Classification models for **arthropathy grades** of **multiple joints** based on **hierarchical** **continual learning**

Yu Rang Park,
Digital Healthcare Lab
Department of Biomedical systems informatics
Yonsei University College of Medicine
yurangpark@yuhs.ac, http://dhlab.org

4 grades
(normal, mild, moderate, severe)

(Internal) knee, elbow, ankle, shoulder
(External) knee, hip

Classification models for arthropathy grades of multiple joints based on hierarchical continual learning

3-level hierarchy
(L1: normal, mild /
L2: normal, low, high /
L3: normal, mild, moderate, severe)

Class-incremental learning model
(DER: Dynamically Expandable Representation)

La radiologia medica
https://doi.org/10.1007/s11547-025-01974-4

ORIGINAL ARTICLE

Classification models for arthropathy grades of multiple joints based on hierarchical continual learning

Bong Kyung Jang[1] · Shiwon Kim[1,2] · Jae Yong Yu[1] · JaeSeong Hong[1] · Hee Woo Cho[3] · Hong Seon Lee[3] · Jiwoo Park[3] · Jeesoo Woo[5] · Young Han Lee[3,4] · Yu Rang Park[1,2,3,4]

Received: 28 June 2024 / Accepted: 14 February 2025

Abstract

Purpose To develop a hierarchical continual arthropathy classification model for multiple joints that can be updated continuously for large-scale studies of various anatomical structures.

Challenges in radiographic assessment of arthropathy

- **Arthropathy** is a various condition that affects the joints, including osteoarthritis, inflammatory arthritis such as rheumatoid arthritis and psoriatic arthritis, lupus arthritis, rotator cuff arthropathy, gouty arthritis.
- **Osteoarthritis (OA)** is a musculoskeletal disorder that primarily affects weight-bearing joints such as the knee and ankle joints, as well as non-weight-bearing joints such as the elbow joints and shoulder.
- Radiography is commonly used to evaluate the joints of the musculoskeletal system, with joint radiography being the primary imaging modality for suspected arthropathy or OA [1].
- **Radiographic assessment of OA severity** is critical for clinical decision making, including diagnosis, treatment monitoring, and research [2].
- However, radiographic classification of OA severity is **a time-consuming task** requiring assessments of joint space width, osteophytes, and subchondral sclerosis.
- Moreover, it is **a subjective evaluation**, coupled with vaguely defined features at various stages of OA progression, resulting in **low inter-observer reliability** [3, 4].

1. Kohn MD, Sassoon AA, Fernando ND (2016) Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res 474(8):1886–1893. https://doi.org/10.1007/s11999-016-4732-4
2. Croft P (2005) An introduction to the atlas of standard radiographs of arthritis. Rheumatology. https://doi.org/10.1093/rheumatology/kei051
3. Culvenor AG, Engen CN, Oiestad BE, Engebretsen L, Risberg MA (2015) Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surg Sports Traumatol Arthrosc.
4. Damen J, Schiphof D, Wolde ST, Cats HA, Bierma-Zeinstra SM, Oei EH (2014) Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study. Osteoarthr Cartil 22(7):969–974.

Limitations of existing AI classification models

- In light of these challenges, **AI is being used to classify the severity of OA** [5, 6], and deep learning tools have been developed to diagnose and assess OA [7-10].
- Nevertheless, the majority of AI models are **only capable of assessing a singular joint,** lacking scalability across multiple joints; In this work, we apply ***continual learning***—a deep learning strategy designed to accommodate dynamic data distributions [11]—to OA severity grading, enabling **simultaneous classification of multiple joints with various morphologies**.
- Although some studies have applied continual learning to medical imaging, existing approaches **overlook the hierarchical structure inherent in many medical annotations**, resulting in low performance [12, 13].
- In this study, we developed and validated a hierarchical and continual learning approach, **Hierarchical Dynamically Expandable Representation (Hi-DER)**, for **enhanced model scalability** and **effective utilization of hierarchical information** in arthropathy classification.

5. Thomas KA, Kidzinski L, Halilaj E, Fleming SL, Venkataraman GR, Oei EHG, Gold GE, Delp SL (2020) Automated classification of radiographic knee osteoarthritis severity using deep neural networks. Radiol Artif Intell 2(2):e190065.
6. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, Cho K, Chang G, Deniz CM (2020) Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. Radiology.
7. Ureten K, Arslan T, Gultekin KE, Demir AND, Ozer HF, Bilgili Y (2020) Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiol 49(9):1369–1374.
8. von Schacky CE, et al. (2020) Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 295(1):136–145. https://doi.org/10.1148/radiol.2020190925
9. Wang Y, Bi Z, Xie Y, Wu T, Zeng X, Chen S, Zhou D (2022) Learning from highly confident samples for automatic knee osteoarthritis severity assessment: data from the osteoarthritis initiative. IEEE J Biomed Health Inform 26(3):1239–1250.
10. Kijowski R, Fritz J, Deniz CM (2023) Deep learning applications in osteoarthritis imaging. Skeletal Radiol 52(11):2225–2238. https://doi.org/10.1007/s00256-023-04296-6
11. Wang L, Zhang X, Su H, Zhu J (2024) A comprehensive survey of continual learning theory method and application. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2024.3367329
12. Pianykh OS, Langs G, Dewey M, Enzmann DR, Herold CJ, Schoenberg SO, Brink JA (2020) Continuous learning AI in radiology: implementation principles and early applications. Radiology 297(1):6–14. https://doi.org/10.1148/radiol.2020200038
13. Dimitrovski I, Kocev D, Loskovska S, Dzeroski S (2011) Hierarchical annotation of medical images. Pattern Recogn 44(10–11):2436–2449. https://doi.org/10.1016/j.patcog.2011.03.026

Model Development

Internal dataset

Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Randomize dataset into stratified 5 folds

(Training set 80% , Test set 20%)

Hierarchically dynamically expandable representation

Step1 Knee → Step2 Elbow → Step3 Ankle → Step4 Shoulder

Hierarchical Classification Model

- Abnormal (L1)
- Low-grade or high-grade (L2)
- Specific grade (L3)

External Validation

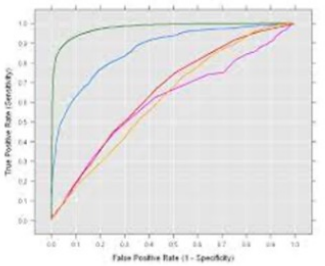
External hip dataset (Training set 80%, Test set 20%) ← Unseen joint

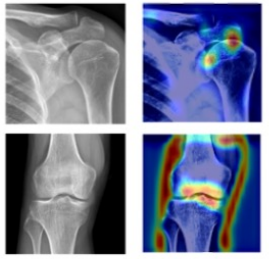
External knee dataset (For testing only) ← Seen joint

Hierarchical Classification Model

- Step5 Hip → Hierarchical Classification Model
- Knee

Evaluation





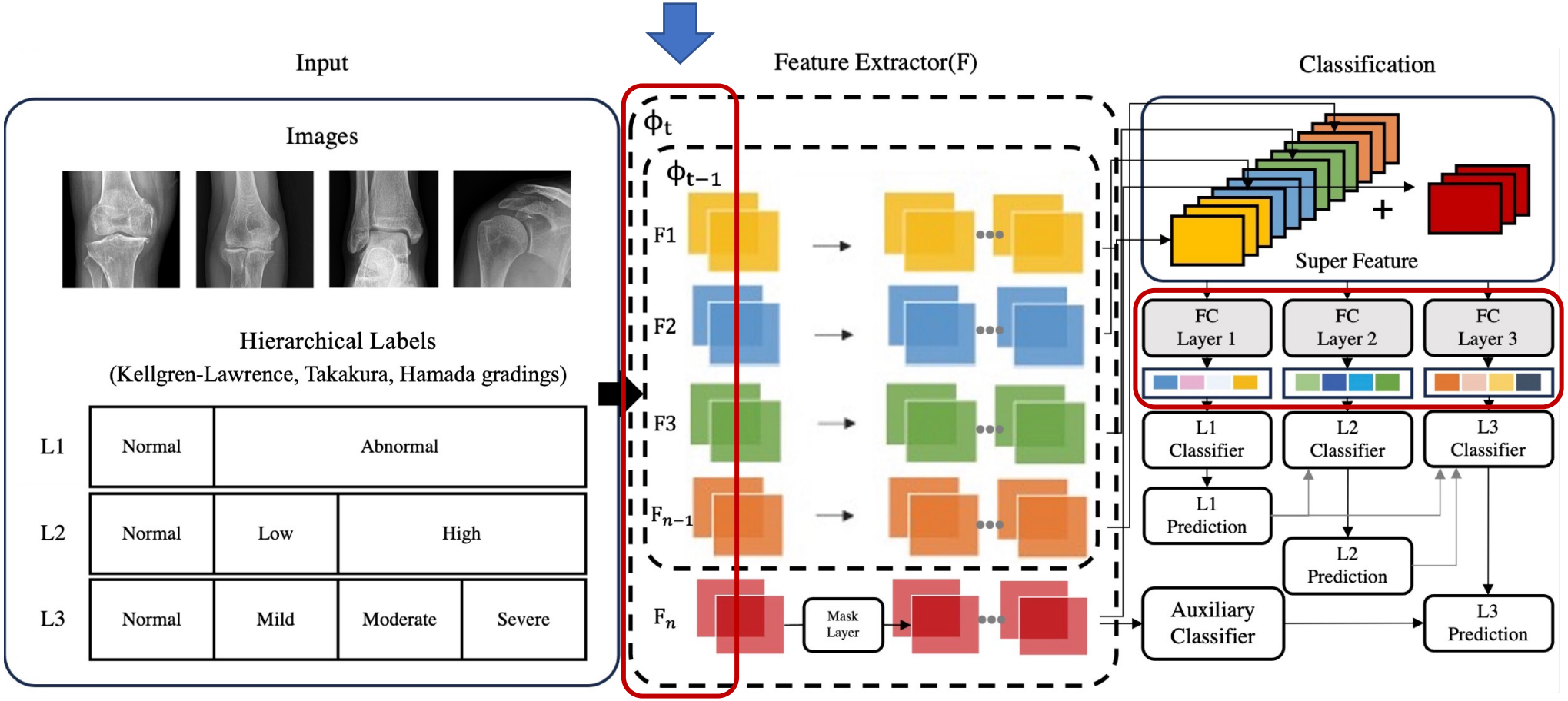
- This study included radiographs of knee, elbow, ankle, shoulder of adult patients (over 18 years of age) from **Sinchon Severance Hospital** in inpatient and outpatient settings from July 1, 2022 to December 31, 2022. This resulted in a total of **934 AP radiographs: 274 knee, 209 elbow, 249 ankle, and 202 shoulder.**
- Radiologic grading of OA was performed by two musculoskeletal imaging fellowship-trained radiologists blinded to clinical information and other imaging results: KL grading for knee and elbow [14], Takakura grading for ankle [15], and Hamada grading for shoulder [16].
- **For external validation**, **125 hip AP radiographs** of adult patients were collected from **Yongin Severance Hospital** in inpatient and outpatient settings from January 1, 2022 to December 31, 2022, and **312 knee AP radiographs** of adult patients were collected from **Gangnam Severance Hospital** in inpatient and outpatient settings from January 1, 2023 to June 30, 2023. Both external datasets were graded using KL grading.
- We used a **three-level hierarchical labeling strategy** based on the annotations of the radiologists:

L1	Normal	Abnormal		
L2	Normal	Low	High	
L3	Normal	Mild	Moderate	Severe

14. Kellgren JH, Lawrence JS (1957) Radiological assessment of Osteo-arthrosis. Ann Rheum Dis 16(4):494–502. https://doi.org/10.1136/ard.16.4.494
15. Takakura Y, Tanaka Y, Kumai T, Tamai S (1995) Low tibial osteotomy for osteoarthritis of the ankle. Results of a new operation in 18 patients. J Bone Joint Surg Br 77(1):50–54
16. Brolin TJ, Updegrove GF, Horneff JG (2017) Classifications in brief: hamada classification of massive rotator cuff tears. Clin Orthop Relat Res 475(11):2819–2823. https://doi.org/10.1007/s11999-017-5340-7

- The network is **trained continually in multiple steps**, expanding the feature extractor at each step. It preserves the old knowledge from previous steps while acquiring new information with a new feature extractor.

At each incremental step, a new feature extractor is generated and integrated with the previous feature extractors. **Each step includes different joints.**



A three-layered training strategy for capturing **hierarchical information** between outcomes.

Objective function

$$loss_{total} = loss_{clf} + \lambda_a loss_{aux} \; (\lambda_a = 0 \; in \; initial \; training)$$

Discriminates the previous and current steps

Auxiliary loss

$$loss_{aux} = -\sum_{i=1}^{|t|+1} y_a^i \log(p_a^i)$$

$$y_a = \{1 \dots |t| \; \textbf{if} \; new \; class \; \textbf{else} \; 0\}$$

Classification loss at each hierarchical layer

Hierarchical classification loss = layer loss ($lloss$) + dependency loss ($dloss$)

$$lloss_l = -\sum_{j=1}^{|l|} y_{tl}^j \log\left(p_{tl}^j\right), \qquad dloss_l = (w_{l-1})^{\mathbb{D}_l \mathbb{I}_{l-1}} (w_l)^{\mathbb{D}_l \mathbb{I}_l} - 1$$

$$\mathbb{D}_l = \left\{1 \; \textbf{if} \; P_{tl} \not\Rightarrow P_{t(l-1)} \; \textbf{else} \; 0\right\}, \qquad \mathbb{I}_l = \{1 \; \textbf{if} \; P_{tl} \neq y_{tl} \; \textbf{else} \; 0\}$$

Forces the hierarchical dependency

Require:
x_t^i is the input image, y_{tl}^i is the true hierarchical label, F_t is the newly created feature extractor at step t, Φ_t is the super feature at step t, H_l is the hierarchical classifier at hierarchical layer l, H^a is the auxiliary classifier, ch is the number of channels of convolutional layer c, FC is the fully connected layer, w is the dependence punishment weight, λ is the hyper-parameter, M_t is the rehearsal memory at step t, and m is the memory sample size per class.

Input: $D_t = \{x_t^i, y_{tl}^i\}; y_{tl}^i \in Y^{tl}; i = 1, \dots, N; t = 1, \dots, T; l = 1, \dots, L$
Output: predicted hierarchical labels of input x_t^i

for *each incremental step* $t = 1, \dots, T$ **do**

Feature Extraction:
append(D_t, M_{t-1}) $(M_0 = \emptyset)$

Model Pruning:
$F_t^P \leftarrow$ add_mask(F_t)
$loss_{spr} \leftarrow \frac{\sum_{c=1}^{|c|} \|mask_{c-1}\| \, \|mask_c\|}{\sum_{c=1}^{|c|} ch_{c-1} ch_c}$

Auxiliary Loss:
$p_a \leftarrow \text{Softmax}(H_t^a(F_t^P))$
$loss_{aux} \leftarrow -\sum_{i=1}^{|t|+1} y_a^i \log(p_a^i)$

$y_a = \{1 \dots |t|$ **if** new class **else** $0\}$

$\Phi_t^P \leftarrow \text{concatenate}([\Phi_{t-1}^p, F_t^P])$ $(\Phi_1^P = F_1^P)$

Classification:
for *each hierarchical layer* $l = 1, \dots, L$ **do**
generate FC_{tl}
$p_{tl} \leftarrow \text{Softmax}(H_{tl}(\Phi_t^P))$
$P_{tl} \leftarrow \text{argmax}(p_{tl})$

Hierarchical Loss Network:
$lloss_l \leftarrow -\sum_{j=1}^{|l|} y_{tl}^j \log(p_{tl}^j)$
$dloss_l \leftarrow -(w_{l-1})^{\mathbb{D}_l \mathbb{I}_{l-1}} (w_l)^{\mathbb{D}_l \mathbb{I}_l}$

$\mathbb{D}_l = \{1$ **if** $P_{tl} \nRightarrow P_{t(l-1)}$ **else** $0\}$
$\mathbb{I}_l = \{1$ **if** $P_{tl} \neq y_{tl}$ **else** $0\}$

$loss_{clf} \leftarrow lloss_l(y_{tl}, p_{tl}) + dloss_l(P_{tl}, y_{tl})$
$loss_{total} \leftarrow loss_{clf} + \lambda_a loss_{aux} + \lambda_s loss_{spr}$ ($\lambda_a = 0$ in initial training)

$M_t \leftarrow$ construct_rehearsal_exemplar(m)

Supplementary Fig.1 Pseudo code of the Hi-DER model.

Hierarchical classification performance of arthropathy grading in multiple joints

Table 2 Overview of the Hi-DER test performance according to hierarchical levels and anatomical locations

Performance Metric	L1: Abnormal Classifications					L2: Low-grade or High-grade Classifications					L3: Specific Grade Classifications				
	Anatomical Location					Anatomical Location					Anatomical Location				
	Knee	Elbow	Ankle	Shoulder	Weighted Average	Knee	Elbow	Ankle	Shoulder	Weighted Average	Knee	Elbow	Ankle	Shoulder	Weighted Average
Accuracy (%)	97.78	89.50	84.72	87.62	90.32	86.15	74.16	77.13	70.83	77.64	73.06	63.67	67.04	66.33	67.84
PPV	0.979	0.897	0.865	0.877	0.908	0.856	0.690	0.780	0.718	0.768	0.752	0.609	0.618	0.649	0.661
NPV	0.996	0.986	0.979	0.987	0.987	0.990	0.984	0.977	0.976	0.982	0.982	0.979	0.977	0.979	0.979
Sensitivity	0.978	0.895	0.847	0.876	0.902	0.861	0.742	0.772	0.707	0.777	0.732	0.634	0.670	0.662	0.678
Specificity	0.996	0.987	0.975	0.981	0.985	0.981	0.977	0.977	0.976	0.978	0.979	0.976	0.967	0.972	0.974
F1 score	0.978	0.895	0.847	0.869	0.900	0.835	0.670	0.766	0.703	0.751	0.716	0.610	0.635	0.640	0.654

Performance metrics

- **Accuracy (%)** = (TP + TN) / (TP + TN + FP + FN)
- **Positive Predictive Value (PPV; Precision)** = TP / (TP + FP)
- **Negative Predictive Value (NPV)** = TN / (TN + FN)
- **Sensitivity (True Positive Rate; Recall)** = TP / (TP + FN)
- **Specificity (True Negative Rate)** = TN / (TN + FP)
- **F1 score** = 2 × (PPV × Sensitivity) / (PPV + Sensitivity) = 2 × TP / (2TP + FP + FN)

Comparison with other deep learning algorithms

Table 3 Algorithm performance comparison: ResNet-50, DER, and Hi-DER

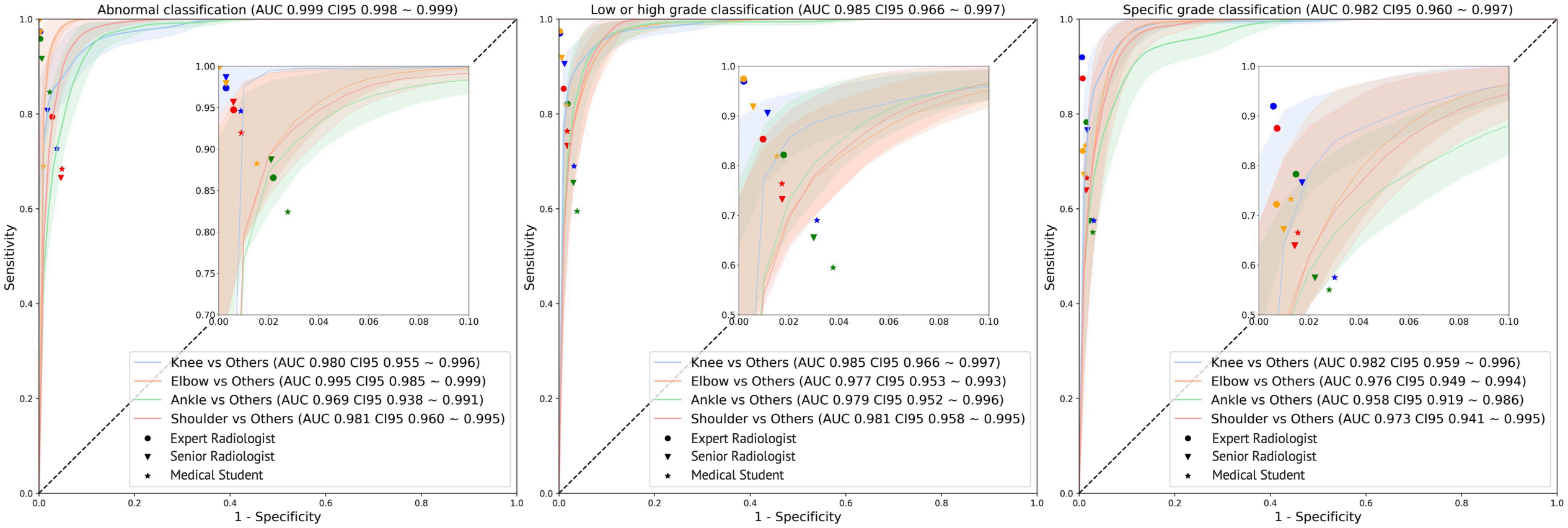
Anatomical Location	Hierarchical Level	ResNet-50				DER				Hi-DER			
		Accuracy (%)	Sensitivity	Specificity	AUC	Accuracy (%)	Sensitivity	Specificity	AUC	Accuracy (%)	Sensitivity	Specificity	AUC
Knee	L1	99.64	0.996	0.998	0.999	88.38	0.884	0.873	0.955	97.76	0.978	0.983	0.997
	L2	91.39	0.914	0.961	0.980	77.19	0.772	0.856	0.881	90.62	0.906	0.942	0.975
	L3	77.13	0.771	0.524	0.935	61.77	0.618	0.869	0.837	81.28	0.813	0.940	0.959
Knee Elbow-	L1	93.70	0.937	0.979	0.995	77.50	0.775	0.920	0.944	94.54	0.945	0.984	0.994
	L2	83.82	0.838	0.971	0.968	60.71	0.607	0.921	0.890	81.30	0.813	0.949	0.979
	L3	71.43	0.714	0.958	0.954	54.40	0.544	0.933	0.881	72.47	0.725	0.958	0.971
Knee Elbow Ankle	L1	90.49	0.905	0.980	0.990	70.74	0.707	0.937	0.942	92.00	0.920	0.986	0.992
	L2	76.42	0.764	0.966	0.979	55.58	0.556	0.942	0.904	79.87	0.799	0.971	0.982
	L3	64.00	0.640	0.958	0.968	50.76	0.508	0.947	0.889	72.13	0.721	0.972	0.973
Knee Elbow Ankle Shoulder	L1	89.96	0.900	0.984	0.992	68.60	0.686	0.957	0.933	90.32	0.902	0.985	0.999
	L2	76.26	0.763	0.975	0.980	57.61	0.576	0.961	0.909	77.64	0.777	0.978	0.985
	L3	54.40	0.644	0.969	0.969	46.93	0.469	0.956	0.893	67.84	0.678	0.974	0.982

Comparison methods

- **ResNet-50** [17] : A static convolution network used as the backbone feature extractor of our model
- **Dynamically Expandable Representation (DER)** [18] : A continual learning method that does not incorporate hierarchical information

17. He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr) pp 770–778. https://doi.org/10.1109/Cvpr.2016.90
18. Yan SP, Xie JW, He XM (2021) DER: Dynamically expandable representation for class incremental learning. Proc Cvpr Ieee. https://doi.org/10.1109/Cvpr46437.2021.00303

Comparison with grading predictions of radiologists



(a) L1 classification (b) L2 classification (c) L3 classification

Fig.2 Visual comparison of the **receiver-operating characteristic (ROC) curves** and **grading results of the radiologists**. The classification results of the radiologists (expert, senior, and medical student) are plotted alongside the ROC curves.

Model explainability: Grad-CAM

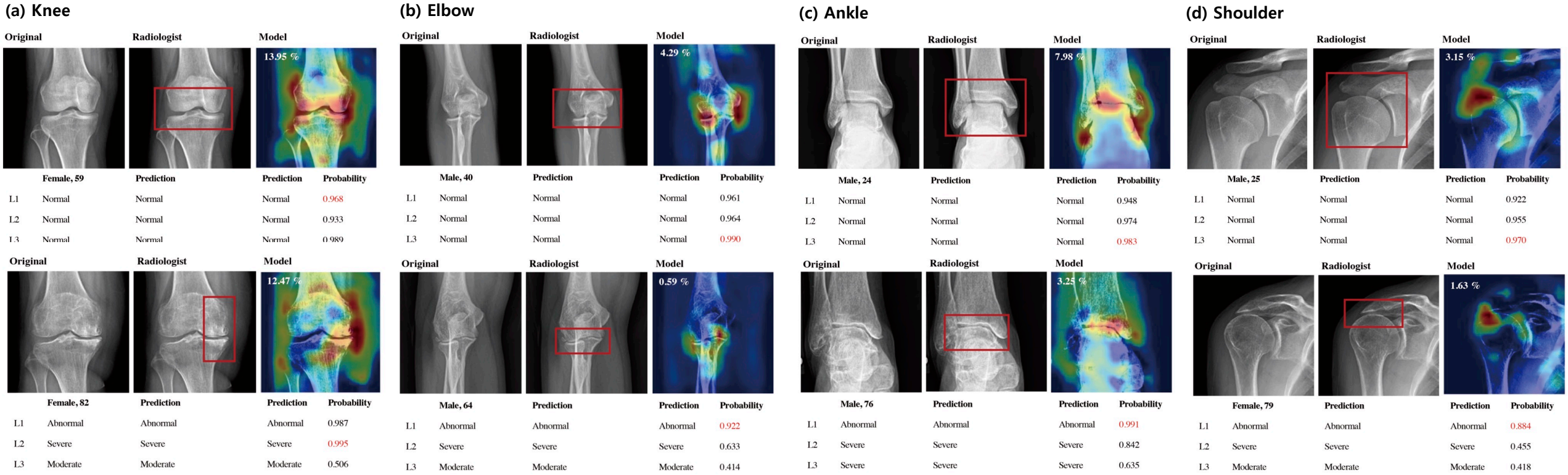
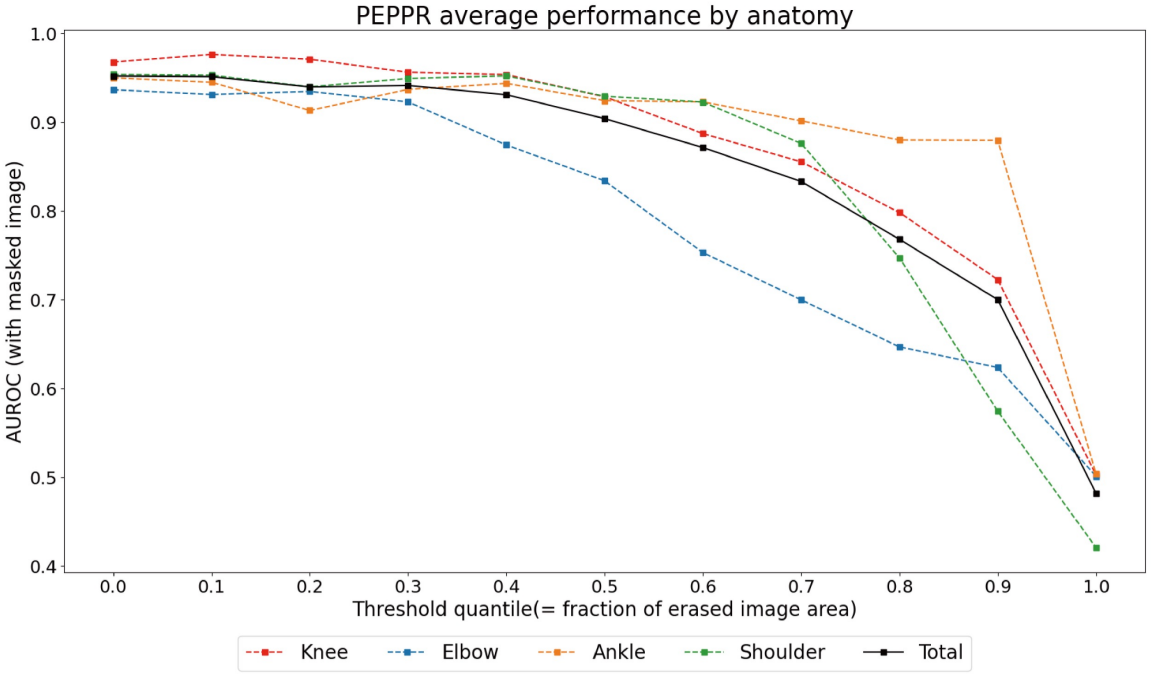


Fig.3 Visual comparison of **gradient-weighted class activation mapping (Grad-CAM) of the model** and **box annotations by the expert radiologist** for normal (first row) and abnormal (second row) cases of the internal dataset.

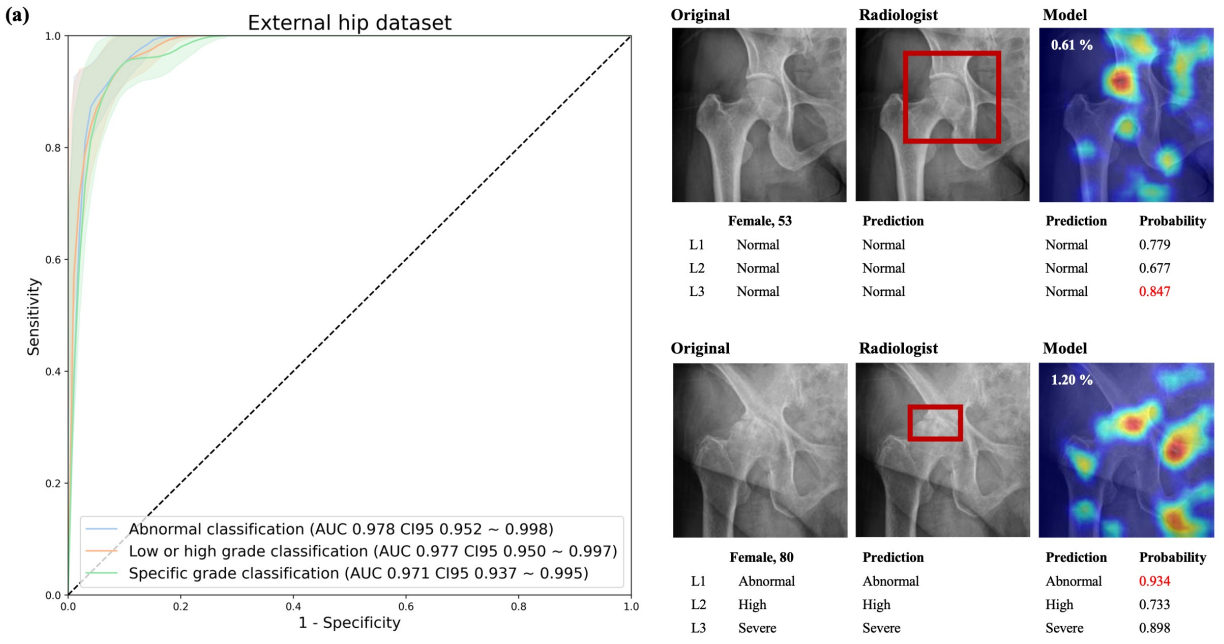


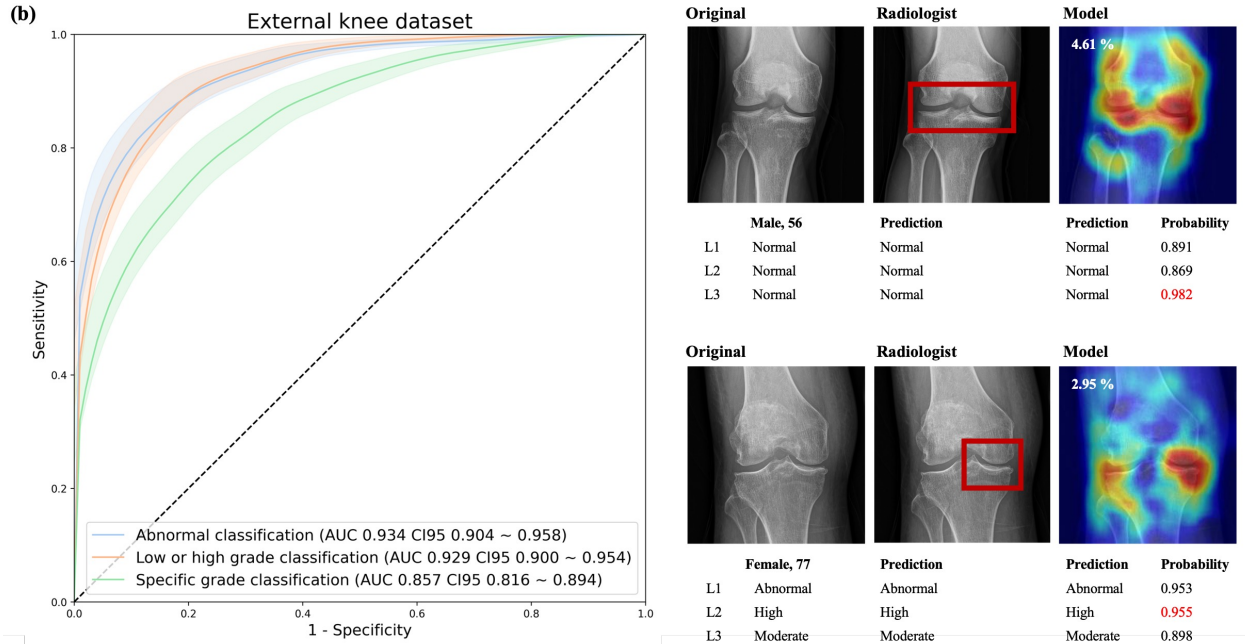
(a) AUROC with masked images

(b) Examples of masked image inputs

Fig.4 Quantitative validation of model explainability using progressive erasing plus progressive restoration (PEPPR).
(a) The average area under the **receiver-operating characteristic curve (AUROC) of L3 classification using masked images.** The x-axis represents the fraction of erased image area (from 0.0 to 1.0 in increments of 0.1).
(b) Masked image inputs are provided to show the progressive erasing of the original radiographs.

External validation





(a) External hip dataset
(Yongin Severance Hospital)

(b) External knee dataset
(Gangnam Severance Hospital)

Incrementally train the Hi-DER model using the external hip dataset (**unseen joint**)

Evaluate the Hi-DER model using the external knee dataset **without additional training** (**seen joint**)

Fig.5 Illustration of the receiver-operating characteristic (ROC) curves for external hip and knee datasets, and comparison of bounding boxes annotated by an expert radiologist and gradient-weighted class activation mapping (Grad-CAM) attention maps.

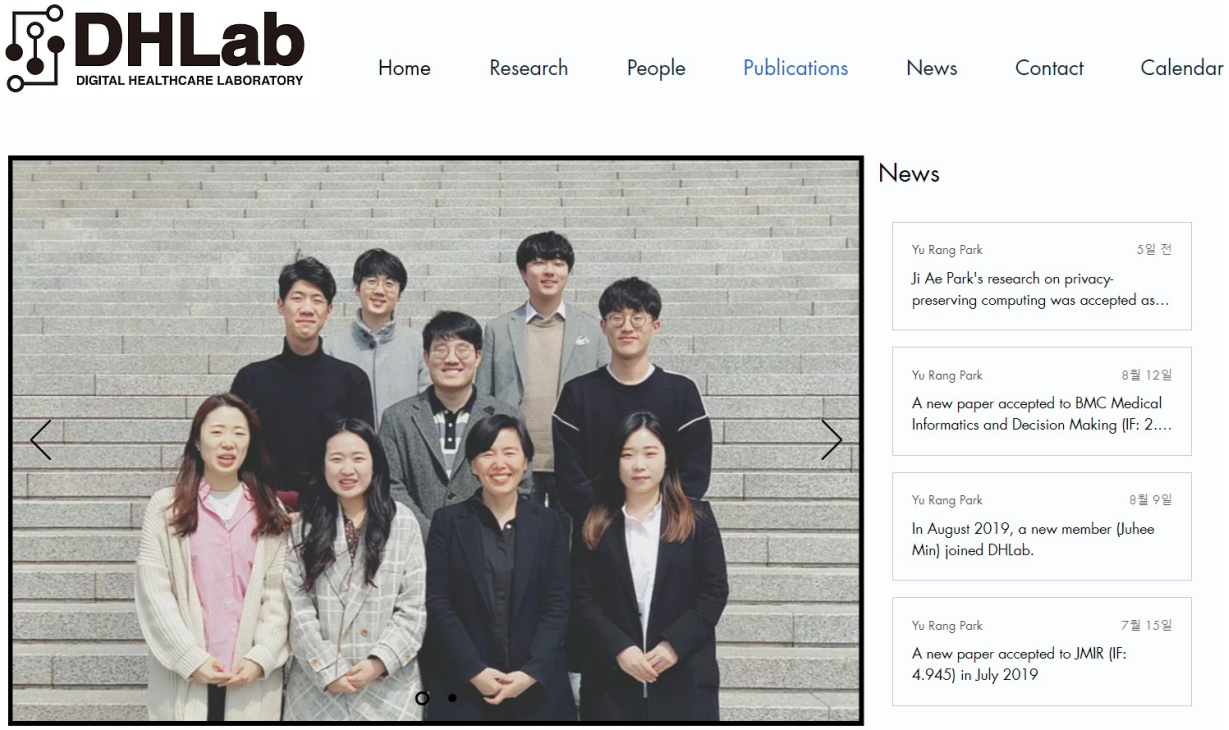
- In this study, we developed **a comprehensive hierarchical arthropathy grade classification model for multiple joints, Hi-DER, with continuously expandable capabilities,** and tested it on internal dataset of knee, elbow, ankle, and shoulder, and external dataset of hip and knee.
- To the best of our knowledge, Hi-DER is the first **hierarchical continual OA classification model** capable of grading the severity of OA in multiple joints based on class hierarchy.
- The results underscore the capability of the continual model trained on representative joints with arthropathy to **generalize effectively across joints with varying anatomical shapes and dimensions.**
- Our model was developed and evaluated on several major joints. In the future, the model could be further trained to classify the arthropathy grades of other joints, including small joints like the fingers or wrists.
- We focused on analyzing primary OA. Further studies on analyzing secondary OA would be beneficial considering its unique characteristics and potential clinical implications.
- We have examined the model's performance in the context of sequential scale classification of KL grade, Hamada grade, or Takakura grade. It would be advantageous to assess the model's effectiveness in non-sequential classifications such as those involving osteogenic, chondrogenic, fibrotic, and other categorizations.

Thank you

이영한
영상의학과 교수

장봉경
의생명시스템정보학과
석사과정

김시원
디지털애널리틱스융합협동과정
석사과정



yurangpark@yuhs.ac

https://www.dhlab.org/