_\theta(\hat{x}_t, h, t)$. During training, histograms are calculated based on the ground-truth values, and during inference, they can be adjusted to control texture variations without extra cartilage-type annotations. Unlike standard diffusion models, which add Gaussian noise over T time steps and train a neural network to predict the noise distribution at step t given the noisy image x_t from the original image x_0, our method ensures the model focuses solely on the cartilage by applying the mask M_f to compute the loss only within the cartilage region.

Furthermore, we employ a multi-channel decomposition joint modeling strategy to capture correlations among multiple cartilage types and generate heterogeneous cartilage patterns within a single sample. Each channel corresponds to a different cartilage type, and the model simultaneously generates textures for all channels. The training objective is defined as follows:

$$\mathbb{E}_{\hat{x}_0, \epsilon \sim \mathcal{N}(0,1), t} \sum_{i=1}^{n} M_f^{(i)} \left\| \epsilon - p_\theta(\hat{x}_t, t) \right\|_2^{(i)}, \tag{7}$$

where ϵ represents the noise sampled from a Gaussian distribution, and i indicates the channel index. To fuse these channels, we calculate: $\sum_{i=1}^{n} \left[M_f^{(i)} o_{t-1}^{(i)} \right]$, where, outside of M_f, the predicted cartilage o_{t-1} is replaced with the true noise background $\hat{x}_{t-1}$. It is worth noting that the diffusion model is pre-trained separately on the training cohort and is only used for inference during the training of PMMG to generate synthetic samples, without participating in the end-to-end gradient backpropagation of the main model. Experimental results demonstrate that the synthetic data generated in this manner effectively enhances the model's performance.

3.4 CONFIDENCE WEIGHTED LOSS

In consideration of prediction uncertainty and the disparity in class distribution, this study introduces Confidence Weighted Loss, designed to mitigate the negative effects of class imbalance during model training by weighting the loss function with sample confidence and class weights. The Confidence Weighted Loss is defined as follows:

$$L_{\mathrm{CWL}} = (1 + \beta_t) \cdot [-\alpha_t (1 - p_t)^{\gamma} \log(p_t)], \tag{8}$$

where p_t represents the model's prediction for the t-th sample, γ is a modulation factor, α_t denotes the class weight for positive samples, β_t is the confidence weight, calculated as:

$$\beta_t = \exp\left(\delta \cdot (H_1 + H_2 + \cdots + H_n)/N\right), \tag{9}$$

with δ as the temperature coefficient, N denoting the total number of tasks (corresponding to patellar cartilage defect grading and femoral cartilage defect grading), and H_n as the confidence for the n-th task, defined as:

$$H_n = -\sum_{j=1}^{c} p_{ij} \cdot \log(p_{ij} + \varepsilon), \tag{10}$$

where c is the class number, p_{ij} is the predicted probability of the i-th sample in the j-th class, and ε is a small constant.

Since both branches in our model are responsible for grading, we apply supervision to both the final output and the two branch outputs, calculating the loss for all tasks. For the predictions from the two branches, we apply cross-entropy loss (CE) to measure the distance between the branch result y_{b} and the ground truth label $\hat{y}$. In addition, we use Confidence Weighted Loss to measure the distance between the final result y and the ground truth label $\hat{y}$, which helps the model focus on difficult samples. We use a hyperparameter λ to balance the loss terms, and the total loss is computed as:

$$L = \lambda \sum_b \sum_{i=1}^{N} L_{\mathrm{CE}}(y_{\mathrm{b}}^i, \hat{y}^i) + \sum_{i=1}^{N} L_{\mathrm{CWL}}(y^i, \hat{y}^i). \tag{11}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. This study recruits 1,173 individuals with knee-related symptoms from two medical centers. The internal cohort is randomly split into training, validation, and test sets with a 7:2:1 ratio. For each subject, we acquire: (1) multi-sequence MRI (sagittal PDW, coronal PDW, axial T2, sagittal T1, sagittal PD); (2) corresponding radiology reports (free-text, without any grades of cartilage defects or joint degeneration); (3) structured clinical data including age, sex, BMI, Lysholm functional score, and Numerical Rating Scale pain score. Two senior musculoskeletal radiologists independently evaluate the severity of defects in the patellar cartilage and femoral cartilage using a 0–4 grading scale. In cases of disagreement, an orthopedic specialist is consulted, and the final annotations are determined and verified through specialist consensus. Detailed dataset information (scanning parameters, patient enrollment, and class proportions) is provided in the **Appendix A.2**.

Baselines and Evaluation Metrics. We compare PMMG with six representative baseline methods: (1) graph-based representation methods: CSNet Zhuang et al. (2022) and KGNet Zhuang et al. (2026), which model cartilage topology via graph structures for defect assessment; (2) multi-sequence and multimodal fusion methods: CoPAS Qiu et al. (2024) and MMF Panfilov et al. (2025), which employ co-plane attention and multimodal Transformers, respectively, to integrate multi-sequence MRI information; and (3) semi-supervised and radiomics methods: DC-MT Huo et al. (2022) leverages dual-consistency semi-supervised learning for cartilage grading, while LBTRBC-M Wang et al. (2025b) employs longitudinal MRI radiomics for progression prediction. We report Accuracy (ACC), Area Under the Curve (AUC), F1 score, and Positive Predictive Value (PPV) for each of the two tasks and their average results. Implementation details of PMMG and the baselines are provided in the **Appendix A.3**.

4.2 MAIN RESULTS

Table 1 presents the comprehensive evaluation results of PMMG across the patellar and femoral cartilage defect grading tasks. Our model achieves top scores on both individual tasks and their averages, with mean ACC, AUC, F1, and PPV reaching 0.739, 0.891, 0.743, and 0.762, respectively.

Table 1: Performance comparison of different methods. For the two grading tasks and their average results, we report ACC (%), AUC (10^{-2}), F1 (10^{-2}), and PPV (10^{-2}). The best results are highlighted in bold, and the second-best results are underlined.

Method	P				F				Mean			
	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV
DC-MT (MIA'22)	53.3	64.1	46.9	48.0	67.8	79.9	60.7	55.7	60.6	72.0	53.8	51.9
CSNet (TMI'23)	58.9	74.1	60.7	62.9	75.6	<u>89.7</u>	77.3	80.7	67.2	81.9	69.0	71.8
CoPAS (NC'24)	51.1	68.3	50.0	49.3	63.3	77.5	60.7	58.6	57.2	72.9	55.4	54.0
MMF (JBHI'25)	50.0	68.3	44.0	41.0	67.8	81.3	57.6	50.2	58.9	74.8	50.8	45.6
LBTRBC-M (PLoS Med'25)	58.9	73.5	52.5	49.1	65.6	83.6	58.9	54.2	62.3	78.6	55.7	51.7
KGNet (TMI'26)	<u>61.1</u>	<u>74.9</u>	<u>65.2</u>	**70.2**	<u>76.7</u>	89.1	<u>78.8</u>	<u>81.5</u>	<u>68.9</u>	<u>82.0</u>	<u>72.0</u>	<u>75.9</u>
PMMG (Ours)	**66.7**	**84.8**	**67.4**	<u>68.5</u>	**81.1**	**93.4**	**81.2**	**83.8**	**73.9**	**89.1**	**74.3**	**76.2**

Compared to KGNet, the second-best competitor, PMMG improves mean ACC by 0.050 and mean AUC by 0.071, demonstrating a consistent and substantial lead in discriminative capability. Notably, both tasks benefit significantly from our approach, with the femoral grading task showing particularly pronounced improvements (ACC: 0.811 vs. 0.767), reflecting the model's enhanced sensitivity to the more complex anatomical structures of the femoral cartilage.

The Receiver Operating Characteristic (ROC) curves in **Figure 3.A** visually substantiate these quantitative findings. PMMG consistently yields higher AUCs and superior ROC curves across both tasks compared to all competing methods. This reflects a clearer decision boundary among different severity levels, especially in distinguishing subtle cartilage abnormalities. This observation is further supported by the confusion matrices of all compared methods shown in **Appendix Figure 6**. Furthermore, the Decision Curve Analysis (DCA) in **Figure 3.B** demonstrates that our model delivers the highest net clinical benefit across a wide threshold range on both tasks. Importantly, while most baseline models exhibit a sharp decline or even turn negative at higher threshold probabilities, PMMG maintains a stable and positive net benefit. This confirms that our approach not only achieves superior performance but also exhibits enhanced robustness to threshold selection, effectively balancing underdiagnosis and overtreatment.

To further assess the generalizability of our method, we additionally evaluate PMMG on an independent external test cohort comprising 269 patients. As shown in the last row of **Table 2**, our model achieves robust performance on this unseen dataset, with mean ACC, AUC, F1, and PPV reaching 0.705, 0.895, 0.726, and 0.800, respectively. Notably, most scores remain consistently high and are comparable to those observed in the internal test cohort, confirming that PMMG maintains robust discriminative performance across different populations and imaging protocols. The PPV values are particularly encouraging, indicating a low false-positive rate in clinical decision-making. **Appendix Table 6** reports 95% confidence intervals (CI) for all four metrics in both cohorts, with relatively narrow intervals for mean AUC indicating reasonably precise estimates of discrimination. These results collectively demonstrate that our model generalizes well to external data, highlighting its potential for real-world clinical deployment.

4.3 Compare to Radiologists

To evaluate the clinical utility of our proposed PMMG model, we compare its performance against radiologists with varying levels of experience. As presented in **Table 2**, PMMG achieves grading performance approaching that of senior radiologists while substantially outperforming junior radiologists across internal test cohort (ITC) and external test cohort (ETC). In the ITC, PMMG achieves a mean ACC of 0.739 and mean AUC of 0.891, closely approaching the senior radiologists' mean ACC of 0.767 and surpassing their mean AUC of 0.845. In the ETC, PMMG maintains strong performance with mean ACC, AUC, F1, and PPV reaching 0.705, 0.895, 0.726, and 0.800, respectively, while the senior radiologist achieves 0.712, 0.808, 0.677, and 0.676. **Appendix Table 7** shows generally moderate-to-substantial agreement, with Senior Radiologist achieving the highest reader–reference weighted Cohen's kappa point estimates. Across both test cohorts, PMMG outperforms senior radiologists on some individual cartilage grading tasks while underperforming on others, but its overall average performance approaches that of senior radiologists. This consistent advantage in AUC and PPV across both cohorts underscores the model's robustness and potential to serve as a reliable assistive tool for radiologists.

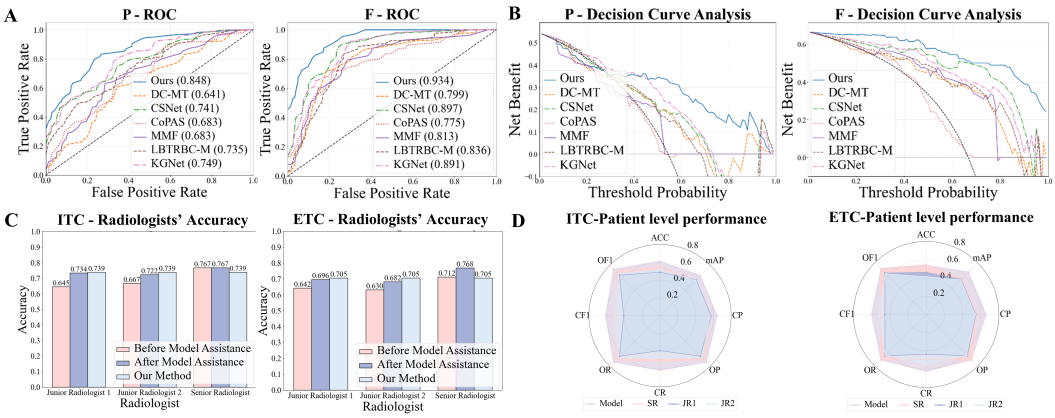

Figure 3: Performance evaluation of PMMG. (A) ROC curves; (B) DCA curves.(C) Bar chart comparing radiologists' accuracy before and after model assistance. (D) Patient-level performance comparison.

Table 2: Comparison of our proposed PMMG with radiologists in internal test cohort (ITC) and external test cohort (ETC). For the two grading tasks and their average results, we report ACC (%), AUC (10^{-2}), F1 (10^{-2}), and PPV (10^{-2}). The best results are highlighted in bold, and the second-best results are underlined.

Cohort	Setting	P				F				Mean			
		ACC	AUC	F1	PPV	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV
ITC	Junior Radiologist 1	61.1	74.1	60.5	61.8	72.2	81.5	70.1	69.6	66.7	77.8	65.3	65.7
	Junior Radiologist 2	56.7	71.1	54.3	61.5	72.2	81.5	66.7	67.7	64.5	76.3	60.5	64.6
	Senior Radiologist	**74.4**	83.0	**72.9**	**73.8**	78.9	85.9	73.8	77.0	**76.7**	84.5	73.4	75.4
	PMMG (Ours)	66.7	**84.8**	67.4	68.5	**81.1**	**93.4**	**81.2**	**83.8**	73.9	**89.1**	**74.3**	**76.2**
ETC	Junior Radiologist 1	60.6	73.7	61.4	63.1	65.4	77.0	69.3	77.2	63.0	75.4	65.4	70.2
	Junior Radiologist 2	58.4	72.2	57.0	61.0	69.9	79.9	69.5	71.3	64.2	76.1	63.3	66.2
	Senior Radiologist	65.8	77.2	61.4	61.3	**76.6**	84.4	**74.0**	73.9	**71.2**	80.8	67.7	67.6
	PMMG (Ours)	**74.7**	**89.8**	**74.0**	**74.6**	66.2	**89.1**	71.1	**85.4**	70.5	**89.5**	**72.6**	**80.0**

As shown in **Figure 3.C**, we compare the grading accuracy of the model and the radiologists before and after assistance from the model. It is evident that with the assistance of our model, the diagnostic accuracy of all radiologists significantly improves, increasing from an average of 0.693 to 0.741, with junior radiologists showing a particularly pronounced improvement from 0.656 to 0.728. With AI assistance, the performance of junior radiologists approaches that of senior radiologists, indicating that their diagnostic accuracy can reach the senior level on the internal test cohort.

Furthermore, **Figure 3.D** illustrates the patient-level evaluation results across accuracy (ACC), mean average precision (mAP), category precision (CP), overall precision (OP), category recall rate (CR), overall recall rate (OR), category F1 score (CF1), and overall F1 score (OF1) for radiologists, where a patient is considered accurately graded only when both tasks are correctly classified. As shown, our model's polygon encompasses all others, achieving the highest scores across all eight metrics. Detailed results are presented in the **Appendix Table 8**. These results validate that PMMG possesses robust and well-balanced multi-task grading capabilities in complex multi-joint grading tasks, ensuring robust diagnostic stability at the patient level.

4.4 Ablation Studies and Visualization

Ablation studies in **Table 3** show that as the core modules are progressively integrated, the overall mean ACC, AUC, F1 scores, and PPV exhibit a consistent upward trend, ultimately reaching the highest levels. Among all components, the Level-wise Attention module contributes the most substantial improvement. Compared to the baseline, it raises mean ACC from 0.572 to 0.706 (+0.134) and AUC from 0.729 to 0.872 (+0.143). Additionally, we conducted ablation studies on the three levels of attention, with the results presented in **Appendix Table 9**. The results show that progres-

Table 3: Ablation study. From top to bottom, each row represents the addition of one component upon the previous row, and the final row is the full PMMG. For the two tasks and their average, we report ACC (%), AUC (10^{-2}), F1 (10^{-2}), and PPV (10^{-2}).

Method	P				F				Avg			
	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV
Baseline	51.1	68.3	50.0	49.3	63.3	77.5	60.7	58.6	57.2	72.9	55.4	54.0
Level-wise Attention	62.2	83.8	59.6	58.9	78.9	90.6	76.4	76.3	70.6	87.2	68.0	67.6
Cartilage-Focused Diffusion	67.8	83.0	65.8	65.3	76.7	90.7	74.0	75.3	72.2	86.8	70.0	70.3
Confidence Weighted Loss	66.7	84.8	67.4	68.5	81.1	93.4	81.2	83.8	73.9	89.1	74.3	76.2

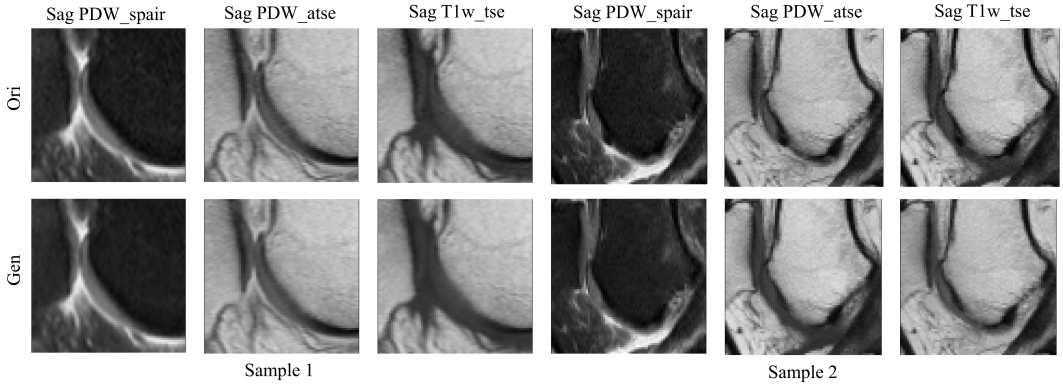

Figure 4: Cartilage-Focused Diffusion sample example.

sively integrating the three attention levels improves overall grading performance, increasing the average ACC from 0.572 to 0.739. Although slight fluctuations are observed in a few individual task metrics during the progressive integration process, the average performance maintains stable growth, demonstrating that the modules complement each other effectively. By extracting and complementing information from diverse dimensions, these modules work synergistically to achieve the final superior performance of PMMG.

Figure 4 presents representative examples of samples generated by our Cartilage-Focused Diffusion strategy alongside real samples. The generated samples realistically synthesize cartilage regions while faithfully preserving the background, demonstrating the effectiveness of our augmentation approach. Here, we train the Cartilage-Focused Diffusion model on the training cohort with cartilage defects, covering all grades of both patellar and femoral cartilage lesions, and subsequently generate 184 synthetic samples. The synthesized dataset consists of 44, 55, and 85 samples for patellar cartilage grades 2, 3, and 4, respectively, and 60, 60, and 64 samples for femoral cartilage grades 2, 3, and 4. **Appendix Figure 7** visualizes the regions of focus learned by PMMG across the three MRI views. The attention maps show that our model precisely localizes cartilage regions in all views (sagittal, coronal, and axial), demonstrating cross-view consistency and mimicking the diagnostic workflow of experienced radiologists. Collectively, these visualizations significantly enhance model interpretability and provide clinicians with intuitive visual cues for reliable diagnostic assessment.

5 Conclusion

This paper presents PMMG, a Clinical Prior-Guided Multimodal Multi-task Cartilage Defect Grading Model for PFOA. Guided by clinical priors, the multi-sequence collaborative learning architecture with level-wise attention jointly grades patellar and femoral cartilage defects. Cartilage-Focused Diffusion and Confidence Weighted Loss are employed to mitigate class imbalance. Extensive experiments on a multimodal PFOA dataset demonstrate that PMMG achieves state-of-the-art performance. Its grading accuracy approaches that of senior radiologists, while it significantly boosts junior radiologists' diagnostic performance when used as an assistive tool and demonstrates generalization to an external test cohort. These results validate that PMMG enables earlier and more precise stepwise interventions, providing a promising tool for personalized clinical decision-making.

AI USE STATEMENT

In this work, we used generative AI tools for improving the readability and clarity of the manuscript. We have not used generative AI tools for any research-related tasks, including hypothesis generation, methodology design, experimental implementation, data analysis, or result interpretation. All AI-assisted modifications were reviewed by the authors, and we take responsibility for the final content of this work.

REFERENCES

Mohamad MA Ashames, Semih Ergin, Omer N Gerek, and H Serhan Yavuz. Attention-enhanced 3d residual networks for knee abnormality classification. *Expert Systems with Applications*, pp. 129858, 2025.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Neslihan Bayramoglu, Miika T Nieminen, and Simo Saarakkala. Automated detection of patellofemoral osteoarthritis from knee lateral view radiographs using deep learning: data from the multicenter osteoarthritis study (most). *Osteoarthritis and Cartilage*, 29(10):1432–1447, 2021.

Neslihan Bayramoglu, Miika T Nieminen, and Simo Saarakkala. Machine learning based texture analysis of patella from x-rays for detecting patellofemoral osteoarthritis. *International journal of medical informatics*, 157:104627, 2022.

GBD 2015 Chronic Respiratory Disease Collaborators et al. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet. Respiratory Medicine*, 5(9):691, 2017.

Jiahui Fu, Lin Mu, Dong Dong, Mingyang Li, Zheng Miao, Xiaochen Huai, Yuhao Zheng, and Huimao Zhang. An mri-based radiomics framework for early identification and progression stratification in knee osteoarthritis: data from the osteoarthritis initiative. *BMC Musculoskeletal Disorders*, 26(1):1–10, 2025.

Nicholas R Fuggle, Roland Chapurlat, Andrea Laslop, Nasser Al-Daghri, Majed Alokail, Jotheeswaran Amuthavalli Thiyagarajan, Ewa Balkowiec-Iskra, Francis Berenbaum, Angie Botto-van Bemden, John-Joseph Borg, et al. Novel approaches to the stratified management of knee osteoarthritis. *Nature Reviews Rheumatology*, 21(11):684–695, 2025.

Anthony A Gatti, Louis Blankemeier, Dave Van Veen, Brian Hargreaves, Scott L Delp, Garry E Gold, Feliks Kogan, and Akshay S Chaudhari. Shapemed-knee: a dataset and neural shape model benchmark for modeling 3d femurs. *IEEE transactions on medical imaging*, 44(3):1140–1152, 2024.

Anna Gounot, Marion Decrouez, Baptiste Dehaine, Valérie Burdin, and Guillaume Dardenne. Statistical shape model for bone-based cartilage prediction: Applicability in healthy and pathological knees. *IEEE Transactions on Biomedical Engineering*, 2026.

JiangRong Guo, Pengfei Yan, Yong Qin, MeiNa Liu, Yingkai Ma, JiangQi Li, Ren Wang, Hao Luo, and Songcen Lv. Automated measurement and grading of knee cartilage thickness: a deep learning-based approach. *Frontiers in Medicine*, 11:1337993, 2024.

Jonathan R Hill, Edwin HG Oei, Kay M Crossley, Hylton B Menz, Erin M Macri, Michelle D Smith, Narelle Wyndow, Liam R Maclachlan, Megan H Ross, and Natalie J Collins. Classification approaches used to grade radiographic patellofemoral osteoarthritis: A scoping review. *Osteoarthritis Imaging*, 5(1):100258, 2025.

Gabrielle Hoyer, Kenneth T Gao, Felix G Gassert, Johanna Luitjens, Fei Jiang, Sharmila Majumdar, and Valentina Pedoia. Foundations of a knee joint digital twin from qmri biomarkers for osteoarthritis and knee replacement. *npj Digital Medicine*, 8(1):118, 2025.

Rui Huang, Zonghai Huang, Hantang Zhou, Qiang Zhai, Fengjun Mu, Huayi Zhan, Hong Cheng, and Xiao Yang. Memory-guided transformer with group attention for knee mri diagnosis. *Pattern Recognition*, 162:111417, 2025.

Jiayu Huo, Xi Ouyang, Liping Si, Kai Xuan, Sheng Wang, Weiwu Yao, Ying Liu, Jia Xu, Dahong Qian, Zhong Xue, et al. Automatic grading assessments for knee mri cartilage defects via self-ensembling semi-supervised learning with dual-consistency. *Medical Image Analysis*, 80:102508, 2022.

NE Lankhorst, Jurgen Damen, EH Oei, JAN Verhaar, Margreet Kloppenburg, SMA Bierma-Zeinstra, and Marienke van Middelkoop. Incidence, prevalence, natural course and prognosis of patellofemoral osteoarthritis: the cohort hip and cohort knee study. *Osteoarthritis and cartilage*, 25(5):647–653, 2017.

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*, 289(1):160–169, 2018.

Wei Liu, Yufei Chen, and Xiaodong Yue. Enhancing multi-view classification reliability with adaptive rejection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18969–18977, 2025.

Hongzhi Lv, Zhiyong Hou, Yanbin Zhu, Ran Sun, Jianzhao Wang, Guangzhao Hou, Bing Yin, Song Liu, Guang Yang, Xiao Chen, et al. Prevalence of radiographic knee osteoarthritis in china: a national survey of thirty thousand, four hundred and fifty five individuals cross-sectional study. *International Orthopaedics*, 49(10):2489–2507, 2025.

Erin Michelle Macri. *Patellofemoral osteoarthritis: characterizing knee alignment and morphology*. PhD thesis, University of British Columbia, 2017.

Paul Margain, Patrick Omoumi, and Julien Favre. Automatic framework for evaluating osteoarthritic cartilage severity: high-resolution cartilage thickness mapping and scoring. *European Radiology*, 36(7):5233–5242, 2026.

Keiley Mead, Tom Cross, Greg Roger, Rohan Sabharwal, Sahaj Singh, and Nicola Giannotti. Mri deep learning models for assisted diagnosis of knee pathologies: a systematic review. *European Radiology*, 35(5):2457–2469, 2025.

Egor Panfilov, Simo Saarakkala, Miika T Nieminen, and Aleksei Tiulpin. End-to-end prediction of knee osteoarthritis progression with multimodal transformers. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6276–6286, 2025.

Zelin Qiu, Zhuoyao Xie, Huangjing Lin, Yanwen Li, Qiang Ye, Menghong Wang, Shisi Li, Yinghua Zhao, and Hao Chen. Learning co-plane attention across mri sequences for diagnosing twelve types of knee abnormalities. *Nature Communications*, 15(1):7637, 2024.

Chandrashish Roy, Mohammed Roshan, Nidhi Goyal, Prerana Rana, Nitin P Ghonge, Amarnath Jena, Raju Vaishya, and Sourabh Ghosh. Mri detection and grading of knee osteoarthritis–a pilot study using an ai technique with a novel imaging-based scoring system. *Biomaterials Science*, 13 (19):5475–5494, 2025.

Maimouna Sanogo, Fengyi Gao, Nickolas Littlefield, Ismaeel A Siddiqui, Luke A Carlson, Amin Rezaei, Pavan Bodanki, Zoe Menezes, Abha Kanda, Mehrnaz Abedian, et al. Kneexnet-2.5 d: a clinically-oriented and explainable deep learning framework for mri-based knee cartilage and meniscus segmentation. *npj Health Systems*, 3(1):18, 2026.

Pankaj Pratap Singh, Shitala Prasad, Anil Kumar Chaudhary, Chandan Kumar Patel, and Manisha Debnath. Classification of effusion and cartilage erosion affects in osteoarthritis knee mri images using deep learning model. In *International Conference on Computer Vision and Image Processing*, pp. 373–383. Springer, 2019.

Joshua J Stefanik, Ali Guermazi, Frank W Roemer, George Peat, Jingbo Niu, Neil A Segal, Cora E Lewis, Michael Nevitt, and David T Felson. Changes in patellofemoral and tibiofemoral joint cartilage damage and bone marrow lesions over 7 years: the multicenter osteoarthritis study. *Osteoarthritis and cartilage*, 24(7):1160–1166, 2016.

Nicola Veronese, Cyrus Cooper, Olivier Bruyère, Nasser M Al-Daghri, Jaime Branco, Etienne Cavalier, Sara Cheleschi, Mario Coelho da Silva Rosa, Philip G Conaghan, Elaine M Dennison, et al. Multimodal multidisciplinary management of patients with moderate to severe pain in knee osteoarthritis: A need to meet patient expectations: N. veronese et al. *Drugs*, 82(13):1347–1355, 2022.

Roman Vuskov, Alexander Hermans, Martin Pixberg, Jonas Müller-Hübenthal, Andreas Brauksiepe, Eric Corban, Malin Cubukcu, Julia Nowak, Aleksandar Kargaliev, Marc von der Stück, et al. Comprehensive deep learning-assisted multi-condition analysis of knee mri studies improves resident radiologist performance. *European radiology*, 36(4):2563–2575, 2026.

Cen Wang, Yingpu Cui, Rile Nai, Yaofeng Zhang, Xiaodong Zhang, Xiaoying Wang, et al. Automatic diagnosis and grading of patellofemoral osteoarthritis from the axial radiographic view: a deep learning-based approach. *Acta Radiologica*, 64(2), 2023.

Dingyu Wang, Chao Liu, Xinwei Du, Xueqing Yu, Jiaxin Liu, Shanggui Liu, Songlin Zhang, Tingting Jiang, and Dong Jiang. Bone morphology as a significant risk factor for anterior cruciate ligament injury. *npj Digital Medicine*, 2026.

Pengyun Wang, Wenpeng Zhang, Liang Li, Xiaowei Zhang, Chengyan Wang, Peiqing Zhao, Tiehui Shen, Shuhong Dai, and Xianbin Wang. Cartisurface: implicit surface reconstruction for anatomically-aware cartilage thickness mapping in knee mri. *npj Digital Medicine*, 8(1):686, 2025a.

Qiuke Wang, Jos Runhaar, Margreet Kloppenburg, Maarten Boers, Johannes WJ Bijlsma, Sita MA Bierma-Zeinstra, CREDO expert group, et al. Diagnosis for early stage knee osteoarthritis: probability stratification, internal and external validation; data from the check and oai cohorts. In *Seminars in arthritis and rheumatism*, volume 55, pp. 152007. Elsevier, 2022.

Ting Wang, Hao Liu, Wenbo Zhao, Peihua Cao, Jia Li, Tianyu Chen, Guangfeng Ruan, Yan Zhang, Xiaoshuai Wang, Qin Dang, et al. Predicting knee osteoarthritis progression using neural network with longitudinal mri radiomics, and biochemical biomarkers: A modeling study. *PLoS Medicine*, 22(8):e1004665, 2025b.

Zhuoyao Xie, Zelin Qiu, Yanwen Li, Xiaoping Chen, Qiang Ye, Menghong Wang, Xiaqing Chen, Yangyang Shao, Caolin Liu, Liwen Song, et al. Development of a multi-task deep learning system for classification of nine common knee abnormalities on mri: a large-scale, multicentre, stepwise validation study. *EClinicalMedicine*, 89, 2025.

Pauline Shan Qing Yeoh, Khin Wee Lai, Siew Li Goh, Khairunnisa Hasikin, Yan Chai Hum, Yee Kai Tee, and Samiappan Dhanalakshmi. Emergence of deep learning in knee osteoarthritis diagnosis. *Computational intelligence and neuroscience*, 2021(1):4931437, 2021.

Lirong Zhang, Zhiwei Che, Yang Li, Meng Mu, Jialin Gang, Yao Xiao, and Yibo Yao. Multi-level classification of knee cartilage lesion in multimodal mri based on deep learning. *Biomedical Signal Processing and Control*, 83:104687, 2023.

Xiaoyu Zhang, Jungang Han, Zhe Li, Jiayu Yang, and Pengyu Zhan. A multi-task learning framework for segmentation and classification of patellofemoral osteoarthritis in multi-parametric magnetic resonance imaging. In *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition*, pp. 449–456, 2022.

Zixu Zhuang, Liping Si, Sheng Wang, Kai Xuan, Xi Ouyang, Yiqiang Zhan, Zhong Xue, Lichi Zhang, Dinggang Shen, Weiwu Yao, et al. Knee cartilage defect assessment by graph representation and surface convolution. *IEEE Transactions on Medical Imaging*, 42(2):368–379, 2022.

Zixu Zhuang, Dongdong Chen, Sheng Wang, Kai Xuan, Xiangyu Zhao, Zhong Xue, Dinggang Shen, Lichi Zhang, Weiwu Yao, and Qian Wang. Enhancing knee disease diagnosis via multi-view graph representation with multi-task pre-training. *IEEE Transactions on Medical Imaging*, 2026.

A Appendix

A.1 Assessment Criteria

Experts evaluate the degree of cartilage defects on MRI images using a 0-4 grading scale. Grade 0 indicates normal cartilage morphology. Grade 1 is characterized by the disappearance of the layered structure of the cartilage with localized low-signal areas, while the surface of the cartilage remains smooth. Grade 2 presents mild to moderate irregularity of the cartilage surface contour, with defect depth less than 50% of the total cartilage thickness. Grade 3 shows severe irregularity of the surface contour, with defect depth exceeding 50% of the total thickness but without complete detachment. Grade 4 refers to full-thickness defects with cartilage exfoliation, exposing the subchondral bone, with or without abnormal signal changes in the subchondral bone.

A.2 Dataset Details

Patient enrollment. As shown in **Appendix Figure 5**, this study recruits 1,173 individuals with knee-related symptoms from two medical centers, according to the inclusion and exclusion criteria. Patients are enrolled according to the following eligibility criteria: (i) Age of 18 years or above; (ii) Skeletal maturity has been reached; (iii) Written informed consent has been obtained, and agree to undergo the required follow-up evaluations. The exclusion criteria are as follows: (i) Systemic joint disease; (ii) Vascular, neurological, or dermatological issues; (iii) MRI contraindications or significant artifacts; (iv) Incomplete data, records, or consent.

MRI scanning parameters. The MRI scans of the 1,173 patients enrolled in our cohort are acquired on two different scanners across two medical centers. Detailed scanning parameters are presented in **Appendix Table 4**.

Class proportions. Appendix Table 5 summarizes the class proportions across the five grading tasks.

A.3 Implementation Details

All experiments are conducted on a workstation equipped with two NVIDIA RTX A100 GPUs. Because grades 0 and 1 show no definite defects (with grade 1 limited to surface roughness) and generally warrant the same clinical management, we temporarily merge them into a single class during both training and inference. Consequently, each original five-class classification task (grades 0–4) is reformulated as a four-class classification task, resulting in two classification tasks, each with four classes.

Data preprocessing includes multi-sequence MRI cropping. Specifically, we train a nnU-Net-based cartilage segmentation model using a subset of the dataset with manually annotated cartilage masks. The trained model is then applied to the entire MRI cohort to generate cartilage segmentation masks. Subsequently, bounding boxes are extracted from the predicted masks and are used to crop the corresponding multi-sequence MRI volumes, thereby reducing irrelevant background regions and focusing the model on anatomically relevant cartilage areas.

During the segmentation phase, the initial step involves resampling the data to a voxel resolution of $3.3 \times 0.3 \times 0.3\ \text{mm}^3$ to optimize memory usage. This is followed by trimming the scans to patches of size 320×320×16 voxels. Additionally, various data augmentation techniques, such as random rotations, scaling adjustments, and random noise, are applied to enhance the diversity and complexity of the training dataset.

The segmentation model is trained exclusively on sagittal images, with the images clipped to a size of 224×224×24, centered on the mask. For coronal and transverse images, the corresponding images

are clipped to 224×224×24 by rotating the mask accordingly. In the prediction phase, image preprocessing converts the original MRI scans into a uniform format across different cohorts, minimizing variability and preparing the data for accurate prediction. The preprocessing includes resampling the data to a voxel resolution of $3.3 \times 0.3 \times 0.3$ mm^3. Furthermore, we apply data augmentation with a 50% probability—such as random cropping, random rotation (0 to 10 degrees), contrast adjustment, and brightness adjustment—during training to ensure model generalization.

The deep learning models in this study are implemented with PyTorch 2.6.0. The nnU-Net model is trained using the AdamW optimizer with an initial learning rate of 0.00001 for 200 epochs with a batch size of 2. The PMMG model is also trained using the AdamW optimizer, with an initial learning rate of 0.00005 for 100 epochs. Early stopping is applied with a patience value of 8, and the batch size is set to 4. The best model checkpoint is selected based on performance on the internal validation cohort, and its predictions are validated on both the internal test cohort and the external test cohort. The M3D model is fine-tuned according to its original configuration, using a batch size of 4. The Cartilage-Focused Diffusion model is trained using the AdamW optimizer with an initial learning rate of 0.0001 for 20,000 epochs with a batch size of 8.

We modify existing methods while preserving their core network architectures. Specifically, to align with our task definition, we extend each model to perform two four-class classification tasks, grading patellar and femoral cartilage lesions separately, consistent with PMMG. Additionally, several baselines are originally tailored to specific MRI sequences, so we only compare them on the sequences available in our dataset. To ensure a fair comparison, all models are trained under the same experimental settings.

DC-MT is a semi-supervised framework based on a mean-teacher classification model that enhances consistency between the teacher and student networks through a novel dual-consistency strategy. The original work demonstrated substantial improvements in the classification and localization of knee cartilage defects. In our implementation, we use fat-suppressed sagittal MR images together with cartilage masks as inputs, train the model in a fully supervised manner, and extend the output to two four-class classification tasks.

CSNet is a non-Euclidean deep learning framework that models knee cartilages as a graph representation and performs defect assessment via surface convolution with self-attention. The original work achieved state-of-the-art performance in knee cartilage defect grading and provided strong 3D interpretability through attention visualization on cartilage surfaces. In our implementation, we use sagittal fat-suppressed MR images together with bone surface-based vertex sampling, train the model with patch-level pre-training followed by subject-level supervision, and extend the output to two four-class classification tasks.

CoPAS is a deep learning model that incorporates coplanar attention across image sequences to classify knee abnormalities, achieving strong performance on 12 abnormality categories, including cartilage lesions. Since additional coronal sequences are unavailable in our data, we remove the cross-sequence attention module for coronal views and retain only the T1 sequence for sagittal cross-sequence attention. Instead of the original plane-aware feature integration based on abnormality probability maps, we directly concatenate features from the three branches and reformulate the task as two four-class classification tasks.

MMF employs a hybrid deep learning encoder to integrate multimodal data for predicting knee osteoarthritis (KOA) progression, providing a unified end-to-end data-driven framework and new insights into multimodal imaging. Following the original architecture, we encode the five MRI sequences and fuse their features but replace the progression prediction objective with two four-class classification tasks.

LBTRBC-M combines radiomics features with biomarkers and clinical variables to predict KOA progression. The original study showed satisfactory performance and highlighted radiomics features as potential imaging biomarkers for KOA progression. In our work, we use sagittal MRI sequences and their corresponding cartilage masks, extract and select features following the original pipeline, and reformulate the progression prediction task into two four-class classification tasks.

KGNet is a graph-based framework originally designed for multi-view MRI diagnosis, which constructs a unified knee graph and employs multi-task pre-training with masked vertex reconstruction and segmentation. The original work achieved strong performance on cartilage defect, ACL tear,

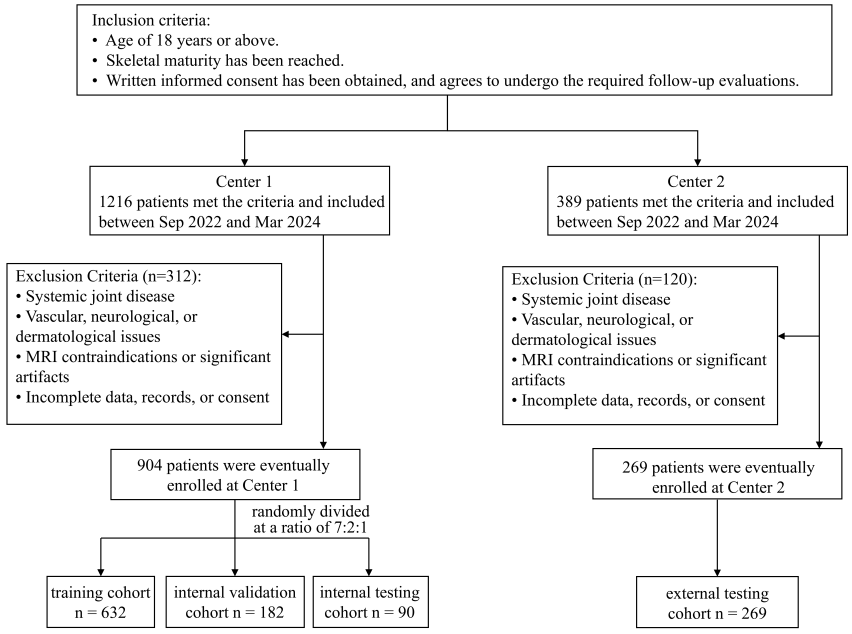

Figure 5: Flowchart of the study enrollment process.

Table 4: The MRI scanning parameters for the two medical centers.

Affiliated Site	Scanner	Sequence	TR/TE (ms)	Slice (mm)	Gap (mm)	FOV (mm)	Matrix size
1	Philips 1.5 T	PD-sag	5000/30	3.0	0.35	160*160	248*206
		T1WI-sag	500/17	3.0	0.35	160*160	320*280
		PDS-sag	3000/30	3.0	0.35	160*160	248*226
		PD-cor	2900/30	3.0	0.35	160*160	320*278
		T2WI-tra	Shortest/100	3.0	0.35	160*160	324*248
2	UI 3.0 T	PD-sag	2603/36.72	3.5	15	160*160	320*85
		T1WI-sag	620/11.86	3.5	15	160*160	320*75
		T2WI-sag	2000/36.72	3.5	15	160*160	320*80
		PD-cor	2096/34.86	3.5	10	160*160	320*85
		T2WI-tra	2162/86.38	3.5	10	150*150	384*80

and knee abnormality tasks using three aligned views. In our implementation, we use only sagittal fat-suppressed MR images together with bone segmentation to construct the knee graph representation, pre-train KGNet with reconstruction and segmentation tasks in a self-supervised manner, and fine-tune it for subject-level cartilage defect grading.

Table 5: Class distribution of patellar (P) and femoral (F) cartilage defect grading tasks across internal test cohort (ITC) and external test cohort (ETC). Data are presented as case numbers with percentages in parentheses.

Cohort	Task	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
ITC	P	256 (28.32%)	236 (26.11%)	157 (17.37%)	124 (13.72%)	131 (14.49%)
	F	469 (51.88%)	145 (16.04%)	121 (13.38%)	84 (9.29%)	85 (9.40%)
ETC	P	37 (13.75%)	98 (36.43%)	54 (20.07%)	36 (13.38%)	44 (16.36%)
	F	138 (51.30%)	55 (20.45%)	27 (10.04%)	35 (13.01%)	14 (5.20%)

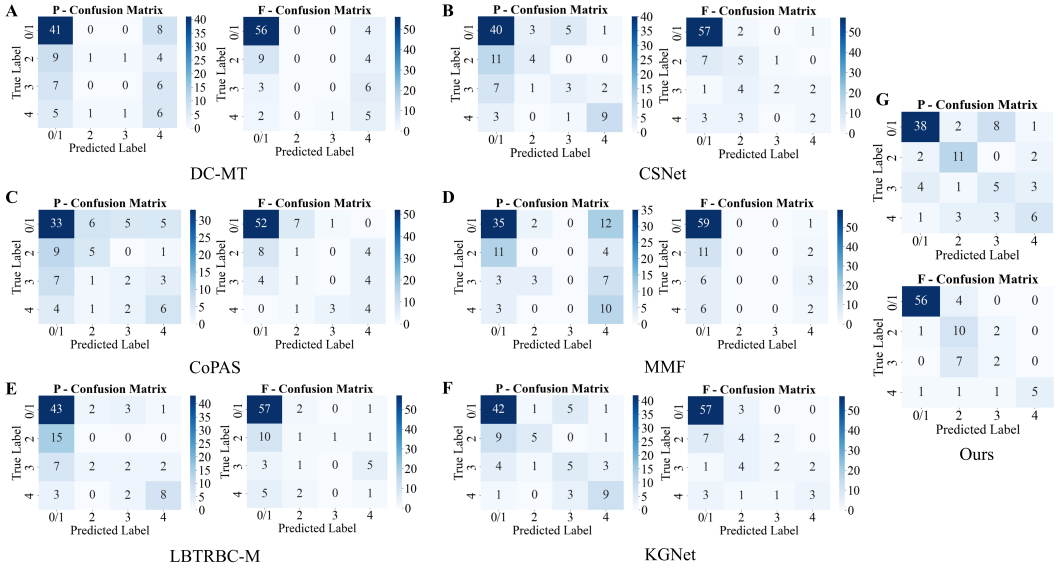

Figure 6: Confusion matrices of PMMG and comparison methods.

Table 6: Reader–reference and inter-reader agreement for patellar (P) and femoral (F) cartilage defect grading in the internal test cohort (ITC) and external test cohort (ETC). We report quadratic-weighted Cohen's kappa (QWK) with 95% confidence intervals (CIs), exact agreement (EA, %), and within-one-grade agreement (W1A, %). CIs were estimated using 20,000 paired record-level bootstrap resamples with the percentile method. W1A denotes an absolute difference of at most one category on the four-level ordinal scale. Junior Radiologist 1 (JR1) and Junior Radiologist 2 (JR2) denote the two junior radiologists, Senior Radiologist denotes the senior radiologist, and GT denotes the consensus reference standard.

Cohort	Comparison	P			F		
		QWK (95% CI)	EA	W1A	QWK (95% CI)	EA	W1A
Reader–reference agreement							
ITC	JR1 vs. GT	0.494 (0.306–0.660)	56.7	85.6	0.594 (0.373–0.766)	72.2	92.2
	JR2 vs. GT	0.636 (0.465–0.769)	61.1	87.8	0.712 (0.507–0.855)	72.2	96.7
	SR vs. GT	0.709 (0.537–0.833)	68.9	90.0	0.746 (0.578–0.856)	72.2	92.2
ETC	JR1 vs. GT	0.691 (0.612–0.759)	58.4	92.9	0.544 (0.409–0.656)	69.9	91.4
	JR2 vs. GT	0.717 (0.637–0.784)	60.6	92.6	0.652 (0.546–0.739)	65.4	93.7
	SR vs. GT	0.765 (0.698–0.820)	65.8	92.2	0.700 (0.590–0.786)	76.6	93.7
Inter-reader agreement							
ITC	JR1 vs. JR2	0.780 (0.652–0.866)	71.1	97.8	0.806 (0.627–0.908)	84.4	97.8
	JR1 vs. SR	0.464 (0.242–0.647)	62.2	85.6	0.588 (0.350–0.776)	77.8	92.2
	JR2 vs. SR	0.605 (0.418–0.746)	63.3	86.7	0.690 (0.475–0.839)	74.4	94.4
ETC	JR1 vs. JR2	0.588 (0.490–0.674)	55.8	90.7	0.508 (0.380–0.618)	61.7	91.8
	JR1 vs. SR	0.697 (0.604–0.772)	66.2	93.3	0.512 (0.353–0.644)	72.9	93.3
	JR2 vs. SR	0.647 (0.558–0.722)	57.2	89.2	0.505 (0.375–0.615)	59.9	90.3

Table 7: Performance of PMMG with 95% bootstrap confidence intervals on the internal test cohort (ITC) and external test cohort (ETC). For the two tasks and their average, we report ACC (%), AUC (10^{-2}), F1 (10^{-2}), and PPV (10^{-2}).

Cohort	Task	ACC	AUC	F1	PPV
ITC	P	66.7 (56.7–76.7)	84.8 (77.9–90.8)	67.4 (57.5–76.9)	68.5 (58.9–79.6)
	F	81.1 (72.2–88.9)	93.4 (89.3–96.9)	81.2 (72.8–88.7)	83.8 (76.1–92.0)
	Mean	73.9 (66.1–81.1)	89.1 (84.6–93.1)	74.3 (66.4–81.5)	76.2 (68.9–84.4)
ETC	P	74.7 (69.5–79.9)	89.8 (86.7–92.6)	74.0 (68.3–79.4)	74.6 (69.1–80.3)
	F	66.2 (60.6–71.7)	89.1 (85.6–92.4)	71.1 (66.0–75.9)	85.4 (81.0–89.7)
	Mean	70.5 (66.2–74.5)	89.5 (86.7–92.0)	72.6 (68.3–76.6)	80.0 (76.4–83.8)

Table 8: Patient-level performance comparison between PMMG and radiologists on the internal test cohort (ITC) and external test cohort (ETC). ACC: accuracy; mAP: mean average precision; CP/CR/CF1: category-level precision, recall, and F1-score; OP/OR/OF1: overall-level precision, recall, and F1-score. All values are reported in (%) or (10^{-2}). The best results are highlighted in bold, and the second-best results are underlined.

Cohort	**Method**	Overall		Precision		Recall		F1-score	
		ACC	mAP	CP	OP	CR	OR	CF1	OF1
ITC	Junior Radiologist 1	48.89	<u>57.31</u>	<u>57.31</u>	64.44	39.21	64.44	40.93	64.44
	Junior Radiologist 2	52.22	56.22	56.22	66.67	48.14	66.67	<u>50.67</u>	66.67
	Senior Radiologist	<u>55.56</u>	56.35	56.35	<u>70.56</u>	<u>51.43</u>	<u>70.56</u>	49.67	<u>70.56</u>
	PMMG (Ours)	**61.11**	**64.05**	**64.05**	**73.89**	**61.31**	**73.89**	**61.20**	**73.89**
ETC	Junior Radiologist 1	46.10	54.50	54.50	64.13	43.54	64.13	45.41	64.13
	Junior Radiologist 2	41.26	53.82	53.82	63.01	<u>55.15</u>	63.01	<u>52.59</u>	63.01
	Senior Radiologist	**54.65**	<u>55.93</u>	<u>55.93</u>	**71.19**	49.49	**71.19**	50.51	**71.19**
	PMMG (Ours)	<u>53.90</u>	**65.79**	**65.79**	<u>70.45</u>	**62.43**	<u>70.45</u>	**59.90**	<u>70.45</u>

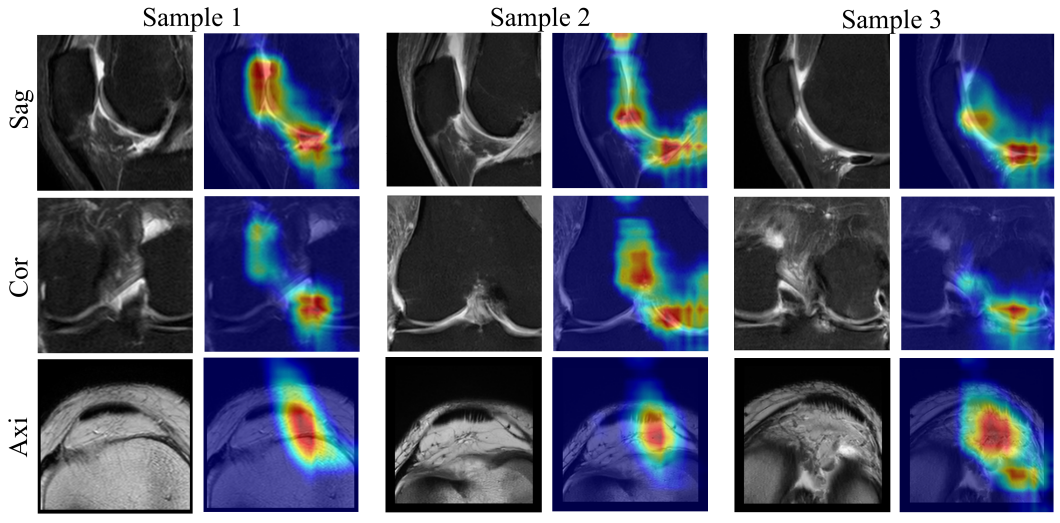

Figure 7: Visualizations of three-view MRI sequences for PFOA patients in three test cohorts using Class Activation Mapping (CAM).

Table 9: Ablation study of the three levels attention (slice, sequence, and modality). From top to bottom, each row represents the addition of one component upon the previous row, and the final row is the full PMMG. For the two tasks and their average, we report ACC (%), AUC (10^{-2}), F1 (10^{-2}), and PPV (10^{-2}).

Method	P				F				Avg			
	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV	ACC	AUC	F1	PPV
Baseline	51.1	68.3	50.0	49.3	63.3	77.5	60.7	58.6	57.2	72.9	55.4	54.0
Multi-slice Attention	60.0	75.1	53.7	50.0	64.4	77.2	63.6	63.4	62.2	76.1	58.7	56.7
Multi-sequence Attention	63.3	72.2	57.9	60.9	65.6	78.3	60.3	55.9	64.4	75.2	59.1	58.4
Multi-modal Attention	66.7	84.8	67.4	68.5	81.1	93.4	81.2	83.8	73.9	89.1	74.3	76.2