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
© Italian Society of Medical Radiology 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.

Materials and methods This study included a total of 1371 radiographs of knee, elbow, ankle, shoulder, and hip joints from three tertiary hospitals. For model development, 934 radiographs of the knee, elbow, ankle, and shoulder were gathered from Sinchon Severance Hospital between July 1 and December 31, 2022. For external validation, 125 hip radiographs were collected from Yongin Severance Hospital between January 1 and December 31, 2022, and 312 knee cases were gathered from Gangnam Severance Hospital between January 1 and June 31, 2023. The Hierarchical Dynamically Expandable Representation (Hi-DER) model was trained stepwise on four joints using five-fold cross-validation. Arthropathy classification was evaluated at three hierarchical levels: abnormal classification (L1), low-grade or high-grade classification (L2), and specific grade classification (L3). The model's performance was compared with the grading predictions of two other AI models and three radiologists. For model explainability, gradient-weighted class activation mapping (Grad-CAM) and progressive erasing plus progressive restoration (PEPPR) were employed.

Results The model achieved a weighted average AUC of 0.994 (95% CI: 0.985, 0.999) for L1, 0.980 (95% CI: 0.958, 0.996) for L2, and 0.973 (95% CI: 0.943, 0.993) for L3. The model maintained an AUC above 0.800 with 70% of the input regions erased. During external validation on hip joints, the model demonstrated a weighted average AUC of 0.978 (95% CI: 0.952, 0.996) for L1, 0.977 (95% CI: 0.946, 0.996) for L2, and 0.971 (95% CI: 0.934, 0.996) for L3. For external knee data, the model yielded a weighted average AUC of 0.934 (95%: CI 0.904, 0.958), 0.929 (95% CI: 0.900, 0.954), and 0.857 (95% CI: 0.816, 0.894) for L1, L2, and L3, respectively.

Conclusion The Hi-DER may enhance the efficiency of arthropathy diagnosis through accurate classification of arthropathy grades across multiple joints, potentially enabling early treatment.

Keywords Arthropathy · Continual learning · Grading prediction · Radiograph

Bong Kyung Jang and Shiwon Kim have contributed equally to this work.

✉ Young Han Lee
sando@yuhs.ac

✉ Yu Rang Park
yurangpark@yuhs.ac

[1] Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2] Department of Digital Analytics, College of Computing, Yonsei University, Seoul, Republic of Korea

[3] Department of Radiology, Research Institute of Radiological Science, and Center for Clinical Imaging Data Science (CCIDS), Yonsei University College of Medicine, Seoul, Republic of Korea

[4] Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea

[5] School of Medicine, CHA University Gyeonggi-do, Pocheon, Republic of Korea

Published online: 24 March 2025

Introduction

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 common 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 severity is dependent on joint space narrowing, osteophytes, subchondral sclerosis and subchondral cysts, which are quantified using scales like the Kellgren-Lawrence (KL) scale [2]. The Hamada classification categorizes rotator cuff arthropathy based on changes observed in the acromion and the humeral head in shoulder [3], and the Takakura classification in ankle has been used for the stratification in ankle OA [4]. Radiographic assessment of OA severity is critical for clinical decision making, including diagnosis, treatment monitoring, and research [5]. However, radiographic classification of OA severity is a time-consuming task requiring assessments of joint space width, osteophytes, and subchondral sclerosis and is a subjective evaluation, coupled with vaguely defined features at various stages of OA progression, resulting in low inter-observer reliability [6, 7].

In light of these challenges, AI is being used to classify the severity of OA [8, 9] and tools using deep learning to diagnose and assess OA have been developed, showing higher consistency and increased accuracy [10–13]. AI can be utilized to rule out rotator cuff tears [14] and to provide automatic ankle OA grading [15]. Nevertheless, the majority of AI models exhibit limitations in that they 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 [16]—to OA severity grading, enabling simultaneous classification across multiple joints with various morphologies. Specifically, we utilize the Dynamically Expandable Representation (DER) [17] method, which continuously expands feature dimensionality to integrate new visual concepts with previously learned knowledge. This characteristic makes DER well-suited for multiple joint OA severity grading, as the structure of severity is consistent across different joints.

Although some studies have applied continual learning to medical imaging [18], most existing approaches overlook the hierarchical structure inherent in many medical annotations. In particular, the severity of arthritis exhibits a consistent hierarchical structure across multiple joints [19], but previous AI models are trained without considering these relationships among classes. Thus, in this study, we developed and validated a Hierarchical Dynamically Expandable Representation (Hi-DER) model that employs a hierarchical and continual classification approach for enhanced model scalability and preservation of hierarchical information in arthropathy classification.

The objective of this study was to formulate a Hi-DER model to classify the severity of multiple arthritis types. This included the use of four different arthritis image datasets, validation with two external datasets from two external sites, a comparative analysis between the areas of significance identified by the model and clinician evaluation, and the identification of critical radiographic features during the classification process.

Materials and methods

Multiple joint osteoarthritis dataset

This study included radiographs of knee, elbow, ankle, shoulder of adult patients (over 18 years of age) in inpatient and outpatient settings between July 1, 2022 and December 31, 2022: 934 radiographs (274 knee, 209 elbow, 249 ankle, and 202 shoulder) from Sinchon Severance Hospital. Radiological grading of OA was performed by radiologists blinded to clinical information and other imaging results: KL grading for knee and elbow [2], Takakura grading for ankle [4], and Hamada grading for shoulder [3]. All grading was performed by two musculoskeletal imaging fellowship-trained radiologists without clinical information.

For external validation, 125 hip AP radiographs and 312 knee AP radiographs were collected and graded using KL grading [2]. External validation on hip included hip AP radiographs of adult patients gathered from Yongin Severance Hospital in inpatient and outpatient settings between January 1, 2022 and December 31, 2022. For external knee validation, knee AP radiographs of adult patients were collected from Gangnam Severance Hospital in inpatient and outpatient settings between January 1, 2023 and June 31, 2023.

We used a three-level hierarchical labeling strategy based on the annotations of the radiologists: L1 for normal / abnormal (mild, moderate, severe); L2 for normal / low-grade (mild) / high-grade (moderate, severe); and L3 for specific grades (normal, mild, moderate, severe). Detailed information on inclusion/exclusion criteria and annotations are provided in Supplementary Method 1.

Study design

We developed a hierarchical continual arthropathy classification model for multiple joints, employing a study design with internal and external validation processes (Fig. 1a). In internal validation, 80% of the dataset was used for model development and 20% was used for testing. For external hip validation, 80% of the external hip dataset was used for training and 20% was used for testing. For external knee validation, all external knee data were used exclusively for testing. Input consisted of 224 × 224 pixel radiographs in both model development and external validation processes.

For model development, we used five-fold cross-validation, adopting a continual learning approach in a stepwise structure across knee, elbow, ankle, and shoulder dataset from Sinchon Severance Hospital, to prevent catastrophic forgetting. The model's performance was verified at each hierarchical level: abnormal classifications (L1), low-grade or high-grade classifications (L2), and specific grade classifications (L3). For external validation on hip, the model trained on internal datasets from Sinchon Severance Hospital

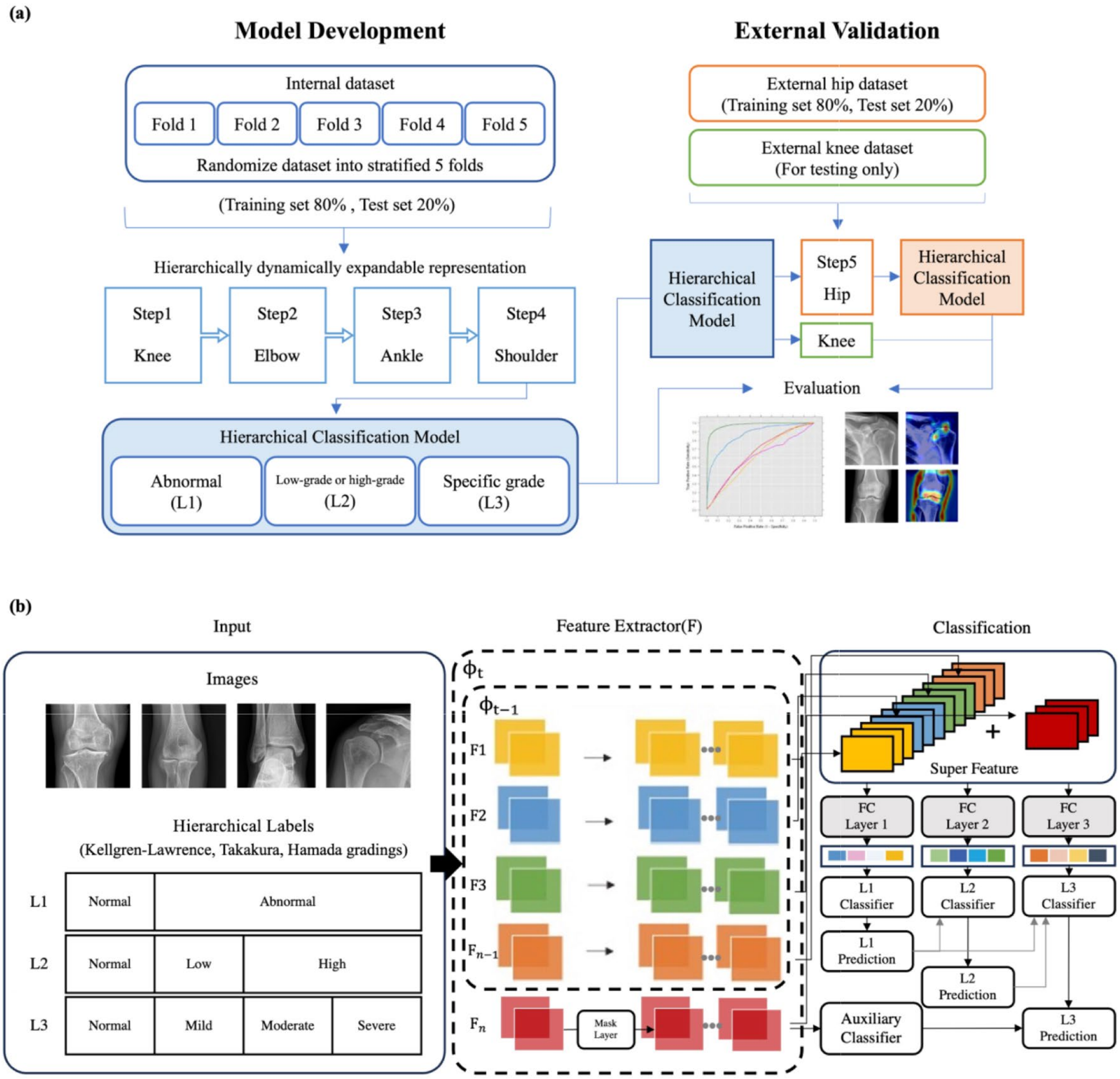


Fig. 1 **a** Illustration of the overall study design. The hierarchical classification model (Hi-DER) was incrementally trained using internal datasets of knee, elbow, ankle, and shoulder osteoarthritis (OA) for model development. The model was additionally trained using hip dataset for external hip validation. For external knee validation, the model was tested on an unseen knee dataset without additional training. **b** Visual explanation of the hierarchical classification model architecture and training process. Input radiographs are hierarchically labeled in three levels based on Kellgren-Lawrence, Takakura, and Hamada gradings. At step t, a new feature extractor F_n is generated and integrated with the previous super-feature extractor Φ_{t-1} to form the t^{th} super-feature extractor Φ_t. Three fully-connected (FC) layers are constructed for classifications at each hierarchical level. The auxiliary classifier discriminates between the old and new concepts

was incrementally trained using the hip OA dataset from Yongin Severance Hospital to evaluate the model's scalability to external data with distinct shapes and findings. For external knee validation, the knee OA dataset from Gangnam Severance Hospital was evaluated without additional training to validate the model's generalizability.

Hierarchical dynamically expandable representation

The Dynamically Expandable Representation (DER) [17] trains the network continually in multiple steps, expanding the feature representations at each incremental step. It preserves the old knowledge from the previous steps while acquiring new information. We developed Hi-DER, a hierarchical continual learning model, which incorporates hierarchical loss at each step, capturing hierarchical information between outcomes. Our method adopts a three-stage training strategy composed of input, feature extraction, and classification. The overall training procedure is shown in Fig. 1b and Supplementary Method 2. A detailed pseudo code can be found in Supplementary Fig. 1.

Comparison with other deep learning algorithms

We conducted a comparative analysis of Hi-DER and other continual and static algorithms, ResNet-50 [20] and DER [17], to validate our model's effectiveness in hierarchical arthropathy classification. ResNet-50 is a static model which is the backbone of our model, and DER is the fundamental continual learning method used for the development of Hi-DER. To effectively compare static and continual methods, we gradually added each joint from single joint classification to the classification of four different joints and observed the capability of each model to continually accommodate the increasing number of classes.

Comparison with grading predictions of radiologists

To evaluate the clinical applicability of our model, we compared its performance with the OA grading results from three radiologists with different levels of experience: L.Y.H., expert, 18 years of experience with musculoskeletal imaging; L.J., senior, 8 years of experience with musculoskeletal imaging; and W.J., medical student, year 4. We used a randomly selected 20% of the multiple joint arthropathy dataset to compare the classification results of our model and the radiologists.

Model explainability

We employed two explainable AI methods, gradient-weighted class activation mapping (Grad-CAM) [21] and progressive erasing plus progressive restoration (PEPPR) [22] to discern key regions in radiographs influencing Hi-DER's decision-making. These methods offer visual and quantitative insights into how different areas of the input radiograph impact on the model's decision, bolstering interpretability.

Performance evaluation

The area under the receiver-operating characteristic curve (AUC) assessed the model's class differentiation ability. The receiver-operating characteristic (ROC) curve was plotted using softmax to visualize the discrimination capabilities. In the evaluation phase, the Hi-DER model yielded three-tiered outcomes using three hierarchy-specific fully-connected layers, each representing the predictions at each hierarchical level (L1, L2, and L3). The outcomes were used to compute the weighted average performance metrics (PPV, NPV, F1 score, sensitivity, specificity) at each level, and scikit-learn version 0.22.2 were used for model evaluation. The metrics were compared to ground truth by expert radiologists, providing a comprehensive analysis of the effectiveness of arthropathy classification at multiple levels.

Results

Patient characteristics

In this study, 1371 radiographs were analyzed, with patients having a median age of 62 years (IQR, 49–71) and 39.46% females; the median BMI was 24.01 kg/m^2 (IQR, 21.85–23.60 kg/m^2). The dataset included Sinchon Severance Hospital (internal validation, n = 934; median age, 62 years; IQR, 48–71 years), Yongin Severance Hospital (external validation, n = 125; median age, 62 years; IQR, 54–70 years) and Gangnam Severance Hospital data (external validation, n = 312; median age, 67 years; IQR, 56–74 years). Internal validation included knee, elbow, ankle, and shoulder radiographs; external validation focused on hip and knee joint radiographs. Arthropathy grading was performed by an expert radiologist, categorized as normal, mild, moderate, or severe (Table 1).

Hierarchical classification performance of arthropathy grading in multiple joints

The performance of Hi-DER on continual arthropathy classification was evaluated using internal test datasets of knee, elbow, ankle, and shoulder joints. In joint-wise comparison, the knee OA classifications had the highest weighted AUCs at all hierarchical levels: 0.999 (95% CI: 0.997, 1.000) for abnormal classifications (L1), 0.985

Table 1 Demographic characteristics of the multiple joint osteoarthritis dataset

Participant characteristics			Internal Validation (Sinchon Severance Hospital)				External Validation (Yongin Severance Hospital)	External Validation (Gangnam Severance Hospital)	Total (n=1371)
			Knee (n=274)	Elbow (n=209)	Ankle (n=249)	Shoulder (n=202)	Hip (n=125)	Knee (n=312)	
Median baseline age, yr (IQR)			66 (58–73)	55 (42–66)	54 (40–66)	65 (53–72)	62 (54–70)	67 (56–74)	62 (49–71)
Median baseline body mass index, *kg/m*^{2} (IQR)			24.37 (22.50–26.50)	24.45 (21.60–26.13)	22.00 (20.21–24.35)	24.49 (22.54–26.44)	24.22 (22.59–26.67)	24.54 (22.43–26.81)	24.01 (21.85–23.60)
	Sex, n (%)								
Female			36 (13.14%)	99 (47.37%)	145 (58.23%)	111 (54.95%)	55 (44.00%)	95 (30.45%)	541 (39.46%)
Male			238 (86.86%)	110 (53.63%)	104 (41.77%)	91 (45.05%)	70 (56.00%)	217 (69.55%)	830 (60.54%)
	Grade, n (%)								
Normal	Normal	Normal	94 (34.31%)	86 (41.15%)	101 (40.56%)	71 (35.15%)	35 (28.00%)	87 (27.88%)	474 (34.57%)
Abnormal	Low	Mild	50 (18.25%)	52 (24.88%)	89 (35.74%)	83 (41.09%)	34 (27.20%)	56 (17.95%)	364 (26.55%)
	High	Moderate	63 (22.99%)	43 (20.57%)	32 (12.85%)	33 (16.34%)	31 (24.80%)	88 (28.21%)	290 (21.15%)
		Severe	67 (24.45%)	28 (13.40%)	27 (10.84%)	15 (7.43%)	25 (20.00%)	81 (25.96%)	243 (17.72%)

(95% CI: 0.966, 0.997) for low-grade and high-grade classifications (L2), and 0.982 (95% CI: 0.959, 0.996) for specific grade classifications (L3). Conversely, the ankle OA classification showed the lowest weighted average AUCs for L1 and L3 (0.990 [95% CI: 0.975, 0.999] for L1; 0.958 [95% CI: 0.919, 0.986] for L3), and the elbow OA classifications were the lowest for L2 (0.977 [95% CI: 0.953, 0.993]). Nevertheless, the results demonstrate the robustness and extendibility of our model in arthropathy classification, with average AUCs above 0.950 for all four differently shaped joints at three hierarchical levels. Detailed results can be found in Fig. 2 and Supplementary Fig. 2.

The Hi-DER showed a weighted average AUC above 0.950 at all hierarchical levels (Fig. 2) with L1 classifications outperforming L2 and L3. The model achieved a weighted average AUC of 0.994 (95% CI: 0.985, 0.999) for L1, 0.980 (95% CI: 0.958, 0.996) for L2, and 0.973 (95% CI: 0.943, 0.993) for L3. The weighted average accuracies were 90.32% for L1, 77.64% for L2, and 67.84% for L3 (Table 2).

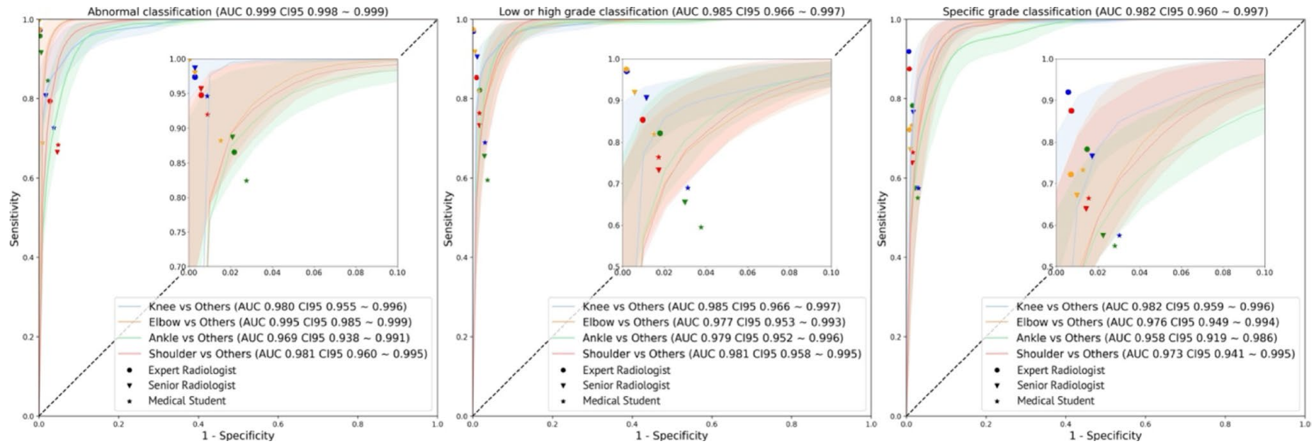


Fig. 2 Visual comparison of the receiver-operating characteristic (ROC) curves and grading results of the radiologists. The three plots represent each hierarchical level of classification. Internal datasets include knee (blue), elbow (yellow), ankle (green), and shoulder (red). 95% confidence intervals (CI95) are illustrated as the boomerang-shaped areas. The classification results of the radiologists (expert radiologist, 18 years in musculoskeletal imaging; senior radiologist, 8 years in musculoskeletal imaging; medical student, year 4) are plotted alongside the ROC curves

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

The performance metrics of the ResNet-50, DER, and Hi-DER using four different dataset combinations are shown in Table 3. The Hi-DER achieved 67.84% accuracy, 0.678 sensitivity, 0.974 specificity, and 0.982 AUC in L3 classification using all anatomical locations for training and evaluation. It outperformed the continual learning method DER by more than 20%p in accuracy, and even surpassed the performance of ResNet-50, a static model trained on the entire dataset at once. While continual learning models like DER tend to offer scalability with limited performance, the Hi-DER overcame this trade-off by surpassing both DER and ResNet-50. Moreover, it retained high performance for anatomically distinct joints like ankle and shoulder, where other models struggled to generalize effectively. When the ankle OA was added, Hi-DER had minimal decrease in L3 performance (accuracy, 0.34%p; sensitivity, 0.004), while ResNet-50 had a significant decrease (accuracy, 7.43%p; sensitivity, 0.074).

When compared to radiologists (Fig. 2), Hi-DER performed similarly with the expert and senior radiologists in L1, and outperformed the medical student in L2 and L3. In the ankle OA classification, our model outperformed the senior radiologist in L2 and L3. For shoulder rotator cuff arthropathy, our model outperformed the senior radiologist in L3. The ROC curves visually illustrate the superiority of the model over the radiologists' performance (Fig. 2).

Explainability of hierarchical dynamically expandable representation

To illustrate the explainability of Hi-DER, Fig. 3 depicts the attention map delineating salient features identified by Hi-DER and the corresponding assessments of significant regions by the radiologist across multiple arthropathy radiographs. Figure 3 indicates that in normal cases, Hi-DER emphasizes expansive regions in the attention map. Controversially in abnormal cases, the emphasis is localized to specific narrow areas that have characteristics features such as joint space narrowing, osteophytes, and subchondral sclerosis. We measured the proportion of the significant regions (attention value > 0.8) for quantitative comparison: knee (normal 13.95%, abnormal 12.47%), elbow (normal 4.29%, abnormal 0.59%) ankle (normal 7.98%, abnormal 3.25%), shoulder (normal 3.15%, abnormal 1.63%) (Fig. 3). Hi-DER highlighted regions consistent with radiologists' markings, indicating it learned relevant features for arthropathy grading in multiple joints.

Using PEPPR (Fig. 4), Hi-DER showed an AUC above 0.800 with only 30% of the original radiographs (threshold 0.7). The ankle had minimal degradation (AUC > 0.900 with 90% masked). All joints except the elbow maintained AUC > 0.800 at threshold of 0.7. The elbow retained

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

AUC > 0.800 with only half of the original radiograph (threshold 0.5).

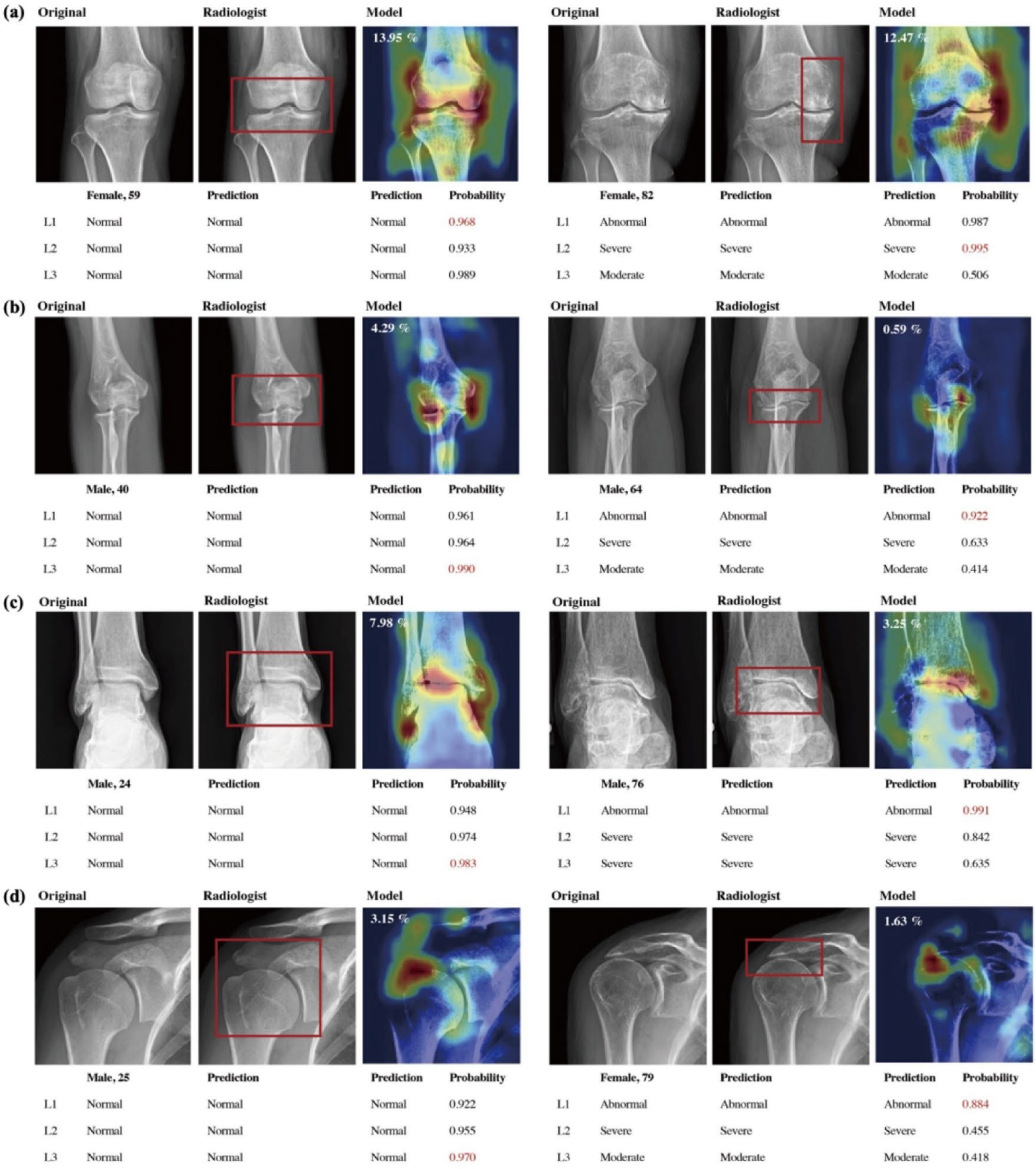
External validation

For hip validation, we incrementally trained Hi-DER using the external hip OA dataset and our model showed consistent performance. As shown in Fig. 5a, the weighted average AUCs surpassed 0.900 for all hierarchical classification levels: 0.978 (95% CI: 0.952, 0.996) for abnormal classifications (L1), 0.977 (95% CI: 0.946, 0.996) for low-grade or high-grade classifications (L2), and 0.971 (95% CI: 0.934, 0.996) for specific grade classifications (L3). Grad-CAM heatmaps presented in Fig. 5a show that our model learned the features related to osteophytes and articular regions. Regions highlighted by the model were consistent with those marked by the radiologists.

For knee validation, we evaluated Hi-DER using external knee OA dataset that was not used during model development. The model exhibited consistent performance in hierarchical arthropathy classification for unseen dataset (Fig. 5b). The weighted average AUCs exceeded 0.900 for L1 and L2, and exceeded 0.850 for L3. Specifically, the model achieved weighted average AUCs of 0.934 (95% CI: 0.904, 0.958) for L1, 0.929 (95% CI: 0.900, 0.954) for L2, and 0.857 (95% CI: 0.816, 0.894) for L3. Grad-CAM attention maps in Fig. 5b demonstrate that even without additional training, our model focused on the features relevant to articular regions and osteophytes in external knee classification. The highlighted regions correspond to those marked by the radiologists, demonstrating the effectiveness of the model.

Discussion

OA can affect joints with various morphologies and pathological changes [23]. The radiographic severity of OA is categorized using various grading systems that consider findings such as joint space width, osteophytes, and subchondral sclerosis. Recent AI-driven radiographic image analysis has shown promise in detecting and grading OA based on KL grading [8, 9, 24, 25], Hamada classification [14], and Takakura classification [15]. Some recent works have adopted semi-supervised methods to effectively learn the representations of radiographs for OA severity grading [26, 27]. However, these methods were trained on a single joint, showing limited extensibility and generalizability. To address this issue, we introduce continual learning [16], a training method that enables continuous and generalized application of the model. Although several studies have applied continual learning to medical imaging [18, 28, 29], these works hardly consider the hierarchical information between the outcomes. Traditional well-performing



(a)

	Female, 59	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.968
L2	Normal	Normal	Normal	0.933
L3	Normal	Normal	Normal	0.989

	Female, 82	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.987
L2	Severe	Severe	Severe	0.995
L3	Moderate	Moderate	Moderate	0.506

(b)

	Male, 40	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.961
L2	Normal	Normal	Normal	0.964
L3	Normal	Normal	Normal	0.990

	Male, 64	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.922
L2	Severe	Severe	Severe	0.633
L3	Moderate	Moderate	Moderate	0.414

(c)

	Male, 24	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.948
L2	Normal	Normal	Normal	0.974
L3	Normal	Normal	Normal	0.983

	Male, 76	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.991
L2	Severe	Severe	Severe	0.842
L3	Severe	Severe	Severe	0.635

(d)

	Male, 25	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.922
L2	Normal	Normal	Normal	0.955
L3	Normal	Normal	Normal	0.970

	Female, 79	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.884
L2	Severe	Severe	Severe	0.455
L3	Moderate	Moderate	Moderate	0.418

Fig. 3 Visual comparison of gradient-weighted class activation mapping (Grad-CAM) heatmaps of the model and box annotations by the expert radiologist for normal and abnormal cases of the internal test datasets **a** knee **b** elbow **c** ankle **d** shoulder. The proportion of the regions with attention value greater than 0.8 are shown on the upper left of the attention maps. Annotations: (left) demographic information and true label; (middle) predictions of the radiologist; (right) predictions and probabilities of the model

continual learning methods [17] also struggle with radiographs due to their inability to capture the hierarchical structure of disease severity [19]. We propose Hi-DER, a hierarchical continual learning method, to overcome this limitation. 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.

With the growth of AI, especially deep learning, transfer learning and pretrained models are being introduced to radiologic imaging. Both continual and transfer learning involve building on pretrained weights and further learning of the model [30]. However, it is essential to continually adjust the updated weights at different stages of learning as new classes emerge. This study successfully applies continual learning to different anatomical joint radiographs. Our model is valuable

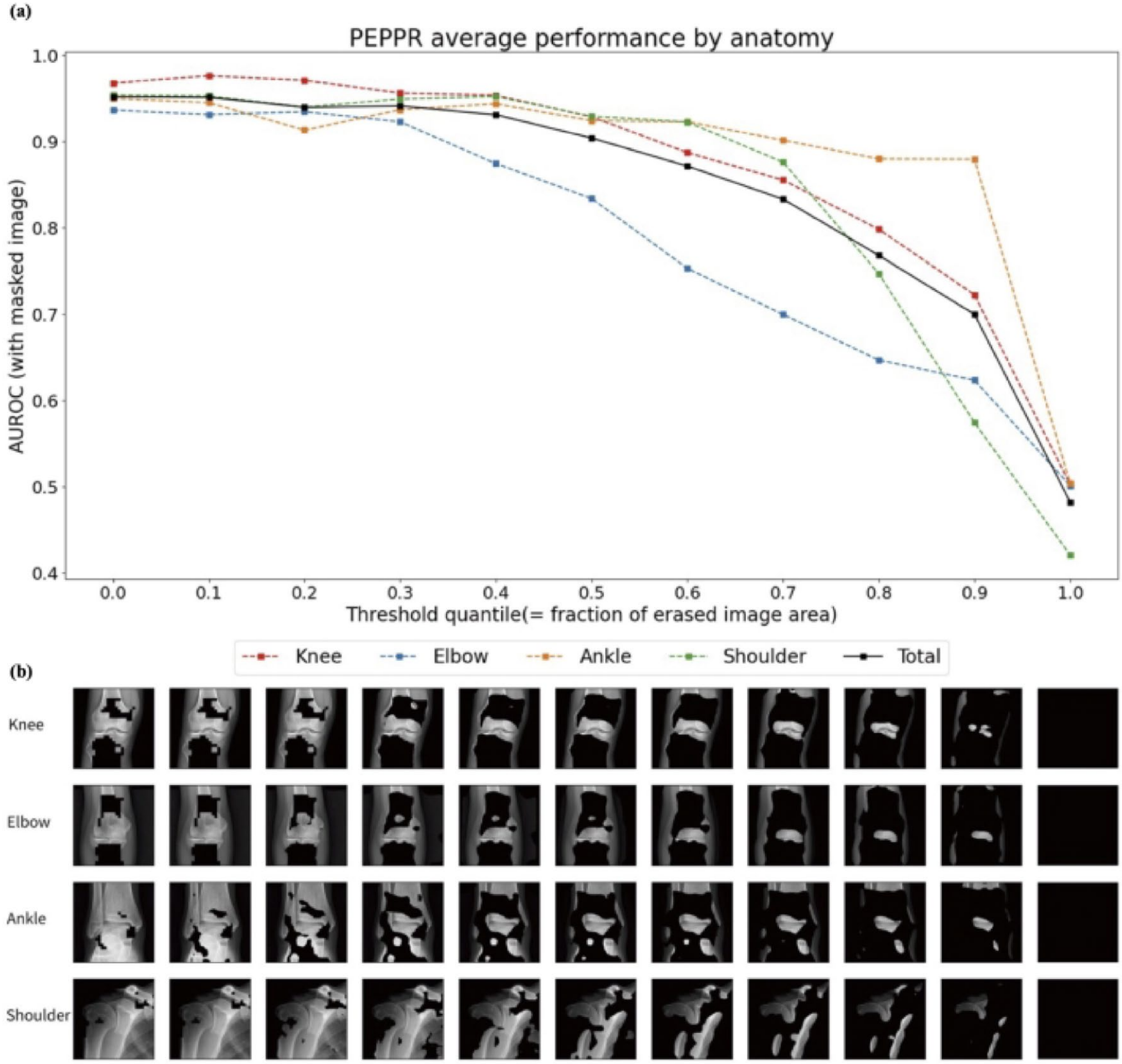
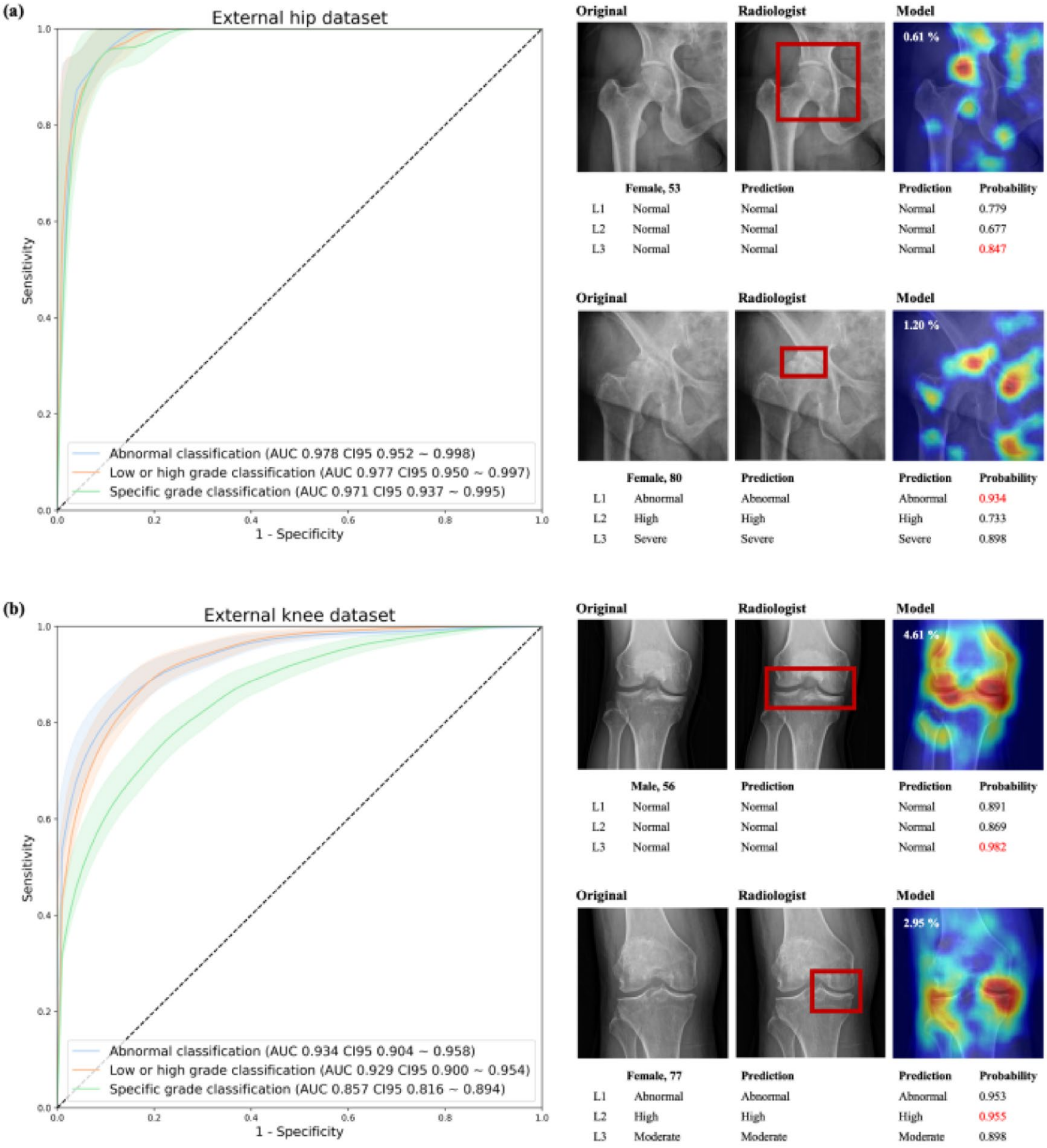


Fig. 4 Quantitative validation of model explainability using progressive erasing plus progressive restoration (PEPPR). **a** Quantitative explanation on which part of the input image contributes the most to the decision made by the model. The average area under the receiver-operating characteristic curve (AUC) of specific grade classifications (L3) was measured using masked input images. Thresholds from 0.0 to 1.0 in increments of 0.1 were used for image masking. **b** Masked images are provided to show the progressive erasing of the original radiographs

for the simultaneous analysis of radiographs of multiple joints. The implementation of continual learning improves the adaptability of the model, which addresses challenges in medical imaging. Existing medical AI algorithms face problems because they are trained on fixed data distributions for specific medical outcome, making them inaccurate and unsustainable when applied in real-world medical settings [31]. Continual learning provides automatic, frequent updates that adapt to local changes in clinical data [18]. It enables practical AI to continually learn and adapt, thereby improving arthropathy grading, supporting clinical decisions, predicting arthropathy prognosis, and enabling hierarchical classification.

In this study, we developed a comprehensive hierarchical arthropathy grade classification model in 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. The experiments demonstrate that Hi-DER can be extended to classify multiple joints with different morphologies in hierarchical arthropathy grade classifications (L1, L2, L3). The model surpassed the weighted average AUCs of 0.950 in all anatomical sites and hierarchical levels for internal joint datasets. Moreover, the performance of our model in knee and ankle OA classifications is equal to or exceeds the values reported in previous studies for single joint OA severity grading [15, 24, 25].



	Female, 53	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.779
L2	Normal	Normal	Normal	0.677
L3	Normal	Normal	Normal	0.847

	Female, 80	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.934
L2	High	High	High	0.733
L3	Severe	Severe	Severe	0.898

	Male, 56	Prediction	Prediction	Probability
L1	Normal	Normal	Normal	0.891
L2	Normal	Normal	Normal	0.869
L3	Normal	Normal	Normal	0.982

	Female, 77	Prediction	Prediction	Probability
L1	Abnormal	Abnormal	Abnormal	0.953
L2	High	High	High	0.955
L3	Moderate	Moderate	Moderate	0.898

Fig. 5 **a** Illustration of the receiver-operating characteristic (ROC) curves and the area under the receiver-operating characteristic curve (AUC) at each hierarchical level for external hip validation (left). Comparison of attention boxes annotated by an expert radiologist and gradient-weighted class activation mapping (Grad-CAM) attention maps for the normal and abnormal cases. **b** The ROC curves at each hierarchical level for external knee validation (left). Comparison of Grad-CAM attention maps of the model and box annotations by the expert radiologist (right). Attention maps are overlaid on the original radiographs

Tiulpin et al. [25] proposed a knee OA grading algorithm (5 grades, KL grade) with a multiclass accuracy of 66.71% and an AUC of 0.930. Similarly, Kim et al. [15] proposed an ankle OA grading model (3 grades, Takakura grade) that yielded the average accuracies of 78.1% and 79.2%. Our model achieved an accuracy of 73.06% and an AUC of 0.982 (95% CI: 0.959, 0.996) for knee OA (4 grades, KL grade). For ankle OA (3 grades, Takakura grade), the accuracy and AUC were 77.13% and 0.979, respectively. In contrast to static models in the previous studies [13, 32], our work demonstrates consistent performance across multiple joints, reinforcing the potential of deep learning models in arthropathy. By applying Hi-DER to arthropathy, our approach successfully handles of arthropathy in multiple joints with different shapes, improving the efficiency of severity assessment for both upper and lower extremities.

We performed step-by-step evaluations of L1, L2, and L3 classifications to assess the hierarchical arthropathy classification performance of Hi-DER in differently shaped joints. 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. This is supported by evidence indicating that the pathologic findings of arthropathy, such as joint space narrowing and osteophyte formation, remain consistent across joints, irrespective of their size or shape. These findings further reinforce the hypothesis that if we train the detection of arthropathy using one joint, then it can also be applied to arthropathy in other joints with different shapes.

However, certain features of OA, such as osteochondral lesions or areas of bone necrosis, may challenge the accuracy of detection models, potentially complicating the model's ability to distinguish OA from other joint disorders by creating overlapping imaging characteristics or variations in pathological expression. Moreover, the model accuracy can be influenced by the underrepresentation of rare and mild cases. For example, ankle valgus OA is significantly less common than varus OA, leading to limited training data and reduced diagnostic accuracy for such atypical cases. Additionally, early OA with subtle minor osteophytes often poses challenges for detection algorithms. These limitations highlight the need for more diverse and balanced datasets to enhance model performance and generalizability. The example radiographs of ankle valgus OA and early OA of the elbow can be found in the "Model Robustness: Addressing Underrepresented Features of Osteoarthritis" section of our GitHub repository: https://github.com/DigitalHealthcareLab/24HiDER.

Our approach accounts for these challenges by focusing on generalized pathological findings across joint types. The generalizability of this approach highlights a key strength of our method, emphasizing its potential scalability for diagnosing arthropathy across a broad spectrum of joints, including those with unique anatomical characteristics. This adaptability positions the model as a promising tool for widespread clinical applications, enabling consistent and reliable detection of arthropathy regardless of joint type. Furthermore, considering the continuation of arthropathy detection, classification of low-grades or high-grades arthropathy, and detailed specific grade classifications, it can be extended to radiographs with sub-grading [33, 34]. This can overcome the challenges of training data for sub-classified diseases associated with the application of AI in medical imaging. This model closely emulates the learning trajectory observed in radiologists. When entering the field of radiology, beginner radiologists initially learn to differentiate between normal and abnormal radiographs. As their expertise develops over time, they gradually become capable of making specific diagnoses and performing specific grading. Our model and the radiologists showed similar performance, both tending to perform better in L1 than L2 and L3. Considering that many pathologies and diseases are classified into finer subcategories, it is believed that this model will be more efficient and robust for further refining disease classification.

We compared the OA classification performance of Hi-DER and the radiologists. In L1, the performance of our model was similar with that of the expert and senior radiologist in most anatomical locations. It presented higher performance than the medical student in L2 and L3. In the ankle OA classifications, our model outperformed the senior radiologist in L2 and L3 in terms of performance. In L3 of the shoulder OA classifications, our model showed higher performance than the senior radiologist. For a visual comparison with the ROC curves of the model, we plotted the radiologists' performance based on their classification performance.

We compared the performance of Hi-DER with ResNet-50 and DER in multiple joint arthropathy. Our model outperforms DER, and even ResNet-50, a static model that uses all available data for training. Continual learning models typically experience greater performance degradation that static models as the number of classes increases, but Hi-DER demonstrates less degradation than ResNet-50. DER's performance drops significantly with each new joint added, highlighting the strong reliance of both ResNet-50 and DER on anatomical shapes. In contrast, Hi-DER remains robust and generalizable across different morphologies. Our model is well suited for the continual learning and classification of multiple joint arthropathy, regardless of their anatomical shapes.

We applied Grad-CAM and PEPPR to demonstrate our model's explainability. Grad-CAM results reveal Hi-DER's focuses on features relevant to arthropathy findings, not just image correlations. Notably, abnormal radiographs' high attention value (>0.8) was confined to smaller areas, contrasting with normal cases covering larger joint regions. This aligns with radiological arthropathy severity gradings, involving specifics like osteophytes, joint space narrowing, and subchondral sclerosis. Grad-CAM show that the model highlights areas consistent with radiologists' judgments, demonstrating learning of relevant features. PEPPR results quantitatively support Hi-DER's learning of arthropathy-related features, achieving AUC >0.800 using only 30% of the input radiographs, with pixels below 0.7 masked.

Finally, we assessed the robustness and expandability of Hi-DER through an external validation process using external hip and knee OA radiographs. For external hip validation, our model still showed consistent performance with additional hip classes from the external hip dataset. The weighted average AUC for L3 of the hip OA classification was 0.971 (95% CI: 0.934, 0.996). For external knee validation, although the external data was exclusively used for testing and not for training, the model retained the weighted average AUCs higher than 0.900 for L1 and L2, and about 0.860 for L3. Grad-CAM results for both the hip and knee show that the areas highlighted by the model are consistent with those marked by the radiologists. This suggests that our model learned the features relevant to OA grading.

Despite the promising results, this study has some limitations. First, our model was developed and evaluated on several major joints. In the future, a continual learning model could be designed to extend the classification of arthropathy data to other joints, including small joints such as the fingers or wrists. While small joints have different anatomical dimensions compared to larger joints, the OA detection model does not rely on joint's specific anatomy but instead focuses on the universal features of arthropathy like joint space narrowing and osteophyte formation. We expect our model to demonstrate robust performance even in smaller joints which have similar pathologic findings. Second, we focused exclusively on analyzing primary OA in this study. The case of secondary OA like inflammatory arthritis, rheumatoid arthritis, post-traumatic arthritis, and post-operative conditions were excluded to maintain a more homogeneous study population and ensure that the findings were specific to primary OA. Further studies on analyzing secondary OA would be beneficial considering its unique characteristics and potential clinical implications. Lastly, we have examined the model performance in the context of sequential scale classification of KL scale, Hamada classification, or Takakura classification. It would be advantageous to assess the model's effectiveness and applicability in non-sequential classifications such as those involving osteogenic, chondrogenic, fibrotic, and other categorizations.

This is the first study to evaluate the performance of a continual learning model on classifying arthropathy grades in multiple joints. Moreover, The algorithm has significant potential for real-world application in the longitudinal evaluation of OA. By tracking changes in radiologic features such as joint space narrowing, osteophyte formation, and subchondral bone changes over time, the model can facilitate the monitoring of disease progression. This is particularly useful for personalized treatment planning, as it allows data-driven adjustments to therapeutic strategies.

Conclusion

In this study, we developed Hi-DER, a continual learning method with a hierarchical structure, and applied it to the training of a hierarchical classification model for grading arthropathy across multiple joints. The model continuously classified arthropathy grades of the knee, elbow, ankle, and shoulder joints, showing classification performance comparable to that of the radiologists. Our algorithm provides a diagnostic process that is efficient and applicable to multiple joint arthritis classifications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11547-025-01974-4.

Author contributions Conceptualization: BKJ, YRP; Methodology: BKJ, SK; Formal analysis and investigation: BKJ, SK; Software: BKJ, SK, JH; Writing—original draft: BKJ, SK, YHL; Writing—review & editing: BKJ, SK, YHL, YRP; Funding acquisition: YRP; Resources and data curation: HWC, HSL, JP, JW, YHL; Supervision: JYY, YRP; Project administration: YHL, YRP.

Funding This research was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (No. P0023675, HRD Program for Industrial Innovation).

Data availability The dataset analyzed in this study was used anonymously and is not publicly available due to privacy concerns but is available as anonymized data from the corresponding author upon reasonable request.

Declarations

Conflict of interests The authors have no relevant financial or non-financial interests to disclose.

Ethical approval This study involving three tertiary centers (Sinchon Severance Hospital, Yongin Severance Hospital, and Gangnam Severance Hospital) was approved by the Institutional Review Board (IRB) of Yonsei University's Health System (4–2023-0614). The need for informed consent was waived because it used anonymized data. All methods were performed in accordance with relevant guidelines and regulations. This study was conducted in compliance with the principles of the Declaration of Helsinki. The authors had complete control of the data and information submitted for publication.

References

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. 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
3. 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
4. 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
5. Croft P (2005) An introduction to the atlas of standard radiographs of arthritis. Rheumatology. https://doi.org/10.1093/rheumatology/kei051
6. 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 23(12):3532–3539. https://doi.org/10.1007/s00167-014-3205-0

7. 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. https://doi.org/10.1016/j.joca.2014.05.007
8. 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. https://doi.org/10.1148/ryai.2020190065
9. 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 296(3):584–593. https://doi.org/10.1148/radiol.2020192091
10. 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. https://doi.org/10.1007/s00256-020-03433-9
11. von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, Posadzy M, Foreman SC, Nevitt MC, Link TM, Pedoia V (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
12. 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. https://doi.org/10.1109/JBHI.2021.3102090
13. 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
14. Iio R, Ueda D, Matsumoto T, Manaka T, Nakazawa K, Ito Y, Hirakawa Y, Yamamoto A, Shiba M, Nakamura H (2023) Deep learning-based screening tool for rotator cuff tears on shoulder radiography. J Orthop Sci. https://doi.org/10.1016/j.jos.2023.05.004
15. Kim H, Choi J, Jang C-Y, Lee JW, Kim S, Han SH (2019) Automatic grading of ankle osteoarthritis based on takakura staging system: a deep learning- based approach. Foot Ankle Orthop. https://doi.org/10.1177/2473011419s00246
16. 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
17. 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
18. 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
19. 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
20. 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
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Ieee Conference on Computer Vision pp 618–626. https://doi.org/10.1109/Iccv.2017.74
22. Engelmann J, Storkey AJ, & Bernabeu MO (2021) Global Explainability in Aligned Image Modalities. arxiv:2112.09591
23. Loeser RF, Goldring SR, Scanzello CR, Goldring MB (2012) Osteoarthritis: a disease of the joint as an organ. Arthritis Rheum-Us 64(6):1697–1707. https://doi.org/10.1002/art.34453
24. Chen PJ, Gao LL, Shi XS, Allen K, Yang L (2019) Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Comput Med Imag Grap 75:84–92. https://doi.org/10.1016/j.compmedimag.2019.06.002
25. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 8(1):1727. https://doi.org/10.1038/s41598-018-20132-7
26. Nguyen HH, Saarakkala S, Blaschko MB, Tiulpin A (2020) Semixup: in- and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs. IEEE Trans Med Imaging 39(12):4346–4356. https://doi.org/10.1109/TMI.2020.3017007
27. Raisuddin A, Nguyen HH, Tiulpin A (2022) Deep semi-supervised active learning for knee osteoarthritis severity grading. I S Biomed Imag. https://doi.org/10.1109/Isbi52829.2022.9761668
28. Ravishankar H, Venkataramani R, Anamandra S, Sudhakar P, Annangi P (2019) Feature transformers: privacy preserving lifelong learners for medical imaging. Lect Notes Comput Sc 11767:347–355. https://doi.org/10.1007/978-3-030-32251-9_38
29. Yang Y, Cui ZY, Xu JJ, Zhong CH, Wang RX, Zheng WS (2021) Continual learning with bayesian model based on a fixed pretrained feature extractor. medical image computing and computer assisted intervention - Miccai 2021. Pt V 12905:397–406. https://doi.org/10.1007/978-3-030-87240-3_38
30. Morid MA, Borjali A, Del Fiol G (2021) A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 128:104115. https://doi.org/10.1016/j.compbiomed.2020.104115
31. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P (2023) Foundation models for generalist medical artificial intelligence. Nature 616(7956):259–265. https://doi.org/10.1038/s41586-023-05881-4
32. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Hum YC, Tee YK, Dhanalakshmi S (2021) Emergence of deep learning in knee osteoarthritis diagnosis. Comput Intel Neurosc. https://doi.org/10.1155/2021/4931437
33. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen H (2021) Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing 437:186–194. https://doi.org/10.1016/j.neucom.2020.03.127
34. Chen H et al. (2019) deep hierarchical multi-label classification of chest X-ray Images. In: International conference on medical imaging with deep learning PMLR

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.