UNCERTAINTY-QUANTIFIED AND EXPLAINABLE AGE- AND SEX-AWARE CONTRASTIVE LEARNING FOR KNEE OSTEOARTHRITIS CLASSIFICATION

Anonymized Submission at 1st Workshop on Medical Video AI Assessment and Uncertainty Quantification: Bridging Research and Practice

Address - Line 1
Address - Line 2
Address - Line 3

Automated knee osteoarthritis (OA) assessment demands models that are not only accurate, but also trustworthy and demographically robust. To meet this need, we propose XAS-SupCon, an uncertainty-aware, explainable, and age- and sex-aware supervised contrastive learning framework. Using plain knee radiographs from the Osteoarthritis Initiative (OAI), we incorporate age and sex directly into the contrastive objective to strengthen representation learning while mitigating demographic bias. Uncertainty is quantified via Monte Carlo Dropout with risk–coverage analysis and selective prediction. Compared with conventional CNN and contrastive baselines, XAS-SupCon achieves the highest accuracy (0.8419) and F1-score (0.8192), while maintaining lower risk across coverage levels, supporting more reliable and explainable AI-driven knee OA assessment.

1 Introduction

Knee osteoarthritis (OA) is a common and disabling musculoskeletal disorder that leads to chronic pain and reduced mobility [1]. Radiographic evaluation using the Kellgren–Lawrence (KL) grading system [2] remains the clinical standard for assessing OA severity. However, KL grading relies on subjective interpretation, resulting in notable inter- and intra-observer variability. Recent advances in artificial intelligence (AI) have shown strong potential for automating knee OA classification from plain radiographs [3–5]. Nevertheless, most existing AI models are population-agnostic and overlook biological factors such as age and sex that influence OA manifestation and progression. In addition, many AI systems provide deterministic predictions without quantifying predictive uncertainty, limiting their reliability in real-world clinical deployment. In safety-critical medical imaging, uncertainty estimation is essential for identifying ambiguous cases and supporting selective prediction.

To address these limitations, we proposed an uncertainty-aware and explainable age- and sex-aware contrastive AI framework for KL grading in knee OA using plain radiographs. The model learns discriminative image representations through supervised contrastive learning while incorporating age and sex to encourage demographic-aware representation learning. We built our proposed AI methods on the publicly available Osteoarthritis Initiative (OAI) dataset [6]. Through Grad-CAM++ visualization [7, 8], the framework improves model explainability by providing interpretable localization of disease-relevant regions. Our study contributes **clinically** by improving personalized and reliable knee OA assessment, and **technically** by integrating demographic awareness, uncertainty quantification, and explainable contrastive learning into a unified computational framework.

2 Materials and Methods

2.1 Dataset

In our study, we used a dataset from the OAI [6], a publicly available longitudinal cohort study of knee OA that includes imaging data (plain radiographs and MRI), clinical assessments, and patient-reported outcomes. The dataset provides KL grades along with structural features such as joint space narrowing, osteophytes, and subchondral changes. Demographic information, including age and sex, is also available and was incorporated into our study. A total of 4,506 patients with Anteroposterior (AP) fixed-flexion plain knee radiographs were included. Among them, 49.45% were 41–60 years old, 30.49% were 61–70 years old, and 20.06% were older than 70 years. The sex distribution was 41.9% female and 58.1% male. Plain radiographs of both left and right knees were included for each patient. KL grades were binarized into non knee OA (KL 0–1) and knee OA (KL 2–4) following established clinical thresholds [2, 9]. This binarization reduces ambiguity between adjacent grades and supports clinically meaningful decision-making. The final dataset contained 8,949 knee-level samples, including 3,906 knee OA cases and 5,043 non knee OA cases. The average radiograph resolution was 766 pixels in width and 630 pixels in height.

2.2 Data Preprocessing

Original AP fixed-flexion knee radiographs contained both knee structures and surrounding background, which introduced noise and reduced model focus [10]. Therefore, region-of-interest (ROI) extraction was performed to isolate the knee joint area for analysis. We manually annotated 1,364 radiographs to localize the knee joint regions. These annotations were used to train a YOLOv11 [11] object detection model for automatic knee ROI detection. After training, the model was applied to radiographs from 4,506 patients to detect and crop left and right knee ROIs. This automated detect-and-crop pipeline produced 8,949 cropped AP knee images. By isolating anatomically relevant regions and removing background structures, this preprocessing step reduced noise and ensured consistent inputs for subsequent model training.

2.3 XAS-SupCon: Age- and Sex-aware Supervised Contrastive Learning

Supervised Contrastive Learning (SupCon) [12] extends contrastive learning by leveraging class labels within the loss function. While standard self-supervised contrastive learning [13] pulls together augmented views of the same image and pushes apart different images, SupCon further groups embeddings from the same class and separates those from different classes. This encourages representations that better reflect class-level semantics. The SupCon loss is defined as:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (1)$$

where I denotes all augmented samples in the batch ($|I| = 2N$ for N images), $z_i = Proj(Enc(x_i))$ represents the normalized projection feature, and τ is a temperature parameter. $P(i)$ denotes same-class positives for anchor i (excluding i), and $A(i)$ includes all other samples in the batch.

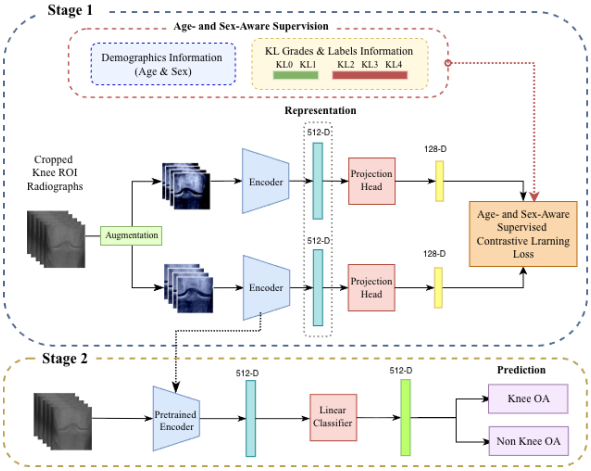

Fig. 1: Overview of the proposed XAS-SupCon framework.

Figure 1 illustrates the proposed Age- and Sex-Aware SupCon (XAS-SupCon). The framework consists of two stages. In **Stage 1**, representations are learned by minimizing a demographic-aware contrastive loss. In **Stage 2**, the encoder is frozen and a linear classifier is trained for binary OA classification. At the representation learning stage, a batch of input images is processed by a data-augmentation module $\mathrm{Aug}(\cdot)$ (cropping, flipping, color jittering), which generates two stochastic views each image, $\tilde{x} = \mathrm{Aug}(x)$. Both views are passed through the shared *Encoder* $\mathrm{Enc}(\cdot)$ to obtain a 512-dimensional normalized representation $r = \mathrm{Enc}(\tilde{x}) \in \mathbb{R}^{D_E} (D_E = 512)$. This representation is then mapped by the *Projection Head* $\mathrm{Proj}(\cdot)$ to a 128-dimensional feature $z = \mathrm{Proj}(r) \in \mathbb{R}^{D_P} (D_P = 128)$, which is also normalized for the contrastive objective. Our XAS-SupCon loss function leverages both labels and demographic information to capture patient-specific variation during representation learning. In plain knee radiographs, images from patients with similar ages and the same sex may exhibit more comparable structural characteristics, suggesting that demographic similarity could influence representation learning. Motivated by this observation, we hypothesize that incorporating demographic similarity into the contrastive objective can improve the quality and stability of learned representations.

To achieve this, we introduce a weighting scheme applied to same-class positive pairs within each batch. Instead of assigning equal contribution to all positives, XAS-SupCon adjusts their influence according to age proximity and sex concordance. Specifically, anchor–positive similarities are scaled by a demographic-aware weight, allowing age-similar and same-sex pairs to contribute more strongly, while pairs with larger age gaps exert less influence. This preserves the contrastive structure while embedding demographic constraints into the representation space, promoting more robust performance across age and sex subgroups. The corresponding XAS-SupCon objective is defined as:

$$\mathcal{L}_{\text{XAS-SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(w_{ip}(z_i \cdot z_p)/\tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \quad (2)$$

where the weight term w_{**} scales the unweighted similarity $z_i \cdot z_p$ such that pairs with more similar demographic attributes contribute more, is calculated as follows:

$$w_{ij} = \left(1 + \lambda_a \, \mathrm{AgeSim}_{ij}\right)\left(1 + \lambda_s \, \mathrm{SexMask}_{ij}\right) \quad (3)$$

where λ_a and λ_s control the contributions of age and sex, respectively. In our study, we set $\lambda_a = \lambda_s = 0.05$. The age similarity from the absolute age gap $\Delta_{ij} = |\mathrm{Age}_i - \mathrm{Age}_j|$ using a Gaussian kernel [14, 15] is quantified as:

$$\mathrm{AgeSim}(\Delta_{ij}) = \exp\left(-\frac{\Delta_{ij}^2}{2\sigma^2}\right) \quad (4)$$

where σ sets the age scale. To calibrate σ, we compute the empirical distribution of pairwise age gaps, select a target

quantile Δ^* and a target similarity s^*, and choose σ such that $\text{age\_sim}(\Delta^*) = s^*$. This defines the age gap threshold below which the similarity becomes meaningful. In our study, using the 60% quantile of the age-gap distribution ($\Delta = 12.0$ years) and a target similarity of $s = 0.5$ yielded $\sigma \approx 10.2$ years. Clinically, knee OA progresses slowly, so age gaps of 10 years are clinically meaningful, whereas smaller gaps are less likely to alter the appearance of plain knee radiograph images. Furthermore, sex concordance is represented by a binary indicator mask defined as:

$$\text{SexMask}(i,j) = \begin{cases} 1, & \text{if } i \text{ and } j \text{ have the same sex,} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

After pretraining with XAS-SupCon, we discard the $Proj(\cdot)$ and use the frozen $Enc(\cdot)$ as a feature extractor. For each image x, we compute the representation $r = Enc(x) \in \mathbb{R}^{D_E}$ and train a *Linear Classifier* on r for OA versus non-OA prediction. Only the classifier parameters are updated in this stage and the $Enc(\cdot)$ remains frozen. In addition, using encoder features rather than projection features is standard practice for downstream classification and has been shown to provide superior performance [13].

2.4 Uncertainty Quantification

To assess predictive reliability, we estimate model uncertainty using Monte Carlo (MC) Dropout [16]. During inference, dropout layers remain active, and each input image is forwarded through the network T times to obtain a distribution of predictive probabilities. The mean probability is used as the final prediction, while predictive entropy is adopted as the primary uncertainty measure. In our experiments, we use $T = 50$ stochastic forward passes with dropout probability $p = 0.1$. To evaluate uncertainty-aware performance, we perform risk–coverage analysis [17]. Samples are ranked by increasing uncertainty, where coverage denotes the proportion of retained (low-uncertainty) samples. Risk is defined as the classification error among the retained samples. We report the Area Under the Risk–Coverage Curve (AURC) and Coverage@95Acc, which quantifies the maximum retained sample proportion while maintaining at least 95% accuracy.

2.5 Experimental Setup and Evaluation

We compared XAS-SupCon with three baseline models: (1) ResNet-18 [18] as a standard convolutional neural network, (2) SimCLR [13] as a self-supervised contrastive learning approach, and (3) SupCon [12] as a supervised contrastive learning method. This comparison allows assessment of the impact of incorporating age and sex information into the contrastive objective. The dataset was split at the patient level into 80% training and 20% testing sets. For XAS-SupCon, SupCon, and SimCLR, the encoder (ResNet-18 backbone) was first pretrained using the training set. The encoder was then frozen, and a linear classifier was trained using five-fold cross-validation to determine the best model configuration. Performance was evaluated using accuracy, precision, recall, and F1-score. Subgroup analyses were conducted across age and sex categories. In addition, uncertainty-aware performance was assessed using risk–coverage analysis and Coverage@95Acc. Grad-CAM++ [7] was applied for visualization to assess whether predictions aligned with anatomically relevant regions. All experiments were implemented in Python (3.10.12) and conducted on a compute node equipped with three NVIDIA Quadro RTX 8000 GPUs (48 GB each), dual 20-core CPUs, and 376 GB RAM.

3 Results

Table 1 and **Table 2** demonstrate that XAS-SupCon achieves the best overall performance on the full test set and maintains consistent advantages across demographic subgroups. On the overall dataset, it obtains the highest accuracy (0.8419), F1-score (0.8192), and recall (0.8207), indicating improved sensitivity in detecting knee OA cases. Across age and sex subgroups, XAS-SupCon achieves the strongest overall results, with particularly notable gains in the older (Age $\geq$ 61) group. Although precision in the female subgroup is slightly lower than SimCLR, XAS-SupCon attains the highest accuracy, recall, and F1-score in most subgroup settings, suggesting improved robustness and generalization across diverse patient populations.

Table 1: Performance comparison of the AI models.

AI Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
ResNet-18	0.8006	0.8011	0.7222	0.7596
SimCLR	0.7642	0.8211	0.5877	0.6851
SupCon	0.8318	0.8077	0.8067	0.8072
XAS-SupCon	**0.8419**	**0.8176**	**0.8207**	**0.8192**

Figure 2 and **Table 3** demonstrate that XAS-SupCon delivers the strongest uncertainty-aware performance among all evaluated models. In the risk–coverage analysis, XAS-SupCon consistently maintains the lowest risk across nearly all coverage levels, indicating more reliable predictions when selectively retaining high-confidence samples. In contrast, baseline models such as SimCLR exhibit substantially higher risk, even at low coverage levels. At a 95% accuracy threshold, XAS-SupCon achieves the highest overall Coverage@95Acc (0.5659), retaining the largest proportion of samples while preserving strict accuracy requirements. It also attains the highest coverage in multiple demographic subgroups, including Age$\leq$60 (0.6296), Female (0.5881), and Male (0.5495), while remaining competitive in the Age$\geq$61 subgroup. Overall, these results indicate that XAS-SupCon not only improves predictive performance but also enhances

selective reliability and demographic consistency, supporting its suitability for risk-aware medical AI applications.

Table 2: Subgroup performance comparison of AI models.

Subgroup	*AI Model*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Age ≤ 60	ResNet-18	0.8242	0.7752	0.7278	0.7508
	SimCLR	0.7909	0.8233	0.5413	0.6531
	SupCon	0.8443	0.7895	0.7798	0.7846
	XAS-SupCon	**0.8465**	**0.7908**	**0.7859**	**0.7883**
Age ≥ 61	ResNet-18	0.7856	0.8329	0.7247	0.7750
	SimCLR	0.7374	0.8198	0.6211	0.7068
	SupCon	0.8193	0.8206	0.8260	0.8233
	XAS-SupCon	**0.8373**	**0.8366**	**0.8458**	**0.8412**
Female	ResNet-18	0.8023	0.7799	0.7477	0.7635
	SimCLR	0.7759	**0.8224**	0.6055	0.6975
	SupCon	0.8317	0.8041	0.8005	0.8023
	XAS-SupCon	**0.8376**	0.8013	**0.8234**	**0.8122**
Male	ResNet-18	0.8086	0.8486	0.6986	0.7663
	SimCLR	0.7487	0.8193	0.5652	0.6690
	SupCon	0.8320	0.8121	0.8145	0.8133
	XAS-SupCon	**0.8477**	**0.8393**	**0.8174**	**0.8282**

Figure 3 presents Grad-CAM++ [7] visualizations generated by different AI models using cropped knee ROI radiographs. Both examples correspond to confirmed knee OA cases, with visible joint space narrowing and marginal osteophytes. Compared with ResNet-18 and SimCLR, SupCon and XAS-SupCon demonstrate more concentrated and anatomically meaningful attention over the tibiofemoral joint space and osteophyte margins, which are clinically relevant regions for OA assessment. Two orthopedic surgeons independently reviewed the visualization overlays and reported that the heatmaps produced by XAS-SupCon were better aligned with established clinical knowledge than those of the other models, providing clearer localization of pathological structures and more interpretable predictions. These findings indicate that incorporating supervised and demographic-aware contrastive learning enhances localization consistency and improves clinical interpretability. All Python codes are available at this GitHub repository [*removed for now because of Anonymized Submission*].

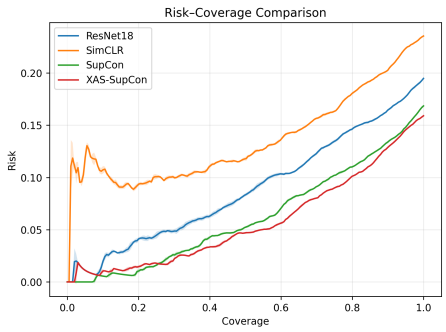

Fig. 2: Risk–Coverage comparison of AI models. Solid lines denote mean risk across 10 randomly selected seeds, and shaded regions indicate standard deviation.

Table 3: Coverage@95Acc comparison across groups.

AI Model	*Overall*	*Age≤60*	*Age≥61*	*Female*	*Male*
ResNet-18	0.3156	0.3415	0.2952	0.2270	0.3919
SimCLR	0.0073	0.0100	0.0180	0.0215	0.0078
SupCon	0.4927	0.4750	**0.5387**	0.4990	0.4219
XAS-SupCon	**0.5659**	**0.6296**	0.4332	**0.5881**	**0.5495**

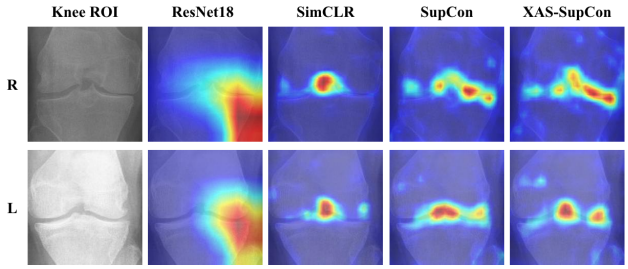

Fig. 3: Grad-CAM++ visualizations for each AI model. Rows correspond to right (R) and left (L) knees, respectively. Knee ROI refers to cropped plain radiograph images of the knee ROI.

4 Discussion, Conclusion, and Outlook

The proposed XAS-SupCon framework demonstrates that incorporating demographic information, such as age and sex, into supervised contrastive learning improves both predictive performance and model interpretability for knee OA classification. By integrating demographic awareness into the representation learning process, the model captures clinically relevant variation across patient populations while maintaining strong overall accuracy. In addition, uncertainty quantification through MC Dropout and risk–coverage analysis shows that XAS-SupCon provides more reliable selective predictions, achieving lower risk across coverage levels and higher Coverage@95Acc compared with baseline methods. These findings indicate that demographic-aware contrastive learning enhances not only classification accuracy but also predictive reliability in safety-critical clinical settings. Grad-CAM++ visualizations further indicate that XAS-SupCon consistently highlights anatomically relevant OA-associated regions, reinforcing model transparency and improving clinical interpretability. Future work will extend this framework by integrating multimodal data sources, including MRI and patient-reported outcomes, to improve generalization across diverse cohorts. We also plan to expand representation across underrepresented demographic groups and explore federated and advanced uncertainty-aware learning strategies to facilitate robust and trustworthy deployment in real-world healthcare environments.

5 References

[1] Anna Litwic, Mark H Edwards, Elaine M Dennison, and Cyrus Cooper, "Epidemiology and burden of osteoarthritis," *British medical bulletin*, vol. 105, no. 1, pp. 185–199, 2013.

[2] Jonas H Kellgren, JS Lawrence, et al., "Radiological assessment of osteo-arthrosis," *Ann Rheum Dis*, vol. 16, no. 4, pp. 494–502, 1957.

[3] Facundo Manuel Segura, Florencio Pablo Segura, María Paz Lucero Zudaire, and Florencio Vicente Segura, "Advances in artificial intelligence for automated knee osteoarthritis classification using the ikdc system," *European Journal of Orthopaedic Surgery & Traumatology*, vol. 35, no. 1, pp. 1–7, 2025.

[4] Nickolas Littlefield, Soheyla Amirian, Jacob Biehl, Edward G Andrews, Michael Kann, Nicole Myers, Leah Reid, Adolph J Yates Jr, Brian J McGrory, Bambang Parmanto, et al., "Generative ai in orthopedics: an explainable deep few-shot image augmentation pipeline for plain knee radiographs and kellgren-lawrence grading," *Journal of the American Medical Informatics Association*, vol. 31, no. 11, pp. 2668–2678, 2024.

[5] Haoming Zhao, Liang Ou, Ziming Zhang, Le Zhang, Ke Liu, and Jianjun Kuang, "The value of deep learning-based x-ray techniques in detecting and classifying kl grades of knee osteoarthritis: a systematic review and meta-analysis," *European Radiology*, vol. 35, no. 1, pp. 327–340, 2025.

[6] Osteoarthritis Initiative, "Osteoarthritis initiative (oai) dataset," `https://nda.nih.gov/oai`, 2025, Accessed: 2025-10-28.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[9] Weiya Zhang, Daniel F McWilliams, Sarah L Ingham, Sally A Doherty, Stella Muthuri, Kenneth R Muir, and Michael Doherty, "Nottingham knee osteoarthritis risk prediction models," *Annals of the rheumatic diseases*, vol. 70, no. 9, pp. 1599–1604, 2011.

[10] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific reports*, vol. 8, no. 1, pp. 1727, 2018.

[11] Rahima Khanam and Muhammad Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[14] David S Broomhead and David Lowe, "Radial basis functions, multi-variable functional interpolation and adaptive networks," Tech. Rep., 1988.

[15] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, 1997.

[16] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[17] Yonatan Geifman and Ran El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

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