Anatomically-Focused Patches for Lightweight and Explainable Knee OA Grading

Tien-En Chang[1,2] and Herve Lombaert[2(✉)]

[1] National Taiwan University, Taipei, Taiwan
[2] Polytechnique Montreal, Montreal, Canada
herve.lombaert@polymtl.ca

Abstract. Current deep learning models for knee osteoarthritis (OA) grading often lack anatomical guidance, limiting their accuracy and explainability. This work proposes a novel framework centered on anatomically-focused patches to overcome these limitations. Our method extracts a set of small image patches from clinically-relevant locations along the joint line, identified by automated landmark detection. These patches are then processed as a bag of instances within an attention-based multiple instance learning (MIL) framework. The MIL model learns to identify and weight the most salient pathological features for an accurate, patient-level diagnosis. Our approach is evaluated on the OAI dataset and achieves state-of-the-art performance with a quadratic weighted Cohen's Kappa of 0.807. This result outperforms larger baselines such as ResNet-34 while using over 85 times fewer parameters. Furthermore, our attention-weighted visualization method produces sharp, clinically meaningful saliency maps that precisely localize features such as osteophytes and joint space narrowing, in contrast to the diffuse heatmaps of prior work. By combining anatomical guidance with an MIL framework, our work presents a lightweight, accurate and trustworthy solution for automated knee OA grading. The code is available at: https://github.com/tien-endotchang/focused-patch-KOA.

Keywords: knee osteoarthritis · Kellgren and Lawrence grading · X-ray · multiple instance learning, anatomical guidance

1 Introduction

Knee osteoarthritis (OA) is a highly prevalent joint disorder, affecting more than 650 million individuals over the age of 40 worldwide [8]. It is primarily characterized by progressive cartilage degeneration, osteophyte formation, and joint space narrowing (JSN). In clinical practice, the severity of knee OA is most commonly assessed using the Kellgren-Lawrence (KL) grading system [14]. This widely-used method provides a semi-quantitative score based on these radiographic features. KL grades have five categories that indicates the level of knee degradation. Despite its widespread clinical adoption, manual KL grading is inherently subjective, labor-intensive, and prone to variability [9], which motivates the development of automated assessment methods.

The automated grading of knee OA from radiographs has been significantly advanced by machine learning, particularly deep learning [2,21,6,20,24]. A common and effective paradigm involves a two-step approach: (1) localizing the knee joint region of interest (ROI), and (2) classifying the localized ROI to predict its KL grade. Early deep learning models, such as those by Antony et al. [2], established this pipeline using a Fully Convolutional Network (FCN) for localization and a Convolutional Neural Network (CNN) for classification. Tiulpin et al. [21] used a SVM based localization [22] and further constrained the model attention to two large patches covering the medial and lateral sides of the knee joint. They employed a Siamese network architecture that shared weights for processing both patches, a design which leverages the joint's symmetry to improve model efficiency. Their model achieved high performance while significantly reducing model complexity and improving explainability compared to larger, monolithic CNNs. Subsequent work aiming for higher performance often incorporated more complex components. For instance, Chen et al. [6] employed a customized object detector (YOLOv2) with a specialized ordinal loss, while a later study by Tiulpin et al. [20] used an advanced CNN backbone with multi-task learning objectives. Yang et al. [24] applied a graph attention network.

However, current methods often operate on coarse, bounding-box regions of the knee. This approach discards the rich anatomical information about bone shape and joint space geometry that is essential for clinical diagnosis. The underutilization of this fine-grained anatomical context represents a critical limitation, often leading to models that are less data-efficient and lack precise explainability. A few studies have attempted to leverage this information by engineering handcrafted shape features [3], but the full potential of using anatomical shape to directly guide deep feature learning remains largely untapped. We hypothesize that by focusing on anatomically-defined regions, we can build a more accurate, efficient, and explainable classification model.

To address this limitation, we introduce anatomically-focused patches, a novel input representation for knee OA grading. Our approach leverages the precise contours of the femur and tibia to define a set of small, overlapping patches along the tibiofemoral joint line. This strategy directs a lightweight CNN to learn discriminative features from the key clinically relevant regions. The generation of these patches relies on stable and reproducible anatomical landmarks of the knee joint. Our experiments use BoneFinder [16,15] for this purpose, which identifies such landmarks using statistical shape models. To effectively aggregate information from this set of patches, an attention-based multiple instance learning (MIL) framework [12] is employed to form a final, patient-level prediction. We demonstrate that this anatomy-guided approach leads to superior classification performance and enhanced explainability, all while using a model with significantly lower complexity. Our primary contributions are:

- We propose a novel method for generating anatomically-focused patches for knee OA grading, which leverages detailed anatomical segmentation data to improve upon heuristic or bounding-box-based ROI extraction.

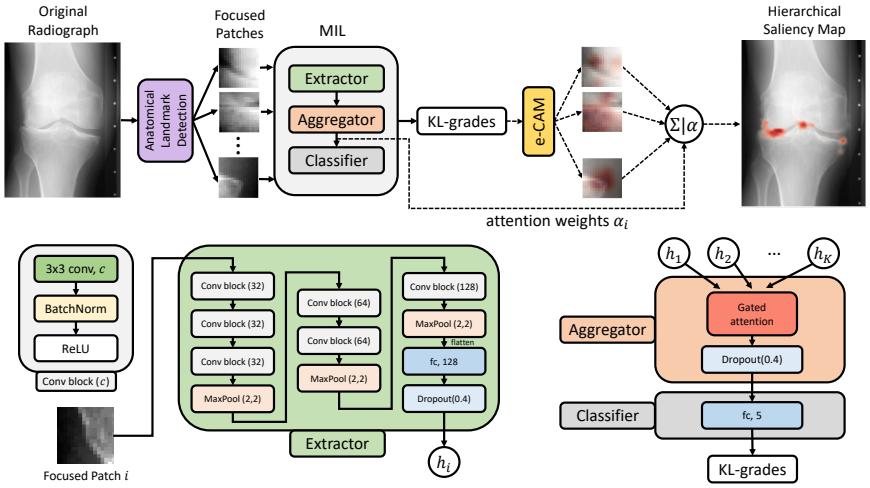


Fig. 1. [Overview] A schematic overview of the proposed method. An input radiograph is processed via anatomical landmark detection to generate a bag of focused patches. These are fed into a MIL model for KL grade prediction. The MIL model consists of three components: **Extractor** (green); **Aggregator** (orange); and **Classifier** (gray). Dashed lines show the explanation pathway: local GradCAMs are weighted by attention scores α_i to form a single, precise saliency map.

- We demonstrate that by feeding these patches into an efficient CNN and aggregating them with an attention-based MIL framework, our model surpasses the performance of influential prior work that relied on less precise, heuristically-defined patches.
- We show qualitatively that our approach yields more precise and clinically meaningful class activation maps, confirming that the model learns from the correct anatomical structures and enhancing its explainability.

2 Methods

Our work introduces a novel framework for knee OA grading centered on anatomically-focused patches, a new input representation designed to guide deep learning models with anatomical shapes. This approach consists of three integrated components (see Fig. 1): (1) a method for generating these focused patches from bone contours; (2) a MIL framework to aggregate features from these patches; and (3) a post-hoc visualization technique to explain the decisions from our model.

2.1 Anatomically-Focused Patches

Our approach moves beyond coarse, bounding-box-based ROI localization by leveraging fine-grained anatomical information. The generation of these patches

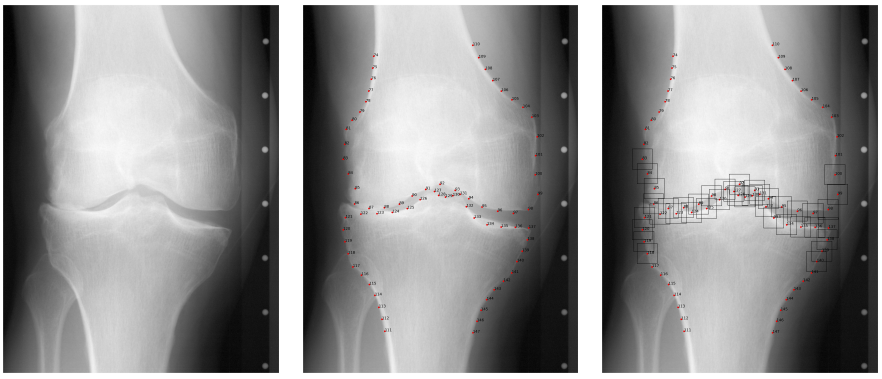
begins with the automated detection of anatomical landmarks that consistently trace the contours of the distal femur and proximal tibia. This provides a detailed, point-based representation of the shape of knee joint, as shown in Fig. 2(b).

Rather than using all available landmarks, a clinically-informed selection strategy is applied. The radiographic assessment of OA, as defined in standardized atlases, focuses on features like osteophyte formation and JSN at the tibiofemoral joint surfaces [1]. Accordingly, only the subset of landmarks located specifically along these articular contours is retained. This selection discards landmarks on the outer bone shafts, which are less relevant for KL grading, and concentrates the attention of model on the regions prone to osteophyte formation and JSN.

Centered on each of these K selected landmarks, a square image patch of size $P \times P$ pixels is extracted. This procedure yields a set of K anatomically-focused patches for each knee radiograph (Fig. 2(c)). Finally, each patch is resized to a uniform resolution, transforming the knee image into a bag of instances $\{X_1, X_2, ..., X_K\}$, which serves as the natural input for a MIL framework.

2.2 Attention-based Multiple Instance Learning

The anatomically-focused patching strategy transforms each knee radiograph into a bag of numerous small instances. A key challenge is that the evidence for OA, such as a small osteophyte, may only be present in one or a few of these instances. A final patient-level KL grade must therefore be inferred from



(a) Original radiograph (b) Detected Anatomical Landmarks (c) Selected landmarks & patches

Fig. 2. [Patches Processing] Generation of anatomically-focused patches. (a) An input knee radiograph. (b) A set of anatomical landmarks are automatically detected, outlining the bone contour. (c) A clinically-informed subset of landmarks is selected along tibiofemoral joint surfaces. A square patch is extracted around each selected landmark, forming the set of focused inputs for our classification model. In this work, the landmark detection was performed using the BoneFinder tool.

this collection of patches, where the importance of each patch is unknown in advance.

This formulation directly corresponds to the problem definition of multiple instance learning. In MIL, a bag (the knee) is labeled as positive (having JSN) if at least one of its instances (the patches) is positive (having JSN). To solve this, we adopt the attention-based deep MIL framework from Ilse et al. [12]. This framework is composed of three components: a feature extractor, an aggregator, and a classifier.

First, each patch X_i is passed through a CNN feature extractor, f, to produce a low-dimensional embedding, $h_i = f(X_i) \in \mathbb{R}^M$. Next, to create a single feature vector z representing the entire knee, the instance embeddings $h_1, ..., h_K$ are combined by a gated attention aggregation. This mechanism learns to assign an attention weight α_i to each patch embedding h_i, effectively allowing the model to focus on patches that are more informative for the final prediction. The aggregated feature vector z is a weighted sum:

$$z = \sum_{i=1}^{K} \alpha_i h_i, \tag{1}$$

where the attention weights α_i are computed as follows:

$$\alpha_i = \frac{\exp\{w^\top(\tanh(Vh_i) \odot \text{sigm}(Uh_i))\}}{\sum_{k=1}^{K} \exp\{w^\top(\tanh(Vh_k) \odot \text{sigm}(Uh_k))\}}. \tag{2}$$

Here, $w \in \mathbb{R}^L, V \in \mathbb{R}^{L\times M}, U \in \mathbb{R}^{L\times M}$ are learnable weight matrices, and $\odot$ denotes Hadamard product. Finally, the aggregated feature vector z is fed into a classifier, g, to produce the final KL grade prediction. The entire framework is trained end-to-end by minimizing a weighted cross-entropy loss function between the predicted probabilities and the ground-truth KL grade.

2.3 Attention-weighted Saliency Map

To visualize the decision-making process of the model, an attention-weighted saliency map is generated. This method integrates the local, pixel-level explanations from individual patches with the global, patch-level importance assigned by the MIL attention mechanism.

Standard class activation mapping (CAM) techniques such as Grad-CAM [19] can produce noisy or incomplete heatmaps [5]. Inspired by [4], an ensemble of CAM variants is computed for each of the K input patches to create a more robust local explanation. This ensembled CAM (e-CAM) is the average of heatmaps generated by multiple techniques, using the final convolutional layer of the feature extractor as the target.

These local e-CAM heatmaps are assembled to form a full-image saliency map. An empty array with the same dimensions as the original radiograph is initialized. Then for each patch, its e-CAM heatmap is resized to its original dimensions and placed back at its anatomical location within the empty array.

Crucially, each placed heatmap is weighted by its corresponding attention score, α_i, from the aggregator (Eq. 2). The resulting attention-weighted sum of all local heatmaps is then normalized and smoothed to produce the final, unified visualization.

3 Experiments and Results

The experiments are designed to validate our primary contributions. We first demonstrate that our proposed framework, which combines anatomically-focused patches with an attention-based MIL model, achieves state-of-the-art classification performance compared to established baselines. We then provide a qualitative analysis of our attention-weighted saliency maps to confirm that its decisions are driven by clinically relevant anatomical features.

3.1 Dataset and Preprocessing

Dataset. This study utilizes data from the Osteoarthritis Initiative (OAI, https://nda.nih.gov/oai), a multi-center, longitudinal, public observational study of knee OA. The full cohort includes 4,796 participants aged 45-79. For our experiments, we used the bilateral PA fixed-flexion knee radiographs from the baseline visit of the OAI database. The KL grades, as provided by the OAI, serve as the ground truth for our classification task. The definitions of KL grades for knee joint are as follows: KL0 (normal), KL1 shows doubtful JSN and possible osteophytes (doubtful), KL2 demonstrates definite osteophytes and possible JSN (minimal), KL3 shows moderate multiple osteophytes, definite JSN (Moderate) and KL4 shows large osteophytes, marked JSN (severe). To ensure consistency, the image contrast were normalized using histogram truncation following [21].

Anatomically-Focused Patch Processing. Our patches cover consistent anatomical areas across images of varying sizes. Their scaling strategy is thus designed to yield normalized patches with a consistent physical size of 10 mm $\times$ 10 mm. Our strategy subsequently builds a set of patches along the bone contours. Our experiments use a set of 74 anatomical landmarks outlining the bone contours, first identified using the BoneFinder tool [16,15]. From this set, $K = 41$ landmarks corresponding to the tibiofemoral articular surfaces are selected. A square patch is extracted around each selected landmark and resized to a final input of 16×16 pixels. This low resolution encourages the model to learn key clinical relevant features.

Data Splits and Augmentation. The dataset was augmented using two strategies as in [21]. First, all images of left knees were horizontally flipped to increase data size. Second, we applied random adjustments to contrast, brightness, and gamma correction. This results in 8,952 unique knee instances. We then performed a knee-level stratified split based on KL grade to create training (60%), validation (20%), and testing (20%) sets. The exact distribution of KL grades across these splits is detailed in Table 1, which highlights the significant class imbalance inherent in the OAI dataset.

Table 1. Distribution of Kellgren-Lawrence (KL) grades in the training, validation, and test sets. The numbers indicate the count of knee images per grade, demonstrating the natural class imbalance of the dataset.

Group	Total	KL0	KL1	KL2	KL3	KL4
Train	5371	2068	959	1424	743	177
Validation	1791	690	319	475	248	59
Test	1790	689	319	475	248	59

3.2 Experimental Setup

Model Architecture. Our MIL framework (Fig. 1) consists of a feature extractor, a gated attention aggregator, and a classifier. The Extractor is a lightweight CNN that processes each 16×16 patch. It contains a sequence of convolutional blocks, separated by 2×2 max-pooling. Each block contains a 3×3 convolution, BatchNorm, and ReLU. The resulting feature map is flattened and passed through a fully-connected layer to produce a feature embedding ($M = 128$). The Aggregator implements the gated attention mechanism [12] with a hidden dimension of $L = 128$, producing attention weights for all $K = 41$ patch embeddings. Finally, the Classifier is a single fully-connected layer mapping the 128-dimensional attention-weighted feature vector to the 5 KL grades.

Training. All models were trained for 100 epochs using the Adam optimizer with an initial learning rate of 1e-4. To mitigate overfitting, a weight decay of 1e-4 and dropout with a rate of 0.4 was applied. We used a batch size of 16 knee image bags. To address the severe class imbalance shown in Table 1, the cross-entropy loss was weighted, with weights set to the inverse frequency of each class in the training set. A learning rate scheduler reduced the learning rate by a factor of 2 if the validation loss did not improve for 10 consecutive epochs. All experiments were conducted with a fixed random seed of 42 for reproducibility. The final model for testing was selected based on the epoch that yielded the highest quadratic weighted Cohen's Kappa [7] on the validation set.

Evaluation Metrics. We evaluated classification performance using three metrics: overall accuracy, weighted F1-score, and the quadratic weighted Cohen's Kappa [7]. As KL grades are ordinal, Kappa is the primary metric for our evaluation, It appropriately penalizes large misclassification errors (e.g., KL0 vs 4) more heavily than small errors (e.g., KL1 vs 2) and is standard in OA literature [21,9]. Besides the metrics on the test set, we also report their standard deviation over 5 bootstrapped samples of the test set.

Baselines for Comparison. We compare our proposed method against two baselines: (1) a lightweight Siamese network in Tiulpin et al. [21], which represents the state-of-the-art in efficient, patch-based OA grading, and (2) a standard ResNet-34 [11] trained on a single large ROI, representing a monolithic deep learning

Table 2. Comparison of classification performance on the test set. Our proposed method (Ours (attention)) is compared against key baselines and an ablation study (Ours (mean)). Performance on the full test set ± std from 5 bootstrap samples. Best results for each metric are in **bold**. # parms denotes the number of trainable parameters.

Methods	# parms	input size	Acc.	F1-Score	Kappa
ResNet-34 [11]	21,287,237	≈ 245 (KB)	**0.648 ± 0.003**	0.616 ± 0.007	0.781 ± 0.013
Tiulpin et al. [21]	595,077	≈ 65 (KB)	0.639 ± 0.008	0.632 ± 0.005	0.772 ± 0.012
Ours (mean)	**215,109**	≈ 21 (KB)	0.617 ± 0.003	0.613 ± 0.004	0.788 ± 0.008
Ours (attention)	**248,262**	≈ 21 (KB)	0.644 ± 0.004	**0.652 ± 0.003**	**0.807 ± 0.006**

approach. Additionally, to validate our specific MIL design, we include an ablation study (Ours (mean)) where the attention-based aggregator is replaced with a simple, non-weighted average pooling of patch features.

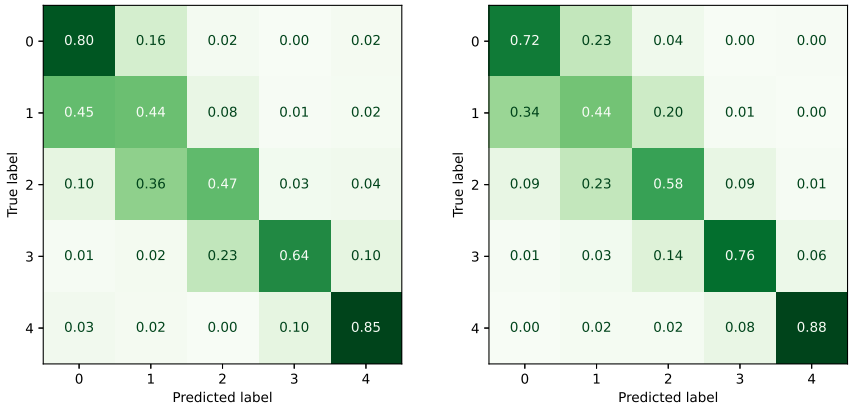
3.3 Classification Evaluation

The quantitative results are summarized in Table 2. Our proposed model, Ours (attention), outperforms all baselines on the primary metric of quadratic weighted Cohen's Kappa and on the F1-score. While ResNet-34 achieves a slightly higher accuracy, this metric could be misleading given the severe class imbalance. The superior Kappa score of our method demonstrates a more clinically meaningful classification ability. Notably, our method also surpasses its direct ablation, Ours (mean), highlighting the critical role of the learned attention mechanism.

To further analyze the performance, Fig. 3 presents the confusion matrices for the baseline Siamese network and our proposed method. Our model demonstrates a visibly stronger diagonal, indicating higher sensitivity for KL grades. The improvement is particularly pronounced for the more advanced and less frequent grades (KL3 and KL4), where the baseline model struggles more. This, combined with the quantitative superiority of the attention model over the mean-pooling ablation, confirms that our framework effectively works in accurate OA grading.

3.4 Qualitative Analysis of Model Explanations

To validate that the improved performance of our model is due to learning clinically relevant features, we visualize its reasoning using our proposed attention-weighted saliency map. Specifically, for each patch, we generate the e-CAM by averaging the normalized heatmaps from five diverse methods in the pytorch-grad-cam library [10]: GradCAM [19], GradCAM++ [5], ScoreCAM [23], LayerCAM [13], and AblationCAM [18]. This ensemble leverages the complementary strengths of each method; for example, LayerCAM provides higher spatial resolution by integrating activations from multiple layers, while AblationCAM is less sensitive to gradient saturation issues, often yielding more complete object localization. These pixel-level e-CAMs provide the local explanations that are subsequently aggregated using the MIL attention weights.



(a) Tiulpin et al. [21] (b) Ours (attention)

Fig. 3. [Reduced Confusion] Normalized confusion matrices for the baseline from Tiulpin et al. [21] (a) and our proposed method (b). Values are normalized by row to represent class-wise sensitivity. Our model shows improved performance, especially for advanced OA (KL3, KL4), with reduced off-diagonal confusion.

Fig. 4 presents a qualitative, case-by-case comparison between the explanations from the baseline model [21] and our method across all five KL grades, with OARSI features [1] from OAI dataset as ground truth. A consistent pattern emerges: the heatmaps of baseline model are often diffuse, correctly identifying the affected knee side but lacking the precision to pinpoint specific pathologies. In contrast, the saliency maps of our model are remarkably focused, highlighting distinct anatomical and pathological features. A detailed analysis follows:

- KL0 (Healthy): While the baseline model shows broad, symmetrical heatmap, the heatmap of our model precisely inspects the joint space and the medial and lateral tibial tubercles. This suggests that our model performs a comprehensive check of all key regions before correctly concluding there are no significant OA features.
- KL1 (Mild): The ground truth indicates medial JSN of grade 1. The baseline model produces a diffuse heatmap in this region. The heatmap of our model is not only more focused on the medial joint space but also highlights the medial tibial tubercle. This aligns with findings that associate tibial spiking with knee OA [17]. Our model appears to have learned that irregularities in this area, including early osteophytes or tibial spiking, are key indicators of knee OA.
- KL2 (Minimal): The ground truth identifies osteophytes in both medial and lateral tibia and medial JSN. The baseline heatmap is again diffuse. Our heatmap, however, produces sharp, distinct focuses on the medial tibial osteophyte and correctly inspects the lateral side for osteophyte formation.

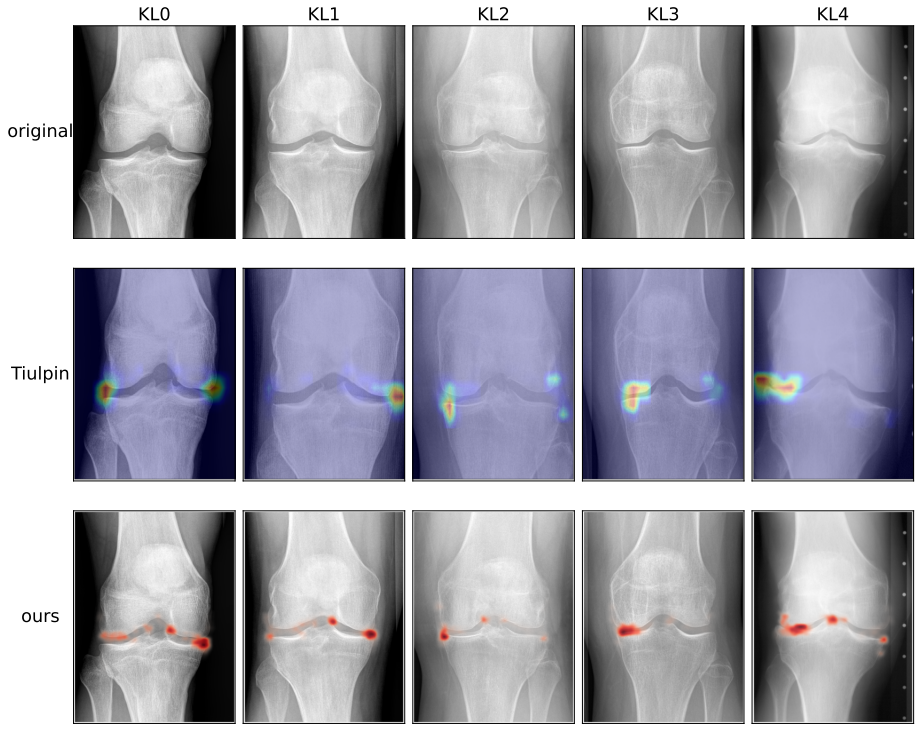


Fig. 4. [Improved Explainability] Qualitative comparison of model explanations across KL grades. Top row:original radiographs for knees with KL grades 0 through 4. **Middle row**: saliency maps from the baseline model by Tiulpin et al. [21], which are often diffuse. **Bottom row**: the attention-weighted saliency maps from our proposed method. Our model explanations are notably more precise, accurately localizing key pathological features such as small osteophytes in early OA (KL1-2) and severe JSN in advanced OA (KL3-4).

This demonstrates a more precise focus on the key features defining this grade.

- KL3 (Moderate): For this case with multiple osteophytes and moderate medial JSN, our model explanation generates a strong, focused heatmap over the narrowed medial joint space and the corresponding osteophytes, whereas the baseline heatmap remains broad and less specific.
- KL4 (Severe): In this example of severe lateral side OA, our model perfectly outlines the narrowed joint space and adjacent osteophytes. This stands in contrast to the unfocused heatmap of baseline, which spills into the background and fails to capture the precise pathological features.

In summary, the visualizations confirm that our anatomy-guided framework learns to identify specific, clinically meaningful features beyond what was possible with the heuristic defined patch baseline. The consistent focus on the tibial

tubercles, an indicator of OA [17], is a novel finding enabled by our anatomy-guided framework. This ability to produce precise and trustworthy explanations for its superior performance is a critical step toward the clinical adoption of automated OA grading systems.

4 Conclusions

This paper introduces an anatomy-guided framework for automated knee osteoarthritis grading. The core of this work is the use of anatomically-focused patches, extracted from bone contours, to direct a multiple instance learning model attention to clinically relevant regions.

This approach achieves promising classification performance on the OAI dataset, reaching a quadratic weighted Cohen's Kappa of 0.807. This result surpasses larger baselines like ResNet-34, while requiring over 85 times fewer parameters (0.25M vs. 21.3M). The efficiency of our model, coupled with its high accuracy, is complemented by its explainability. The attention-weighted saliency maps precisely localize key pathological features such as osteophytes, joint space narrowing and tibial tubercles, confirming that the model learns clinically valid representations.

The small size of our model makes it potentially deployable on resource-constrained hardware, suitable for point-of-care applications. A current dependency of the framework is the initial, accurate detection of bone contours. Future work could thus focus on enhancing the robustness of this anatomical localization step. More broadly, the principles of this anatomy-guided, patch-based framework are not limited to knee OA. This methodology could be extended to other medical imaging tasks where diagnosis relies on localized features within a broader anatomical context, such as grading pathologies in dental radiographs or identifying abnormalities in soft-tissue imaging. This work therefore presents a new direction for creating more accurate, efficient, and trustworthy diagnostic models by directly embedding anatomical information into the learning process.

Acknowledgments. This work was partially supported by the 2025 Polytechnique Montreal Winter Research Internship Program. The data used in this study was obtained from the Osteoarthritis Initiative (OAI), a public-private partnership funded by the National Institutes of Health (NIH) and managed by the Foundation for the NIH with contributions from several private funding partners.

References

1. Altman, R.D., Gold, G.: Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis and cartilage (2007)
2. Antony, J., McGuinness, K., Moran, K., O'Connor, N.E.: Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In: Machine Learning and Data Mining in Pattern Recognition (2017)

3. Bayramoglu, N., Nieminen, M.T., Saarakkala, S.: A lightweight CNN and joint shape-joint space (JS^2) descriptor for radiological osteoarthritis detection. In: Medical Image Understanding and Analysis (2020)
4. Bobek, S., Bałaga, P., Nalepa, G.J.: Towards model-agnostic ensemble explanations. In: International conference on computational science (2021)
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV) (2018)
6. Chen, P., Gao, L., Shi, X., Allen, K., Lin, Y.: Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Computerized Medical Imaging and Graphics (2019)
7. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin (1968)
8. Cui, A., Li, H., Wang, D., Zhong, J., Chen, Y., Lu, H.: Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. EClinicalMedicine (2020)
9. Culvenor, A.G., Engen, C.N., Øiestad, B.E., Engebretsen, L., Risberg, M.A.: Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surgery, Sports Traumatology, Arthroscopy (2015)
10. Gildenblat, J., contributors: Pytorch library for CAM methods. Available at https://github.com/jacobgil/pytorch-grad-cam (2025/06/19)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
12. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning (2018)
13. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing (2021)
14. Kellgren, J., Lawrence, J.: Radiological assessment of osteo-arthrosis. Annals of the Rheumatic Diseases (1957)
15. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE transactions on pattern analysis and machine intelligence (2014)
16. Lindner, C., Thiagarajah, S., Wilkinson, J.M., Wallis, G.A., Cootes, T.F., arcOGEN Consortium, et al.: Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE transactions on medical imaging (2013)
17. Patron, A., Annala, L., Lainiala, O., Paloneva, J., Äyrämö, S.: An automatic method for assessing spiking of tibial tubercles associated with knee osteoarthritis. Diagnostics (2022)
18. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: proceedings of the IEEE/CVF winter conference on applications of computer vision (2020)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (2017)
20. Tiulpin, A., Saarakkala, S.: Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. Diagnostics (2020)

21. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Scientific reports (2018)
22. Tiulpin, A., Thevenot, J., Rahtu, E., Saarakkala, S.: A novel method for automatic localization of joint area on knee plain radiographs. In: Image Analysis (2017)
23. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (2020)
24. Yang, X., Tang, P., Zou, K., Dai, D.: HCGN: Hierarchical convolution and graph network for predicting knee osteoarthritis. In: 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI) (2024)