Arthritis Research & Therapy

RESEARCH

Open Access

Automatic knee osteoarthritis severity grading based on X-ray images using a hierarchical classification method

Jian Pan[1†], Yuangang Wu[2†], Zhenchao Tang[3,4,5†], Kaibo Sun[2], Mingyang Li[2], Jiayu Sun[6], Jiangang Liu[1,3,4*], Jie Tian[3,4,5*] and Bin Shen[2*]

Abstract

Background This study aims to develop a hierarchical classification method to automatically assess the severity of knee osteoarthritis (KOA).

Methods This retrospective study recruited 4074 patients. Clinical diagnostic indicators and clinical diagnostic processes were applied to develop a hierarchical classification method that involved four sub-task classifications. These four sub-task classifications were the classification of Kellgren-Lawrence (KL) grade 0–2 and KL grade 3–4, KL grade 3 and KL grade 4, KL grade 0 and KL grade 1–2, and KL grade 1 and KL grade 2, respectively. To extract the features of clinical diagnostic indicators, four U-Net models were first used to segment the total joint space (TJS), the lateral joint space (LJS), the medial joint space (MJS), and osteophytes, respectively. Based on the segmentation result of TJS, the region of knee subchondral bone was generated. Then, geometric features were extracted based on segmentation results of the LJS, MJS, TJS, and osteophytes, while radiomic features were extracted from the knee subchondral bone. Finally, the geometric features, radiomic features, and combination of geometric features and radiomic features were used to construct the geometric model, radiomic model, and combined model in KL grading, respectively. A strict decision strategy was used to evaluate the performance of the hierarchical classification method in all X-ray images of testing cohort.

Results The U-Net models achieved relatively satisfying performances in the segmentation of the TJS, the LJS, the MJS, and the osteophytes with the dice similarity coefficient of 0.88, 0.86, 0.88, and 0.64 respectively. The combined models achieved the best performance in KL grading. The accuracy of combined models was 98.50%, 81.65%, 82.07%, and 74.10% in the classification of KL grade 0–2 and KL grade 3–4, KL grade 3 and KL grade 4, KL grade 0 and KL

†Jian Pan, Yuangang Wu and Zhenchao Tang share the role of first author.

*Correspondence:
Jiangang Liu
jgliu@buaa.edu.cn
Jie Tian
tianj@buaa.edu.cn
Bin Shen
shenbin_1971@163.com

Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

grade 1–2, and KL grade 1 and KL grade 2, respectively. For all X-ray images of the testing cohort, the accuracy of the hierarchical classification method was 65.98%.

Conclusion The hierarchical classification method developed in the current study is a feasible approach to assess the severity of KOA.

Keywords Knee osteoarthritis, U-Net, Machine learning, X-ray image

Background

Knee osteoarthritis (KOA) is a common degenerative disorder [1, 2] characterized by the loss of cartilage [3, 4], formation of osteophytes [3, 4], and alterations in the subchondral bone [4, 5]. The severity of KOA is often assessed using the Kellgren–Lawrence (KL) grading system [6, 7]. Currently, visual interpretation of X-ray images is the common approach in clinical practice for determining the KL grade in KOA diagnosis [8, 9]. However, this method relies on the clinician's experience, which is subjective [8], time-consuming [10], and inconsistent among different clinicians [10, 11]. Therefore, the use of automatic KL grading based on X-ray images can provide an objective and reproducible diagnosis, while also improving diagnostic efficiency.

Previous studies have shown that deep learning is a promising approach for automatic KL grading using X-ray images. Chen et al. [12] conducted a groundbreaking study that combined knee joint detection and classification to achieve automatic KL grading. In this study, the VGG-19 model with the proposed ordinal loss achieved the highest classification accuracy of 70.4%. In more recent studies, Zhang et al. [13] and Wang et al. [14] employed a similar method and achieved accuracies of 74.81% and 78%, respectively. These deep learning models utilized convolutional layers to extract complex and abstract features for diagnosing KL grades. However, the process of feature extraction can be influenced by the image background, leading to the omission of essential features. In the KL grading system, subchondral bone alterations, osteophyte formation, and joint space narrowing are crucial factors for assessing the severity of KOA [15] and serve as clinical diagnostic indicators. By incorporating these indicators into the construction of an automatic KL grading model, the classification accuracy of KL grades can be improved.

Given that joint space narrowing is a crucial indicator in KL grading, a pioneering study [16] focused on extracting the joint space width, which reflects the degree of joint space narrowing, to classify KL grades. However, this study neglected to consider the features of osteophytes and the subchondral bone, which also play significant roles in assessing the severity of KOA. The comprehensive inclusion of features extracted from the joint space, osteophytes, and subchondral bone can effectively reflect differences in KL grades and aid in their distinction.

Radiomics, a technique extensively employed in medical imaging studies [17–20], has shown promise for the diagnosis of KOA. Pioneering studies have demonstrated that radiomic features can be utilized as a potential method for diagnosing KOA. For example, Hirvasniemi et al. [21] constructed an elastic net model using radiomic features extracted from the tibial bone to classify knees with and without osteoarthritis, achieving an area under the receiver operating characteristic (ROC) curve (AUC) of 0.80. Similarly, Anifah et al. [22] extracted four types of features (contrast, correlation, energy, and homogeneity) and employed them to construct a self-organizing map, resulting in high classification accuracy for diagnosing KL grades 0 and 4. However, this study did not explore other radiomic features and had a limited number of radiographic images, necessitating the validation of radiomics methods for automatic KL grading.

The objective of this study was to develop a hierarchical classification method for automatically diagnosing the severity of KOA. Geometric and radiomic features were initially extracted through the segmentation results of U-Net models. Subsequently, geometric, radiomic, and combined models were constructed to classify the KL grades. We hypothesized that the hierarchical classification method would be a feasible approach for distinguishing between different KL grades.

Methods

Overview

As depicted in Fig. 1, the overall procedure for automatically grading KOA severity based on X-ray images consists of three main components: segmentation, feature extraction, and classification using a hierarchical classification method.

Participants

The present study adhered to the principles outlined in the Declaration of Helsinki and received approval from the West China Hospital, Sichuan University. Informed consent forms were signed by all participants included in the study.

A total of 5317 knee joint X-ray images from 4074 patients (mean ± standard deviation (SD) age = 52.08 ± 14.35 years; 1268 males) between July 2009

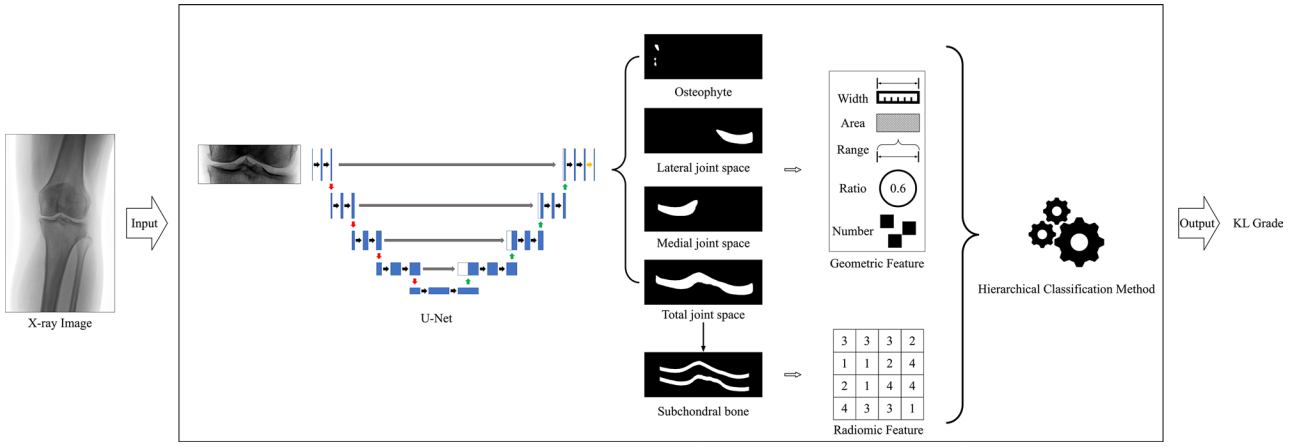

Fig. 1 Overview of the automatic knee osteoarthritis severity grading based on X-ray images. The total joint space, the lateral joint space, the medial joint space, and osteophyte were segmented using U-Net model, respectively. The mask of subchondral bone was generated based on the segmentation result of total joint space. A hierarchical classification method was employed to assessment KOA severity using geometric and radiomic features. KOA: knee osteoarthritis. KL: Kellgren-Lawrence

Table 1 Demographic characteristics of training and testing cohorts

Characteristics	Training cohort (*n* = 3259)	Testing cohort (*n* = 815)	Total (*n* = 4074)
Age (year), mean ± SD	52.10 ± 14.33	52.02 ± 14.41	52.08 ± 14.35
Gender (male/female)	1009 / 2250	259 / 556	1268 / 2806
Number of X-ray images	4247	1070	5317
KL grades			
0	274	66	340
1	1214	275	1489
2	1686	451	2137
3	714	172	886
4	359	106	465

KL: Kellgren–Lawrence

SD: Standard deviation

and April 2021 were analyzed in the current study. Among the 4074 patients, 2848 had only one X-ray image, while the remaining patients had multiple X-ray images. The training cohort comprised 80% of the patients, randomly selected (3259 patients, 4247 radiographs), while the remaining patients formed the testing cohort (815 patients, 1070 radiographs). All X-ray images from a patient were included in only one cohort. Table 1 provides an overview of the demographic characteristics of the training and testing cohorts.

All patients included in the study met the following inclusion criteria: (1) availability of clinical information and (2) availability of images. Patients who had (1) missing images, (2) low image quality, (3) incomplete clinical information, or (4) a history of knee joint trauma or tumor were excluded.

Image acquisition

Each patient in the study underwent radiography using a digital radiography system. The imaging parameters were standardized, with a fixed tube voltage of 50 kV, an exposure condition of 0.8 mAs, a source-film distance of 60 cm, and an exposure index ranging from 1400 to 1800. Given that the Rosenberg view is more sensitive than Antero-posterior view in identifying the joint space narrowing which played an important role in KL grading [23], the field of the knee radiograph encompassed the entire knee joint, leg, and thigh was scanned using Rosenberg view.

Image reading and manual segmentation

Lateral joint space (LJS), medial joint space (MJS), total joint space (TJS), and osteophyte were manually segmented using ITK-SNAP software [24], respectively. The manual segmentation was performed by three senior orthopedic doctors who have more than ten years' experience in X-ray image reading and clinical practice. If there was inconsistency in the manual segmentation, a final decision was reached through discussion among three senior doctors. In addition, these three doctors made diagnosis of KL grade of KOA based on the interpretation of the X-ray images. To ensure accuracy, three

doctors independently reviewed all the acquired X-ray images. The KL grade, which assesses joint space and osteophytes in KOA, was determined based on the X-ray images [25]. In cases where there was inconsistency in the KL grade assigned by the doctors, a final decision was reached through discussion among them. These three senior orthopedic doctors were blinded to the patients' clinical information during manual segmentation and KL grade diagnosis.

Segmentation based on U-Net model

To reduce the time-consuming and labor-intensive process of precise segmentation for doctors, four U-Net models (as depicted in Fig. 2) were used to segment LJS, MJS, TJS, and osteophyte, respectively. The U-Net [26] architecture consists of an encoder and a decoder. The encoder includes four downsample blocks. A downsample block consists of two convolutional layers, two batch normalization layers, a rectified linear unit (ReLU), and a max-pooling layer. The decoder, on the other hand, comprises four upsample blocks. A upsample block consists of a transpose convolutional layer, two convolutional layers, two batch normalization layers, and a ReLU. This structure is constructed to generate the final segmentation output. Two dropout2d layers were added between the encoder and decoder to prevent overfitting.

To normalize the X-ray images, the following equation was applied:

$$y = \frac{x - min}{max - min}$$

where x and y represent the original and normalized intensity of each pixel, respectively. Further, min and max denote the minimum and maximum value of the intensity of all pixels in the corresponding original X-ray image, respectively. Due to the large size of the X-ray image (as depicted in Fig. 2), both the normalized X-ray image and the corresponding manually segmented masks were cropped. During cropping, a smallest rectangle that contained LJS, MJS, TJS, and osteophyte was first generated. Then, a large rectangle was generated by expanding 50 pixels outward from the edges of the smallest rectangle. Thus, the normalized X-ray image and the corresponding manually segmented mask were cropped along the boundary of the large rectangle.

Given that the size of cropped normalized X-ray image and manually segmented mask fed into the U-Net model was not consistent between patients, spline interpolation [27–29] was used to resize the cropped normalized X-ray image and manually segmented mask. Thus, the size of the cropped normalized X-ray images and corresponding cropped manual segmentation masks was changed to 224×512, and then these cropped normalized X-ray images and corresponding cropped manual segmentation masks were used to construct the U-Net model. To validate the constructed U-Net model, approximately one-fourth of the patients randomly selected from the training cohort were designated as the validation cohort (815 patients, 1041 X-ray images), while the remaining patients were used for training (2444 patients, 3206 X-ray images). Each patient's X-ray images were included in only one cohort.

The U-Net model was trained using the cropped normalized X-ray images and their corresponding cropped

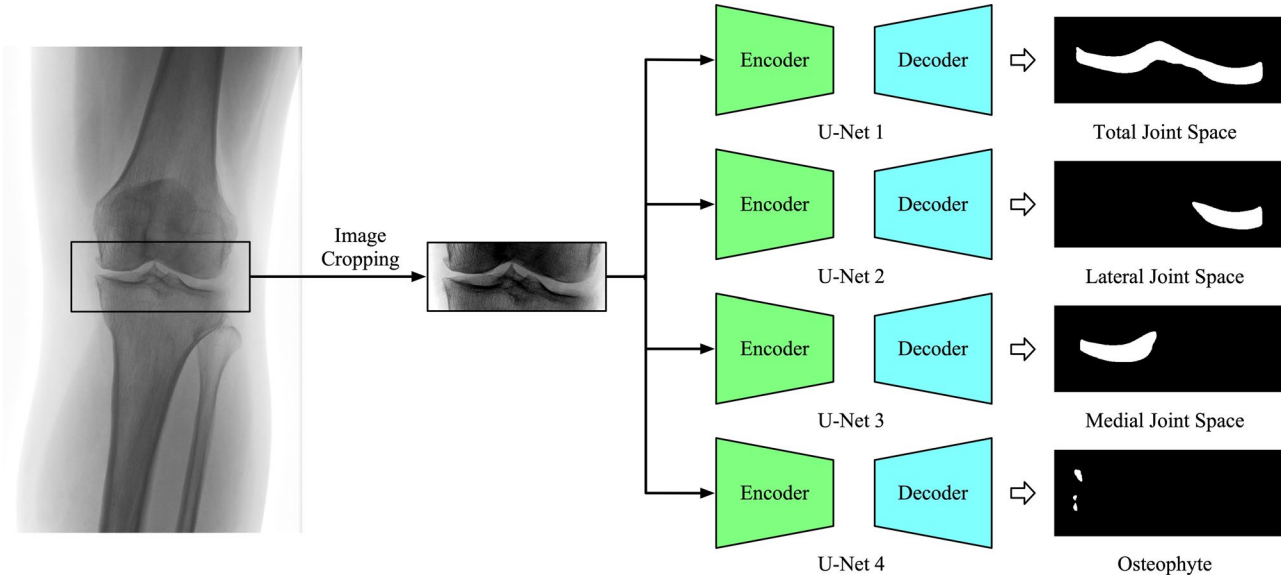

Fig. 2 The flowchart of the segmentation using the U-Net models. Four U-Net models were used to segment the total joint space, the lateral joint space, the medial joint space, and the osteophyte, respectively

manual segmentation masks from the training cohort. Subsequently, the cropped normalized X-ray images and their corresponding cropped manual segmentation masks from the validation cohort were employed to select the best-performing U-Net model. Finally, the cropped normalized X-ray images and their corresponding cropped manual segmentation masks from the testing cohort were used to evaluate the performance of the selected U-Net model. Segmentation result was obtained based on constructed U-Net model. The size of the segmentation result was restored to the size of the corresponding original cropped manual segmentation mask.

For the osteophyte segmentation, X-ray images without osteophyte were excluded from the training, validation, and testing cohorts. Thus, cropped normalized X-ray images with osteophyte in training cohort were used to construct the U-Net model, and cropped normalized X-ray images with osteophyte in validation cohort were used to select the best-performing U-Net model. The cropped normalized X-ray images with osteophyte from the testing cohort were used to evaluate the performance of the selected U-Net model. In addition, *Supplementary Method*1 provided details of another U-Net model that constructed using total training cohort including X-ray images without osteophyte to segment osteophyte.

The U-Net models were constructed using PyTorch on a Windows computer. The GPU used was an NVIDIA Quadro RTX 5000 with 16 GB of memory, while the CPU was an Intel Core i9-9980XE with 128 GB of memory. The training process involved setting the number of epochs to 100, the batch size to 4, and the initial learning rate to 0.0001. During training, the learning rate was updated every 20 epochs by dividing the previous learning rate by 10. The optimization of the U-Net model was performed using adaptive moment estimation (Adam). Further details regarding the evaluation of the U-Net model can be found in *Supplementary Method*2.

Postprocessing for segmentation

After LJS, MJS, TJS, and osteophyte were segmented, all restored segmentation results were generated. As Fig. 3A shown, we found that some evident false positive regions existed in some segmentation results. To avoid the effect on feature extraction, these false positive regions were removed. Specifically, for each patient, after the restored segmentation results obtained, the number of pixels of connected regions in the restored segmentation results was calculated. For the TJS, the regions with the number of pixels less than 2500 were removed. For the LJS and MJS, only the largest connected region was retained. For the osteophyte, the opening was used to remove false positive regions. Thus, the pixel values of these false positive areas were zeroed (Fig. 3B) in the postprocessed segmentation results.

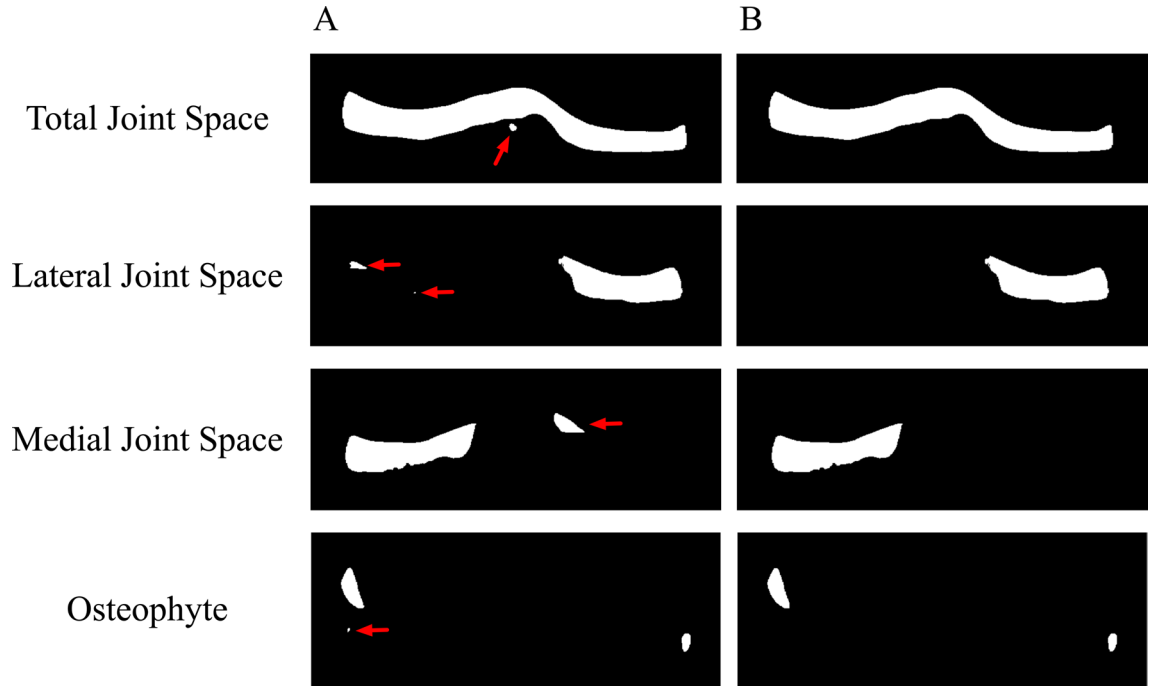

Fig. 3 Examples of the false positive regions in the restored segmentation and postprocessed results. **A**: The restored segmentation results of U-Net models were not performed postprocessing. The area marked by the red arrow was a false positive area. **B**: The restored segmentation results of U-net models were performed postprocessing

Geometric features extraction

In accordance with the current KL grading system, joint space narrowing and osteophyte formation are crucial factors in diagnosing KL. For this study, 27 geometric features were extracted from the postprocessed segmentation results of LJS, MJS, TJS, and osteophytes. These geometric features encompassed width (n=9), area (n=5), range (n=6), ratio (n=6), number (n=1). For a detailed overview of these geometric features, please refer to *Supplementary Method*3.

Radiomic feature extraction

The original X-ray images were used to extract radiomic features. To extract radiomic features from the knee subchondral bone, a region of interest was generated by expanding 20 pixels outward from the upper and lower edges of postprocessed segmentation results of TJS. The open-source pyradiomics package [30] was utilized to extract radiomic features from the original X-ray images. Prior to feature extraction, the original X-ray images underwent preprocessing using seven filters: Square Root, Wavelet, Square, Gradient, Logarithm, 2D Local Binary Pattern, and Exponential. Shape, textural and first-order features were extracted from both the original and preprocessed X-ray images, resulting in a total of 1032 features extracted from each X-ray image. Detailed information about these radiomic features can be found in the pyradiomics documentation (https://pyradiomics.readthedocs.io/en/latest/features.html). For a detailed overview of the procedure of feature extraction using Pyradiomics, please refer to *Supplementary Method*4.

Hierarchical classification method

Figure 4 illustrates the hierarchical classification method developed for predicting the severity of KOA. This method involves four sub-task classifications. First, low-level KL grades (KL grade 0–2) and high-level KL grades (KL grade 3–4) are classified. Then, the classification of KL grade 0 versus KL grade 1–2, and the classification of KL grade 3 versus KL grade 4 are performed. Finally, KL grades 1 and 2 are assigned to their respective groups.

In the hierarchical classification method, the geometric, radiomic, and combined models were constructed using geometric features, radiomic features, and the combination of geometric and radiomic features, respectively. Figure 5 provides an illustration of these models, which were used to compare the performance of different feature sets in classifying KL grades. Further details on the evaluation of the classification models can be found in *Supplementary Method*2.

Geometric model construction

In this study, logistic regression (LR) models with L2 regularization (LR_L2), LR with elastic-net regularization (LR_EN), support vector machines (SVM), random forest (RF) and extreme gradient boosting (XGBoost) were evaluated to determine the best classifier for predicting the severity of KOA based on clinical and geometric features.

Figure 5A demonstrates the construction of the geometric model, which utilized clinical features (age and gender) and geometric features. Feature selection was performed using Pearson's chi-square test for discrete features (gender) and Student's t-test for continuous features. Subsequently, the feature values in the training cohort were scaled to a range of 0 to 1 based on the minimum and maximum values of features in the training cohort. The same scaling was applied to the corresponding features in the testing cohort. The least absolute shrinkage and selection operator (LASSO) was employed for further feature selection. A random undersampling method [31] (namely, RandomUnderSampler [31, 32]) was employed in the training cohort to overcome the problem of the imbalance between majority and minority classes.

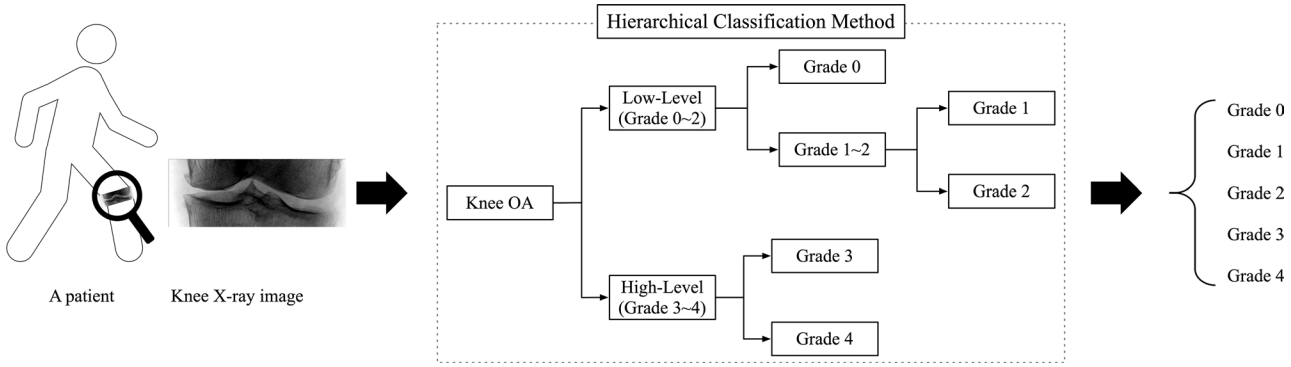

Fig. 4 The structure of hierarchical classification method. In the hierarchical classification method, there were four sub-task classifications. The low-level KL grades (i.e., KL grade 0–2) and the high-level KL grade (i.e., KL grade 3–4) were first classified. Then, the classification of the KL grade 0 and KL grade 1–2, and the classification of the KL grade 3 and KL grade 4 were performed. Last, the KL grade 1 and the KL grade 2 were classified. OA: Osteoarthritis. KL: Kellgren-Lawrence

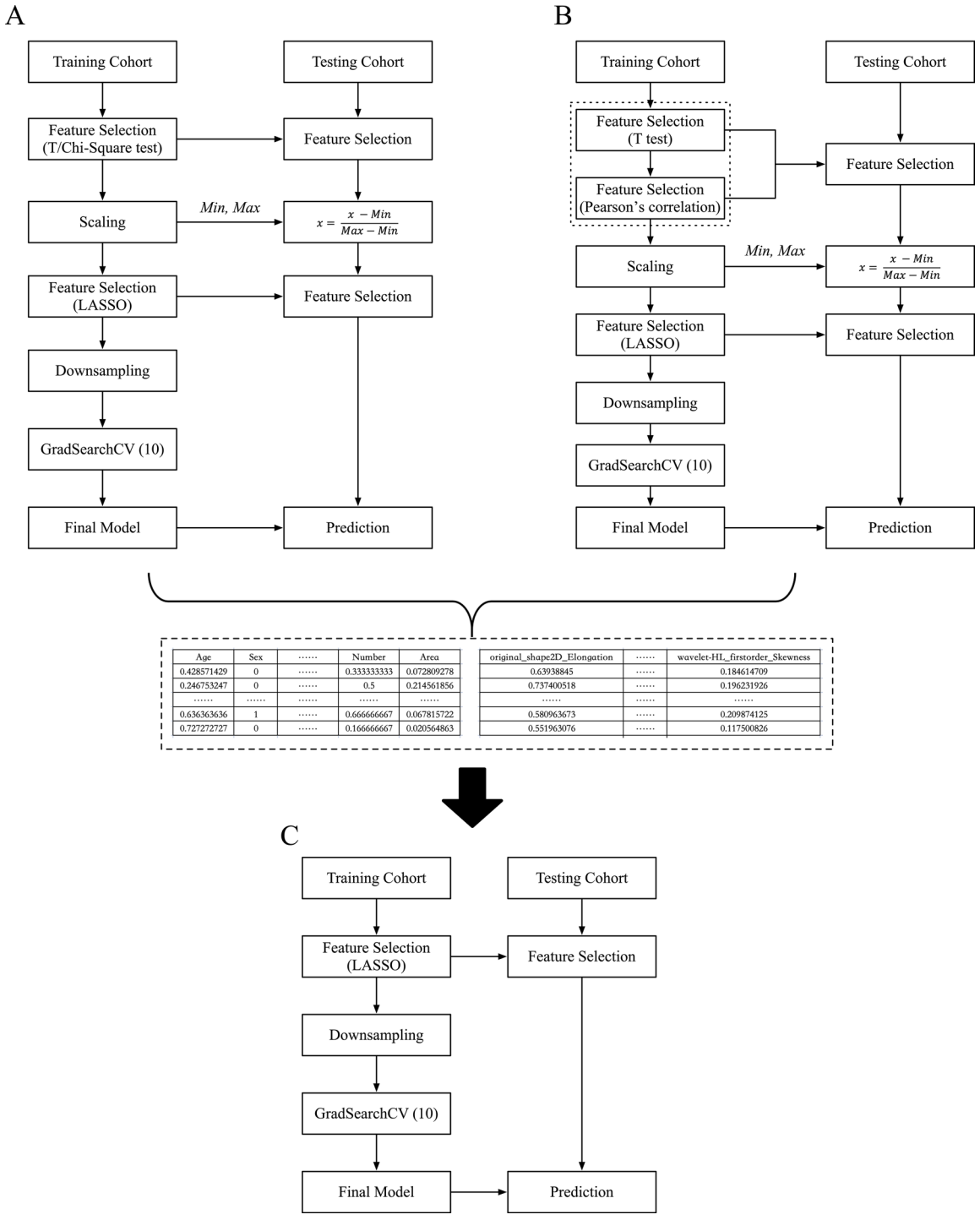

Fig. 5 Schematic overview of training and testing in the geometric model, radiomic model, and combined model. **A**: The flowchart of the construction of the geometric model. **B**: The flowchart of the construction of the radiomic model. **C**: The flowchart of the construction of the combined model. LASSO: least absolute shrinkage and selection operator. GradSearchCV (10): grid search with 10-fold cross-validation. T test: Student's *t*-test. Chi-Square test: Pearson's Chi-Square test

Optimal hyper-parameters for the classifiers were determined through a grid search with 10-fold cross-validation (CV) in the training cohort. For LR_L2, the regularization parameter λ was selected from 21 values $[2^{-10}, 2^{-9}, ..., 2^{9}, 2^{10}]$. For LR_EN, the λ and $\updownarrow_1$ ratios were selected from 21 values $[2^{-10}, 2^{-9}, ..., 2^{9}, 2^{10}]$ and 10 values [0.1, 0.2, ..., 0.9, 1], respectively. For SVM, the *C* parameter was selected from 21 values $[2^{-10}, 2^{-9},$

…, 2^9, 2^{10}]. For RF, the number of estimators and maximum depth was selected from 21 values [1, 5, 10, 15, …, 95, 100] and 19 values [10, …,95, 100], respectively. For XGBoost, the number of estimators, maximum depth and learning rate was selected from 20 value [5, 10, 15, …, 95, 100], 20 values [5, 10, 15, …, 95, 100], and 4 values [0.2, 0.1, 0.01, 0.001], respectively. Finally, classifiers with their corresponding optimal hyper-parameters were constructed using the undersampled training cohort, and these classifiers were then tested using the testing cohort. Based on the highest AUC of training cohort, the best geometric model was selected in each sub-task classification. The Scikit-learn library (version 0.23.2) was utilized for implementing the classification models.

Radiomic model construction

In the current study, LR_L2, LR_EN, SVM, RF and XGBoost models were evaluated to determine the best classifier for predicting the severity of KOA based on radiomic features.

Figure 5B illustrates the construction of the radiomic model, which utilizes radiomic features. The process of constructing the radiomic model is similar to that of the geometric model. To weaken the multicollinearity, the Pearson's correlation analysis [33, 34] was used before radiomic features scaling. Based on Pearson's correlation analysis, these radiomic features with absolute Pearson's correlation coefficient lager 0.8 were removed. In addition, LASSO was employed for further feature selection prior to applying RandomUnderSampler. After the LR_L2, LR_EN, SVM, RF and XGBoost were constructed, the best radiomic model was selected based on the highest AUC of training cohort in each sub-task classification.

Combined model construction

As depicted in Fig. 5C, based on the highest AUC value of the training cohort, the geometric features from the best geometric model and the radiomic features from the best radiomic model were stacked together to form combined features. These combined features were obtained before performing RandomUnderSampler in their respective models. LASSO was employed for feature selection prior to applying RandomUnderSampler. Optimal hyper-parameters for the classifiers were determined through a grid search with 10-fold CV in the training cohort. The hyper-parameters setting was consistent with those of the geometric model. LR_L2, LR_EN, SVM, and RF models were evaluated to determine the best classifier for predicting the severity of KOA based on the combined features.

Table 2 Results of segmentation

	Training cohort	Validation cohort	Testing cohort
Total Joint Space	0.90	0.89	0.88
Lateral Joint Space	0.89	0.86	0.86
Medial Joint Space	0.90	0.88	0.88
Osteophyte	0.75	0.64	0.64

Hierarchical classification method evaluation

To evaluate the hierarchical classification method, a strict decision strategy was employed in this study. According to this strategy, if all four sub-task classifications are accurately predicted, the output of the hierarchical classification method was considered a correct prediction. Otherwise, it was considered an incorrect prediction.

Significant features selection

For each sub-task classification in the hierarchical classification method, features that play important roles in the corresponding classification task were extracted based on the weights of the combined model. Thus, we first ranked all weights according to their absolute values and then selected the top 10% of weights. Features corresponding to these top 10% of weights were selected as significant features, which play important roles in the classification task.

Results

Patient demographics

Table 1 summarizes the patient characteristics. A total of 5317 knee joint X-ray images from 4074 patients were analyzed in the current study. Among the 5317 knee joint X-ray images, 340 were confirmed as KL grade 0, 1489 as KL grade 1, 2137 as KL grade 2, 886 as KL grade 3, and 465 as KL grade 4.

Segmentation on knee osteoarthritis

Table 2 provides an overview of the performance of the U-Net models. For TJS, the Dice similarity coefficient (Dice) values were 0.90, 0.89, and 0.88 in the training, validation, and testing cohorts, respectively. For LJS, the Dice values were 0.89, 0.86, and 0.86 in the training, validation, and testing cohorts, respectively. For MJS, the Dice values were 0.90, 0.88, and 0.88 for the training, validation, and testing cohorts, respectively. Regarding the segmentation of osteophytes, the Dice values were 0.75, 0.64, and 0.64 for the training, validation, and testing cohorts, respectively. Examples of the U-Net models' outputs can be observed in Fig. 6. In addition, *Supplementary Result*1 provided the performance of U-Net model that was constructed using training cohort including X-ray images without osteophyte.

Classification of KL grades

Table 3 presents the performances of the geometric, radiomic, and combined models for KL grading. The

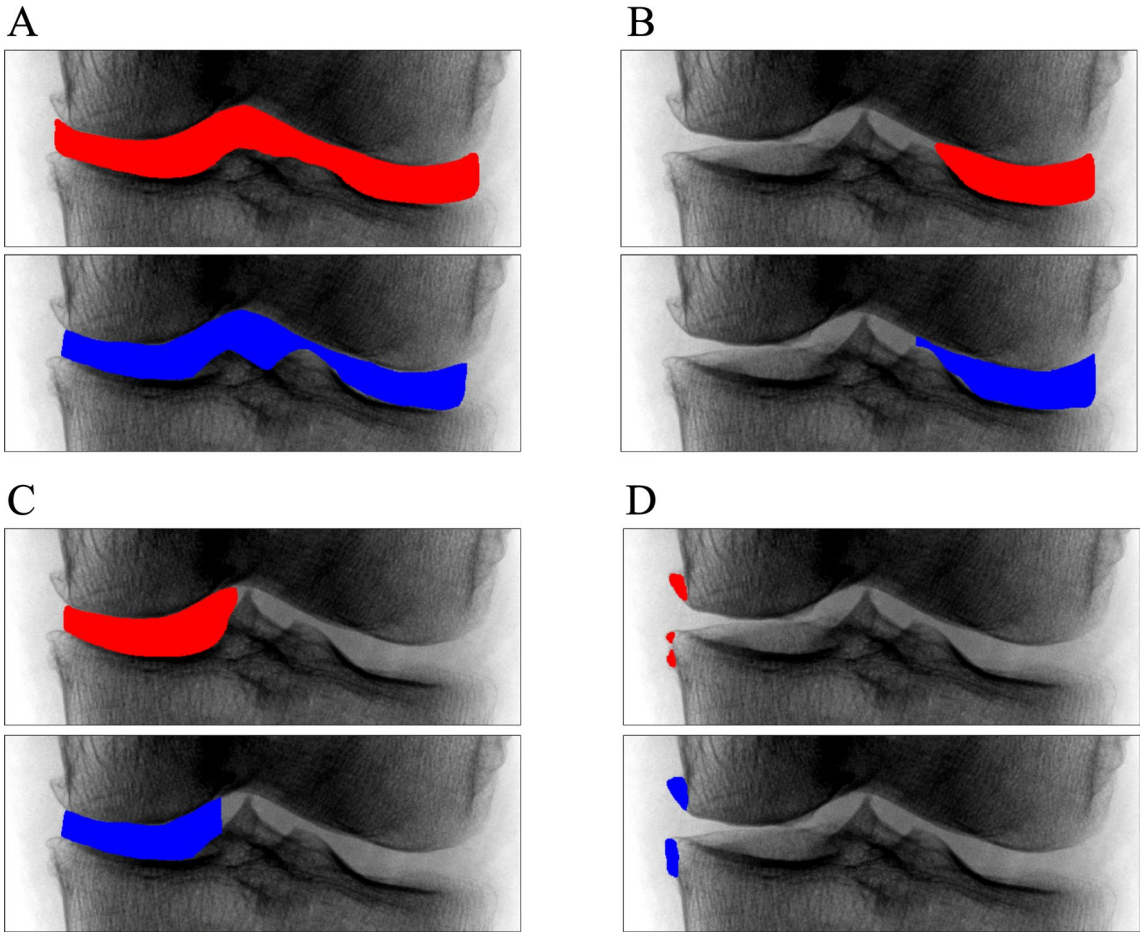

Fig. 6 Examples of segmentation results in four U-Net models. **A**: the segmentation result of the total joint space, **B**: the segmentation result of the lateral joint space, **C**: the segmentation result of the medial joint space, **D**: the segmentation result of the osteophyte. The rad: the mask segmented manually by doctors. The green: the mask segmented by the U-Net model

Table 3 Performance of geometric, radiomic, and combined models at KL grading

Classification	Model name	Accuracy	Sensitivity	Specificity	AUC
Grade 0–2 VS Grade 3–4	Geometric Model	98.41%	94.24%	99.87%	0.990
	Radiomic Model	75.14%	73.74%	75.63%	0.852
	Combined Model	98.50%	94.24%	100.00%	0.992
Grade 3 VS Grade 4	Geometric Model	81.29%	85.85%	78.49%	0.911
	Radiomic Model	72.66%	77.36%	69.77%	0.806
	Combined Model	81.65%	85.85%	79.07%	0.909
Grade 0 VS Grade 1–2	Geometric Model	81.57%	81.68%	80.30%	0.872
	Radiomic Model	77.65%	77.55%	78.79%	0.879
	Combined Model	82.07%	82.23%	80.30%	0.893
Grade 1 VS Grade 2	Geometric Model	73.00%	73.61%	72.00%	0.805
	Radiomic Model	69.01%	68.29%	70.18%	0.760
	Combined Model	74.10%	74.94%	72.73%	0.816

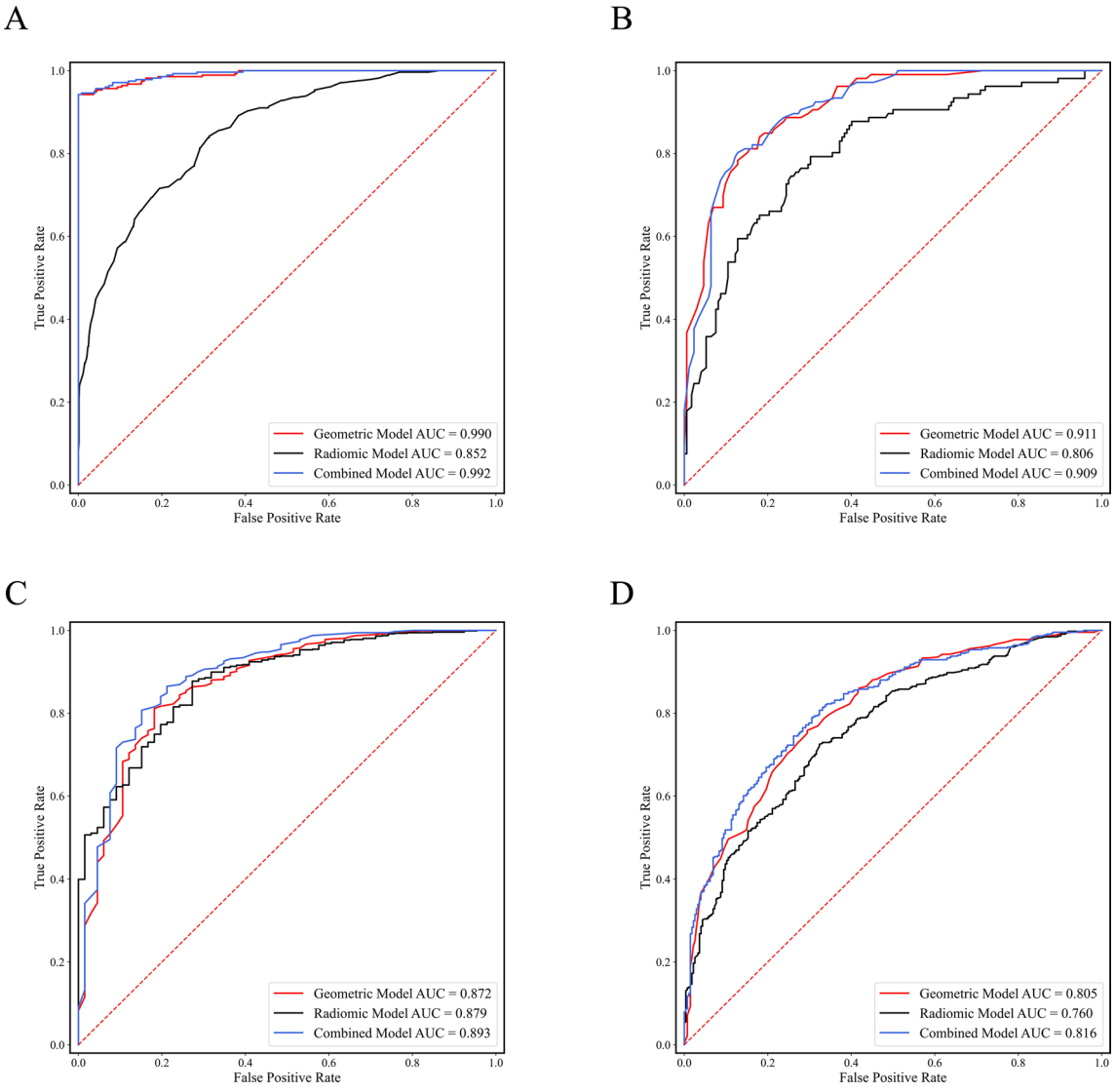

Fig. 7 The performance of the four sub-task classifications. **A**: The predictive performance in the classification of the low-level KL grade and high-level KL grade. **B**: The classification of the KL grade 3 and KL grade 4. **C**: The classification of the KL grade 0 and KL grade 1–2. **D**: The classification of the KL grade 1 and KL grade 2. KL: Kellgren-Lawrence. ROC: the receiver operating characteristic curve. AUC: the area under the ROC curve

Table 4 Significant features corresponding to these top 10% of weights in four sub-task classifications

Classification	Features name
Grade 0–2 VS Grade 3–4	Mean area of all osteophytes
	Number of osteophytes
Grade 3 VS Grade 4	Area of all osteophytes
	Min range of joint space width of TJS
Grade 0 VS Grade 1–2	Original_shape2D_Elongation
Grade 1 VS Grade 2	Mean joint space width of MJS
	Original_shape2D_Elongation

results indicate that the combined model outperformed both the geometric and radiomic models. The accuracies of the combined model were 98.50% for the classification of low-level grades (KL grades 0–2) and high-level grades (KL grades 3–4), 81.65% for KL grades 3 and 4, 82.07% for KL grades 0 and 1–2, and 74.10% for KL grades 1 and 2. The ROC curves for these classification tasks can be seen in Fig. 7.

Predictive performance of the hierarchical classification method

Regarding the hierarchical classification method, the accuracies of the geometric, radiomic, and combined models were 63.36%, 40.84%, and 65.98%, respectively.

Significant features in the hierarchical classification method

Table 4 presents the significant features corresponding to these top 10% of weights in four sub-task classifications. The significant features were listed in the Table 4 in order of absolute weight values. Mean area of all osteophytes and number of osteophytes played important roles in the classification of low-level grades and high-level grades. Area of all osteophytes and Min range of joint space width of TJS played important roles in the classification of KL grades 3 and 4. Original_shape2D_Elongation played an important role in the classification of KL grades 0 and 1–2. Mean joint space width of MJS and Original_shape2D_Elongation played important roles in the classification of KL grades 1 and 2.

Discussion

In this study, we developed a hierarchical classification method that incorporated segmentation of knee joint tissues, aiming to replicate the clinical diagnostic process and enhance the accuracy of KOA severity diagnosis. Initially, we observed satisfactory segmentation performance of the U-Net models for TJS, LJS and MJS with Dice values of 0.88, 0.86, and 0.88, respectively. Subsequently, geometric features were extracted from the segmented masks, while radiomic features were extracted from the subchondral bone of the knee joint. Finally, we constructed geometric models based on geometric features, radiomic models based on radiomic features, and combined models that integrated both types of features. These models exhibited high accuracy in classifying different levels of KOA severity. Notably, the combined model demonstrated superior classification performance compared to the geometric and radiomic models.

This study revealed the significance of geometric features in the combined model. Geometric features extracted in this study capture information about the knee joint space and osteophytes. Therefore, accurate segmentation of the knee joint space and osteophytes is crucial for precise extraction of geometric features. The U-Net model employed in this study demonstrated satisfactory performance in segmenting TJS, LJS and MJS. This segmentation performance facilitated the calculation of joint space width which is a critical factor in assessing the severity of KOA [35, 36]. The distinct calcium content between the knee joint space and the surrounding bones, such as the tibia and femur, leads to notable differences in radiation absorption [10, 37]. As a result, the knee joint space appears distinct from the tibia and femur in X-ray images. The U-Net model effectively captures the distinguishing characteristics between the knee joint space and knee bones, enhancing segmentation performance. Previous studies [38–40] have also utilized segmentation models to segment the tibia and femur bones. While segmentation of these bones aids in calculating joint space width, direct segmentation of the joint space enables straightforward calculation of different features for assessing joint space narrowing, as achieved in the present study.

The segmentation performance of osteophytes was comparatively lower than that of the knee joint space in this study. Osteophyte, being a type of bone with high calcium levels, appear as white areas in X-ray images. They exhibit minimal differentiation from other types of bone, such as the tibia and femur, making their identification challenging using segmentation models. Furthermore, the size of osteophytes is typically small compared to the total knee joint space, which may lead to the U-Net model struggling to extract relevant features for osteophyte segmentation. Consequently, although the U-Net model achieved favorable results in segmenting osteophytes, its performance was relatively lower compared to that of knee joint space segmentation.

In clinical practice, manual segmentation of the knee joint space and osteophytes relies on visual assessment. However, this approach of manual segmentation is time-consuming, labor-intensive, and subject to inter-physician variability based on their expertise and clinical experience. The proposed method in this study not only yielded excellent segmentation outcomes but also facilitated accurate classification of different KL grades for KOA severity based on these segmentation results. This suggests that the segmentation method employed significantly reduces the workload of clinicians during X-ray imaging and mitigates inter-physician inconsistency in review results. Particularly, it effectively assists less experienced physicians in performing precise knee tissue segmentation, enabling accurate assessment of KOA severity.

The current study found that both sub-task classifications and hierarchical classification method of combined model outperform that of geometric and radiomic models. Wang et al. [14] employed deep learning model to automatically diagnose the severity of knee osteoarthritis. Compared with study of Wang et al. [14], our combined model achieved higher accuracy in identifying the KL grade 2 and KL grade 4. The combined model fused geometric and radiomic features which quantified joint space narrowing and osteophyte formation from multiple perspectives. Thus, the geometric and radiomic features can provide more comprehensive information for combined model, which may be a potential reason for better performance in identifying the KL grade 2 and 4 in combined model.

However, the study of Wang et al. [14] achieved higher accuracy than that of the hierarchical classification method. Wang et al. [14] employed an end-to-end method that could directly output the prediction of KL

grade. In the hierarchical classification method, all four binary classification tasks were accurately predicted, the output of the hierarchical classification method was considered a correct prediction, which was a very strict decision strategy. In addition, Wang et al. [14] grouped KL Grade 0 and 1 into one category and included more X-ray images. Thus, the strict decision strategy, more categories and relatively small number of X-ray images may be potential reasons for relatively low in the performance of hierarchical classification method.

In addition, the current study focused on extracting geometric and radiomic features from segmented knee joint tissues to replicate the clinical diagnostic logic of KL grades and achieve accurate prediction of KOA severity, particularly in the classification of higher KL grades (KL 0–2 vs. KL 3–4). Invasive treatments, such as total knee replacement [41], are often required for patients with KL grades 3 and 4, while conservative treatments, including weight loss and pharmacotherapy [42], are more suitable for patients with KL grades 1–2. Thus, the hierarchical classification method developed in this study can assist clinicians in making personalized treatment decisions.

The study utilized various geometric features to quantify the degree of knee joint space narrowing and osteophyte formation. The findings indicated that these geometric features play crucial roles in predicting the severity of KOA. This suggests that the knee joint space and osteophyte formation can serve as significant indicators for diagnosing KL grades [25], and the geometric features extracted from these areas can be potential biomarkers for assessing the severity of KOA. Moreover, the study revealed that the geometric features extracted from the knee joint space and osteophytes have distinct roles in classifying different KL grades. Specifically, geometric features based on osteophytes are particularly important in distinguishing between low and high KL levels and between the two high levels, while geometric features based on the joint space are crucial in classifying low KL grades, namely, KL grade 1 and grade 2. The observed difference in the roles of different knee tissues in the diagnosis of various KL grades aligns with the progression of KOA. As KOA advances from mild to severe stages, the joint space gradually narrows, while the size and number of osteophytes increase [8]. Consequently, the characteristics of the joint space and osteophytes assume crucial roles in the diagnosis of mild and severe KOA, respectively. The findings from this study offer valuable insights for clinicians to adopt tailored diagnostic strategies based on the specific condition of each patient.

Furthermore, as KOA progresses, the narrowing of the knee joint space is often accompanied by subchondral bone alterations [25]. Therefore, features extracted from the subchondral bone indirectly reflect the extent of knee joint space narrowing. The Original_shape2D_Elongation played an important role in classifying low KL grades. Additionally, in comparison to the geometric model that solely utilizes geometric features, the combined model incorporating radiomic features from the subchondral bone and geometric features from the knee joint space and osteophytes exhibited improved accuracy. The combined model effectively harnesses and integrates multiple levels of information present in X-ray images, thereby assisting clinicians, particularly those with less experience, in making accurate diagnoses of KOA.

There are several limitations in this study. First, the number of X-ray images between KL grades was imbalance, especially, the number of X-ray image with KL grade 0 was too small. In future work, we will collect as many as possible X-ray images to overcome the problem of imbalanced sample sizes between KL grades. Second, the risk of overfitting was existed in model construction. To avoid the possibility of overfitting, we will collect more X-ray images, particularly the X-ray images from multiple institutes, to train model which can help the model generalize better. Third, the performance for clinical application was insufficient. In future work, we will include more clinical features and integrate features extracted by CNN into combined model to improve the performance of the hierarchical classification method.

Conclusion

This study demonstrated the feasibility of the hierarchical classification method as a viable approach for assessing the severity of KOA. The findings suggest that this method holds potential for clinical application, providing an objective means to diagnose the KL grade.

Abbreviations

Adam	Adaptive moment estimation
AUC	Area under the ROC curve
CV	Cross-validation
Dice	Dice similarity coefficient
KL	Kellgren-Lawrence
KOA	Knee osteoarthritis
LASSO	Least absolute shrinkage and selection operator
LJS	Lateral joint space
LR	Logistic regression
MJS	Medial joint space
ReLU	Rectified linear unit
RF	Random forest
ROC	Receiver operating characteristic
SVM	Support vector machines
TJS	Total joint space
XGBoost	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13075-024-03416-4.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
YW, KS, ML, and JS performed the data acquisition. ZT, JL, JT and BS performed the study design. JP analyzed and interpreted data. JP, ZT and JL were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Funding
The current study was supported by Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001), National Natural Science Foundation of China (No. 82272068, No. 82402867), Science and Technology Project of Chengdu City (No.2024-YF05-00217-SN), National Key R&D Program of China (No.2021YFA1301603), Key-Area Research and Development Program of Guangdong Province (No. 2021B0101420005), Capital's Funds for Health Improvement and Research (No. 2022-1-2061), the Fundamental Research Funds for the Central Universities (No.YWF-23-Q-1074), and the grant (GYYX24013) from 1·3·5 Project of Center for High Altitude Medicine, West China Hospital, Sichuan University.

Data availability
The original X-ray images will not be shared, because they contain private patient health information. The result data that support the findings of the present study are available on reasonable requests.

Declarations

Ethics approval and consent to participate
The present study adhered to the principles outlined in the Declaration of Helsinki and received approval from the West China Hospital, Sichuan University. Informed consent forms were signed by all participants included in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2] Department of Orthopedics Surgery, Orthopedic Research Institute, and Center for High Altitude Medicine, West China Hospital, Sichuan University, Chengdu, China
[3]School of Engineering Medicine, Beihang University, Beijing, China
[4]Key Laboratory of Big Data-Based Precision Medicine, Beihang University, Ministry of Industry and Information Technology of the People's Republic of China, Beijing, China
[5]CAS Key Laboratory of Molecular Imaging, Institute of Automation, Beijing 100190, China
[6]Department of Radiology, West China Hospital, Sichuan University, Chengdu, China

Received: 8 July 2023 / Accepted: 15 October 2024
Published online: 18 November 2024

References

1. Dainese P, Wyngaert KV, De Mits S, Wittoek R, Van Ginckel A, Calders P. Association between knee inflammation and knee pain in patients with knee osteoarthritis: a systematic review. Osteoarthritis Cartilage. 2022;30:516–34.
2. Kan HS, Chan PK, Chiu KY, Yan CH, Yeung SS, Ng YL, Shiu KW, Ho T. Non-surgical treatment of knee osteoarthritis. Hong Kong Med J. 2019;25:127–33.
3. Sharma L. Osteoarthritis of the knee. N Engl J Med. 2021;384:51–9.
4. Bijlsma JW, Berenbaum F, Lafeber FP. Osteoarthritis: an update with relevance for clinical practice. Lancet. 2011;377:2115–26.
5. Katz JN, Arant KR, Loeser RF. Diagnosis and treatment of hip and knee osteoarthritis: a review. JAMA. 2021;325:568–78.
6. Emrani PS, Katz JN, Kessler CL, Reichmann WM, Wright EA, McAlindon TE, Losina E. Joint space narrowing and Kellgren-Lawrence progression in knee osteoarthritis: an analytic literature synthesis. Osteoarthritis Cartilage. 2008;16:873–82.
7. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of Osteoarthritis. Clin Orthop Relat Res. 2016;474:1886–93.
8. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Hum YC, Tee YK, Dhanalakshmi S. Emergence of Deep Learning in Knee Osteoarthritis Diagnosis. Comput Intell Neurosci. 2021; 2021:4931437.
9. Braun HJ, Gold GE. Diagnosis of osteoarthritis: imaging. Bone. 2012;51:278–88.
10. Teoh YX, Lai KW, Usman J, Goh SL, Mohafez H, Hasikin K, Qian P, Jiang Y, Zhang Y, Dhanalakshmi S. Discovering Knee Osteoarthritis Imaging Features for Diagnosis and Prognosis: Review of Manual Imaging Grading and Machine Learning Approaches. J Healthc Eng. 2022; 2022:4138666.
11. Kose O, Acar B, Cay F, Yilmaz B, Guler F, Yuksel HY. Inter- and intraobserver reliabilities of four different Radiographic Grading scales of Osteoarthritis of the knee Joint. J Knee Surg. 2018;31:247–53.
12. Chen P, Gao L, Shi X, Allen K, Yang L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Comput Med Imaging Graph. 2019;75:84–92.
13. Zhang BF, Tan JM, Cho KY, Chang G, Deniz CM, Ieee. Attention-based CNN for KL Grade Classification: Data from the Osteoarthritis Initiative. In: *IEEE 17th International Symposium on Biomedical Imaging (ISBI): Apr 03–07 2020; Iowa, IA*; 2020: 731–735.
14. Wang CT, Huang B, Thogiti N, Zhu WX, Chang CH, Pao JL, Lai F. Successful real-world application of an osteoarthritis classification deep-learning model using 9210 knees-An orthopedic surgeon's view. J Orthop Res. 2023;41:737–46.
15. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957;16:494–502.
16. Cheung JC, Tam AY, Chan LC, Chan PK, Wen C. Superiority of multiple-joint space width over Minimum-Joint Space Width Approach in the machine learning for Radiographic severity and knee osteoarthritis progression. Biology (Basel). 2021; 10.
17. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, Sun K, Tang Z, Jiang H, Li H, et al. Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to neoadjuvant chemotherapy in breast Cancer: a Multicenter Study. Clin Cancer Res. 2019;25:3538–47.
18. Liu Z, Meng X, Zhang H, Li Z, Liu J, Sun K, Meng Y, Dai W, Xie P, Ding Y, et al. Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. Nat Commun. 2020;11:4308.
19. Wei G, Jiang P, Tang Z, Qu A, Deng X, Guo F, Sun H, Zhang Y, Gu L, Zhang S, et al. MRI radiomics in overall survival prediction of local advanced cervical cancer patients tread by adjuvant chemotherapy following concurrent chemoradiotherapy or concurrent chemoradiotherapy alone. Magn Reson Imaging. 2022;91:81–90.
20. Tang Z, Zhang XY, Liu Z, Li XT, Shi YJ, Wang S, Fang M, Shen C, Dong E, Sun YS, et al. Quantitative analysis of diffusion weighted imaging to predict pathological good response to neoadjuvant chemoradiation for locally advanced rectal cancer. Radiother Oncol. 2019;132:100–8.
21. Hirvasniemi J, Klein S, Bierma-Zeinstra S, Vernooij MW, Schiphof D, Oei EHG. A machine learning approach to distinguish between knees without and with osteoarthritis using MRI-based radiomic features from tibial bone. Eur Radiol. 2021;31:8513–21.
22. Anifah L, Purnama IK, Hariadi M, Purnomo MH. Osteoarthritis classification using self organizing map based on gabor kernel and contrast-limited adaptive histogram equalization. Open Biomed Eng J. 2013;7:18–28.
23. Rosenberg TD, Paulos LE, Parker RD, Coward DB, Scott SM. The forty-five-degree posteroanterior flexion weight-bearing radiograph of the knee. J Bone Joint Surg Am. 1988;70:1479–83.
24. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006;31:1116–28.
25. Schiphof D, Boers M, Bierma-Zeinstra SM. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann Rheum Dis. 2008;67:1034–6.
26. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9*, 2015, Proceedings, Part III 18: 2015: Springer; 2015: 234–241.

27. Heising L, Angelopoulos S. Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN. BMJ Health Care Inf. 2022; 29.
28. Kolarik M, Burget R, Travieso-Gonzalez CM, Kocica J. Planar 3D transfer learning for end to end unimodal MRI unbalanced data segmentation. In: 2020 *25th International Conference on Pattern Recognition (ICPR): 2021*: IEEE; 2021: 6051–6058.
29. Tan JW, Wang L, Chen Y, Xi W, Ji J, Wang L, Xu X, Zou LK, Feng JX, Zhang J, et al. Predicting Chemotherapeutic Response for Far-advanced gastric Cancer by Radiomics with Deep Learning semi-automatic segmentation. J Cancer. 2020;11:7224–36.
30. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin J-C, Pieper S, Aerts HJ. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104–7.
31. Xie C, Du R, Ho JW, Pang HH, Chiu KW, Lee EY, Vardhanabhuti V. Effect of machine learning re-sampling techniques for imbalanced datasets in (18) F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. Eur J Nucl Med Mol Imaging. 2020;47:2826–35.
32. Gou W, Zhang H, Zhang R. Multi-classification and Tree-Based Ensemble Network for the intrusion detection system in the Internet of vehicles. Sens (Basel). 2023; 23.
33. Zheng Y, Han X, Jia X, Ding C, Zhang K, Li H, et al. Dual-energy CT-based radiomics for predicting invasiveness of lung adenocarcinoma appearing as ground-glass nodules. Front Oncol. 2023;13:1208758.
34. Wang X, Dai Y, Lin H, Cheng J, Zhang Y, Cao M, Zhou Y. Shape and texture analyses based on conventional MRI for the preoperative prediction of the aggressiveness of pituitary adenomas. Eur Radiol. 2023;33:3312–21.
35. Ghouri A, Muzumdar S, Barr AJ, Robinson E, Murdoch C, Kingsbury SR, Conaghan PG. The relationship between meniscal pathologies, cartilage loss, joint replacement and pain in knee osteoarthritis: a systematic review. Osteoarthritis Cartilage. 2022;30:1287–327.
36. Gale DR, Chaisson CE, Totterman SM, Schwartz RK, Gale ME, Felson D. Meniscal subluxation: association with osteoarthritis and joint space narrowing. Osteoarthritis Cartilage. 1999;7:526–32.
37. Momose A. X-ray phase imaging reaching clinical uses. Phys Med. 2020;79:93–102.
38. Kim YJ, Lee SR, Choi JY, Kim KG. Using Convolutional Neural Network with Taguchi Parametric Optimization for Knee Segmentation from X-Ray Images. Biomed Res Int. 2021; 2021:5521009.
39. Zhou Z, Zhao G, Kijowski R, Liu F. Deep convolutional neural network for segmentation of knee joint anatomy. Magn Reson Med. 2018;80:2759–70.
40. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magn Reson Med. 2018;79:2379–91.
41. Jang S, Lee K, Ju JH. Recent updates of diagnosis, pathophysiology, and treatment on Osteoarthritis of the knee. Int J Mol Sci 2021; 22.
42. Michael JW, Schluter-Brust KU, Eysel P. The epidemiology, etiology, diagnosis, and treatment of osteoarthritis of the knee. Dtsch Arztebl Int. 2010;107:152–62.

Publisher's note

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