Progress in Medical Devices
ISSN: 2957-5478

Zentime

Research Article https://doi.org/10.61189/517419bnpxap

Research on knee osteoarthritis grading based on multidimensional feature fusion

He Ren*, Yina Zhang*, Yutong Xie*, Anqi Wu, Xianglun Kong, Chenxiao Bai, Miao Yu, Yimeng Wang, Ping Li

Faculty of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China.
*The authors contribute equally.

Corresponding author: Ping Li.

Declaration of conflict of interest: None.

Declaration of ethics: This study uses an open-source dataset, which has been processed in accordance with relevant ethical guidelines during its collection and release. The data provider has completed the necessary ethical review procedures to ensure that data collection complies with internationally recognized ethical norms such as the Declaration of Helsinki. Strict de-identification processing has been applied to information involving individuals to fully protect the privacy and rights of data providers. Since this study does not directly contact original individuals and only analyzes de-identified open-source data, it does not involve additional risks to human participants, and thus there is no need to apply for ethical review again.

Declaration of patient consent: As the open-source data used in this study has been de-identified and cannot be traced back to specific patients, there is no risk of violating patients' privacy. Therefore, it is not necessary to obtain informed consent from patients.

Received July 17, 2025; Accepted September 29, 2025; Published December 31, 2025

Highlights

- Based on 8,110 OAI X-ray images, 18 radiomic features were processed and selected, with SMOTE combined with category weight balancing employed to address data imbalance.
- Among eight machine learning models, the SVM achieved the best performance.
- SHAP and LIME analyses enhanced model interpretability by identifying key radiomic features influencing predictions.

Abstract

Objective: This study aimed to apply machine learning approaches to the Kellgren-Lawrence (KL) grading of knee osteoarthritis, develop an effective automatic KL grading technique, and provide a methodological reference for clinical diagnosis and research. **Methods:** Data were obtained from the Osteoarthritis Initiative (OAI) knee X-ray image dataset, comprising 8,110 images from the folders of *auto*_test, *train*, and *val*. All images were first subjected to inversion processing, followed by extraction of two-dimensional radiomic features. Feature selection was then conducted using a combination of variance thresholding and analysis of variance (ANOVA), yielding 18 key features. To address class imbalance in the original dataset, this synthetic minority over-sampling technique (SMOTE) and class weight balancing were jointly applied. Eight machine learning models—Decision Trees (DT), Logistic Regression (LR), Random Forests (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Light Gradient Boosting Machine (LightGBM)—were trained for KL grading of knee osteoarthritis. Model performance was evaluated using accuracy, precision, recall, F1-score, and the area under the curve (AUC). For the optimal SVM model, global and local interpretability analy-

Address correspondence to: Ping Li, Faculty of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, No. 279, Zhouzhu Highway, Pudong New Area, Shanghai 201318, China. Tel: +86-13764055848. E-mail: lip@sumhs.edu.cn.

ses were further conducted using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to identify the key factors influencing model decisions. **Results:** The support vector machine (SVM) model achieves the best performance. **Conclusion:** This study establishes an effective machine learning-based method for automatic KL grading of knee osteoarthritis, providing valuable support for clinical diagnosis and research applications.

Keywords: Osteoarthritis, knee osteoarthritis, Kellgren-Lawrence (KL) grading, machine learning, interpretability

Introduction

Arthritis is a category of diseases primarily characterized by joint pain, swelling, stiffness, deformity, and functional impairment. Its major types include osteoarthritis, rheumatoid arthritis, ankylosing spondylitis, and gouty arthritis [1-5]. Among these, osteoarthritis is the most prevalent form. Knee osteoarthritis, a common subtype, is characterized by degeneration of knee cartilage and osteophyte formation, predominantly affecting middle-aged and elderly individuals. Due to pain, many patients with knee osteoarthritis reduce their physical activity, leading to impaired lower-limb circulation [6, 7]. The release of inflammatory factor further compromises vascular endothelial function, causing a hypercoagulable state and an increased risk of lower-limb venous thrombosis [8]. Once detached, thrombi may travel to the lungs via the bloodstream, potentially causing pulmonary embolism and fatal complications [9, 10].

Conventional diagnostic methods for knee osteoarthritis, including physical examination, X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and arthrography, each have limitations. Physical examination outcomes rely on physician experience and subjective judgment, lacking standardized criteria. X-ray and CT are limited in detecting early cartilage wear, bone changes, or soft-tissue lesions [11-13]. MRI, though sensitive, is costly, contraindicated in patients with metal implants, and prone to metal artifacts that degrade image quality. Arthrography is invasive, technically demanding, and carries risks such as infection. Additionally, the high prevalence of depression among knee osteoarthritis patients contributes to substantial socioeconomic burdens [14]. The Kellgren-Lawrence (KL) grading system is a classic radiological tool for assessing the severity of osteoarthritis. However, its accuracy is highly dependent on the subjective interpretation of radiologists, potentially affecting diagnostic consistency [15]. Artificial intelligence (AI) models can reduce subjectivity by enabling standardized feature extraction. Yet, conventional deep learning models often lack interpretability, limiting their acceptance as diagnostic evidence in clinical practice [16].

Machine learning, value for its interpretability, has been increasingly used in disease diagnosis and classification. However, most existing studies suffer from small sample sizes, leading to a higher risk of model overfitting and limited generalization. Additionally, in studies utilizing multiple classifiers, the best classification accuracies were 85.6% and 82.61% [14, 17]. These results fall short of the clinical goal of supporting consistent diagnostic decision-making.

Building upon prior research, this study aims to develop a novel diagnostic method with improved diagnostic accuracy and cost-effectiveness to advance osteoarthritis diagnosis. Specifically, quantitative radiomic features were extracted from knee X-ray images using PyRadiomics to establish a database for model training. Eight machine learning classifiers, including Support Vector Machine (SVM) and AdaBoost, were applied for KL grading. Furthermore, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) were employed to interpret the model outputs from both global and local perspectives, enhancing the transparency and clinical interpretability for the results.

Materials and methods

Datasets

Data for this study were sourced from the Osteoarthritis Initiative (OAI) knee osteoarthritis dataset, which includes X-ray images annotated with corresponding KL severity grades. The dataset contains 9,786 knee X-ray images used for knee joint detection and KL grading, organized into four folders: auto_test, test, train, and val. Each folder contains images classified into five KL grades: Grade 0 (Normal), Grade 1 (Suspicious), Grade 2 (Mild), Grade 3 (Moderate),

Table 1. Kellgren-Lawrence (KL) grading and corresponding imaging features for knee osteoarthritis

Class	Kellgren-Lawrence Grading	Imaging Features	Number of X-ray images
Grade 0	Normal	Normal joint space, no osteophytes, no subchondral sclerosis.	3212
Grade 1	Doubtful	Possible joint space narrowing, with possible osteophyte formation in a lip - like shape.	1474
Grade 2	Minimal	Definite osteophyte formation, possible joint space narrowing.	2117
Grade 3	Moderate	Multiple osteophytes, definite joint space narrowing, mild subchondral sclerosis.	1063
Grade 4	Severe	Large osteophytes, marked joint space narrowing, severe subchondral sclerosis.	244

and Grade 4 (Severe). This study selected 8,110 X-ray images from the auto_test, train, and val folders for analysis (**Table 1**).

Data preprocessing

Image processing

Color X-ray images were first converted to grayscale. Image inversion was applied to enhance bone structures for threshold-based segmentation. Subsequently, Otsu's thresholding method was applied to segment the image into foreground (bone) and background regions. Only foreground pixel values were retained, with background pixels being set to 0 for subsequent refinement. A second Otsu thresholding was performed on the foreground to eliminate low-intensity noise and preserve the high-intensity bone core area. A binary mask was generated to locate the Region of Interest (ROI). The upper and lower boundaries of the knee joint were determined using row-pixel statistics combined with a sliding-window search. A quadrilateral ROI consistent with the knee's anatomical structure was delineated through endpoint positioning and spacing adjustment. Finally, polygon filling was applied to generate the mask and define the effective area for feature extraction.

Feature extraction

2D radiomic features were extracted from the preprocessed knee X-ray images using the *Pyradiomics* library. Initially, the ROI was delineated through inversion and binarization of the original image, followed by connected component analysis to remove noise and retain the largest bone-connected region. Subsequently, vertical artifacts were excluded based on column-pixel statistics, generating a rectangular mask that encompassed the core knee joint structure (**Figure 1**). The preprocessed grayscale images and ROI masks were then converted to the *SimpleITK* format to ensure spatial alignment. Using a predefined parameter configuration, 18 first-order statistical features, 14 shape-based features, and 48 texture features were extracted, totaling 80 quantitative features. Finally, feature extraction was performed on each ROI, generating a dataset with feature values, image filenames, and diagnostic labels for subsequent analysis.

Feature selection

To select features with significant influence on the target variable, a two-step selection process combining variance thresholding and analysis of variance (ANOVA) were used. Features with a variance greater than 0.01 were first retained. Subsequently, the SelectKBest function with the *f_classif* criterion was applied to determine features most strongly correlated with the target variable. To further enhance the accuracy and reliability of feature selection, this study integrated variance thresholding and correlation analysis. Specifically, features meeting the variance threshold and showing strong correlation with the target variable were selected, yielding 18 optimal features, which served as the final feature set for subsequent modeling.

Data balancing

The dataset used in this study exhibits class imbalance, with varying sample sizes across categories. This imbalance can bias the model toward majority classes, resulting in suboptimal performance on minority classes. To address this issue, synthetic minority over-sampling technique (SMOTE) was applied to augment the

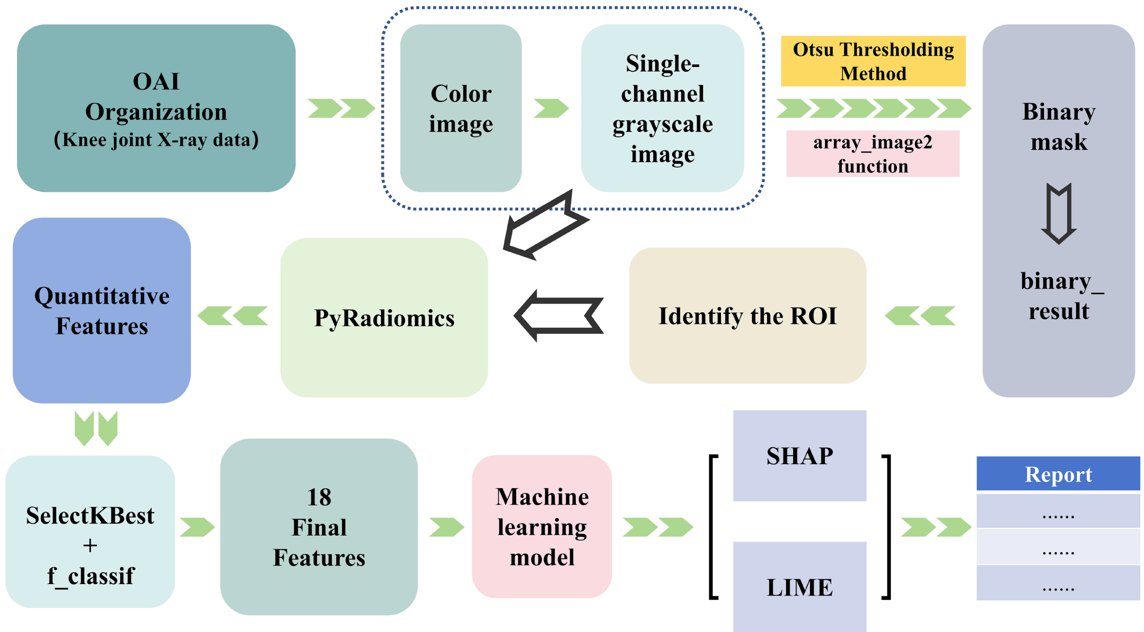


Figure 1. A comprehensive analytical pipeline integrating image processing, radiomic feature extraction, machine-learning model training, and model interpretation. The workflow is designed to extract meaningful information from medical imaging data and facilitate disease diagnosis.

minority class samples, as illustrated in **Figure 2**. SMOTE generates synthetic samples between existing minority - class samples to balance the dataset. Additionally, during model training, different weights were assigned to each class to make the model focus more on the minority - class samples.

To avoid data leakage and ensure the model's generalization to real-world scenarios, the SMOTE algorithm was exclusively applied to the training set to generate synthetic minority samples, while the validation and test sets retained their original class distributions. This strategy mitigated overfitting by preventing the model from learning synthetic patterns that do not exist in real data, enabling an unbiased evaluation of model performance on naturally imbalanced clinical datasets.

Model construction

Support vector machine (SVM)

SVM is characterized by strong generalization ability. It can accurately classify patients based on limited data. Also, its robustness allows tolerance to a certain degree of outliers and noise, contributing to the development of a stable five–class classification model. Thus, SVM performs well in this study. As a supervised learning algorithm based on statistical learning theory, SVM distinguishes samples of different classes by finding a hyperplane that maximizes the margin. Its fundamental principle involves mapping data into a high-dimensional space to identify an optimal hyperplane. This hyperplane separates samples of different classes with the maximum margin.

Other models

In addition to SVM, several other machine learning algorithms were evaluated in this study. The Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) models demonstrated outstanding performance. LightGBM uses a histogram-based algorithm and a leaf-wise growth strategy to enhance training efficiency. It can automatically handle missing values and assess feature importance, thereby assisting in feature selection. Its precise predictions, strong generalization ability, and effectiveness in addressing data imbalance make it particularly suitable for this research.

XGBoost, a gradient boosting decision tree algorithm, incorporates parallel computing and optimization, delivering superior prediction accuracy and automatically handling missing data. These features aid in understanding the rela-

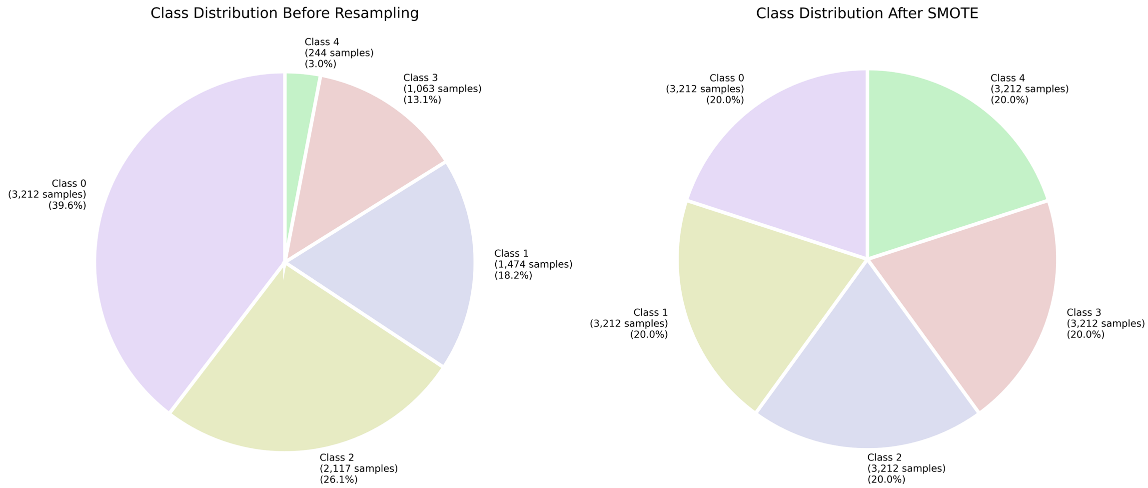


Figure 2. Comparison of the class distribution in the dataset before and after balancing.

tionship between features and model outputs, making them suitable for the five-class classification of knee osteoarthritis.

The K-nearest neighbors (KNN) is simple and intuitive but sensitive to high-dimensional data and noise. Categorical Boosting (CatBoost) performs well in handling imbalanced data by automatically managing class weights and mitigating overfitting. Decision Tree (DT) models offer good interpretability but are prone to overfitting. Random Forest (RF) enhances stability by integrating multiple decision trees but is less efficient in handling imbalanced data. Logistic Regression (LR), a linear model, struggles to capture nonlinear relationships, leading to insufficient classification accuracy. However, it offers good interpretability, simplicity, and ease of use for clinical applications and decision support. Additionally, it provides probability predictions for richer diagnostic information. These advantages make it valuable in specific contexts.

Model evaluation

Model performance was evaluated using five standard metrics: accuracy, precision, recall, F1 score, and area under the curve (AUC). Accuracy represents the proportion of correctly classified instances, indicating overall model correctness. High accuracy indicates that the model performs well in classification. Precision reflects the proportion of true positives among all predicted positives, where higher precision means fewer false positives and improved diagnostic reliability. Recall (sensitivity) evaluates a model's ability to correctly identify actual positive samples; high recall indicates good classification performance. The F1 score balances precision and recall, helping to identify the optimal model. AUC measures a model's ability to distinguish between classes. A higher AUC value indicates better performance and is not affected by class imbalance. These metrics provide a comprehensive evaluation of model performance for the five-class knee osteoarthritis classification task.

Model explainability

To enhance model transparency and clinical trust, interpretability analyses were conducted using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP quantifies feature contributions to model predictions based on Shapley values. It assigns importance values to features, where positive values indicate a promotion effect and negative values indicate an inhibitory effect on the prediction. This reflects the influence of each feature on the model's output.

LIME, on the other hand, explains individual predictions by constructing local surrogate linear models around sample points, simplifying complex models and visually demonstrating feature impacts. In LIME analysis, high - weight features significantly influence predictions of specific arthritis types. These techniques help doctors understand model reasoning, increasing diagnostic credibility and patient acceptance. They also improve the reliability and clinical applicability of models in diagnosing knee osteoarthritis.

Table 2. Comparison of model performance

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM	0.9284	0.9271	0.9282	0.9270	0.9906
DT	0.7905	0.7872	0.7900	0.7870	0.8872
XGBoost	0.9128	0.9116	0.9125	0.9118	0.9874
KNN	0.8776	0.8840	0.8761	0.8626	0.9558
CatBoost	0.8238	0.8202	0.8234	0.8193	0.9596
LightGBM	0.9141	0.9138	0.9141	0.9139	0.9894
LR	0.4038	0.3819	0.4056	0.3856	0.7183
RF	0.5682	0.5587	0.5687	0.5579	0.8460

Results

Comparison of model performance

In this study, eight machine learning models were applied to the five-class classification task of knee osteoarthritis. The models were compared based on accuracy, precision, recall, F1 score, and AUC. Additionally, receiver operating characteristic (ROC) and Precision-Recall (PR) curves were plotted for visualization. SHAP and LIME were used to illustrate the model's interpretability through multiple plots, providing a comprehensive assessment of the models' performance.

As shown in **Table 2**, SVM outperformed other models across multiple metrics. It achieved an accuracy of 0.9284, with precision and recall at 0.9271 and 0.9282, respectively, indicating excellent classification consistency. The SVM model also yielded an F1 score of 0.9270 and an AUC of 0.9906, further demonstrating its robustness and discriminative capability in the medical grading task.

XGBoost and LightGBM, as representatives of gradient boosting frameworks, also performed well. XGBoost achieved an accuracy of 0.9128, while LightGBM achieved an accuracy of 0.9141, both with AUC values exceeding 0.98. LightGBM achieved an F1 score of 0.9139, approaching that of SVM, indicating its robustness in handling complex feature interactions.

The KNN algorithm demonstrated acceptable overall performance but had a relatively lower F1 score (0.8626), showing slight model imbalance. CatBoost and DT models achieved accuracies of 0.8238 and 0.7905, respectively, with AUC values between 0.88 and 0.96. CatBoost, despite its advantages in handling categorical features, performed moderately in this task. DT was prone to overfitting. In contrast, RF and LR had the lowest accuracy at 0.5682 and 0.4038.

The performance of the eight models was further evaluated by plotting their ROC and PR curves. The ROC curve reflects each model's discriminative ability across different classification thresholds. **Figure 3A** shows that the curves of SVM, LightGBM, and XGBoost are closer to the upper left corner, indicating superior ability to accurately identifying positive samples while minimizing false-positive predictions. In contrast, the ROC curves of the other models are closer to the diagonal, indicating weaker discriminative ability. PR curves display a model's precision across varying recall levels. **Figure 3B** shows that SVM, LightGBM, and XGBoost maintain high precision across most recall ranges, indicating a higher probability that identified positive samples are truly positive. By contrast, the PR curves for LR and RF showed a faster decline in precision as recall increases. This suggests these models have more misclassifications when identifying positive samples, resulting in relatively poorer performance.

Figure 4 and **Table 3** show that the balancing strategy used in this study effectively improved recognition performance for minority classes. Grade 4 achieved a recall rate of 99.84%, with no false-negative predictions in the confusion matrix, indicating the model's excellent ability to identify severe lesions. For Grade 1 and Grade 3, the recall rates were 96.77% and 97.13%, respectively, indicating reliable identification of both early and moderate disease stages while minimizing missed or misdiagnosed cases. Grade 0 has a relatively lower recall rate of 78.67%, and most misclassifications were attributed to normal samples being misclassified as Grade 1 (35 cases) or Grade 2

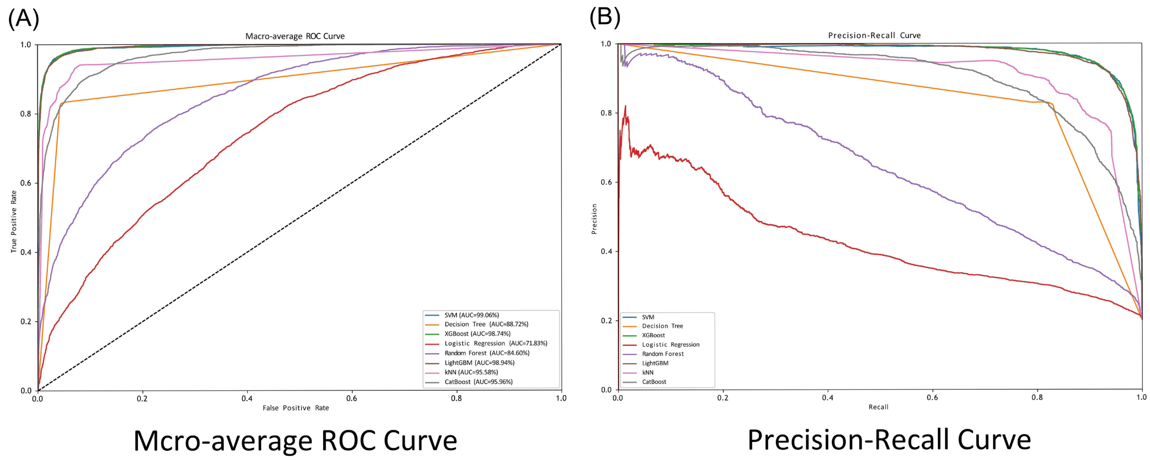


Figure 3. Receiver operating characteristic (ROC) and precision-recall (PR) curves. (A) Macro-averaged ROC curves illustrating the classification performance of eight models across different thresholds; (B) PR curves illustrating the relationship between precision and recall, reflecting the accuracy and completeness of model predictions at varying thresholds.

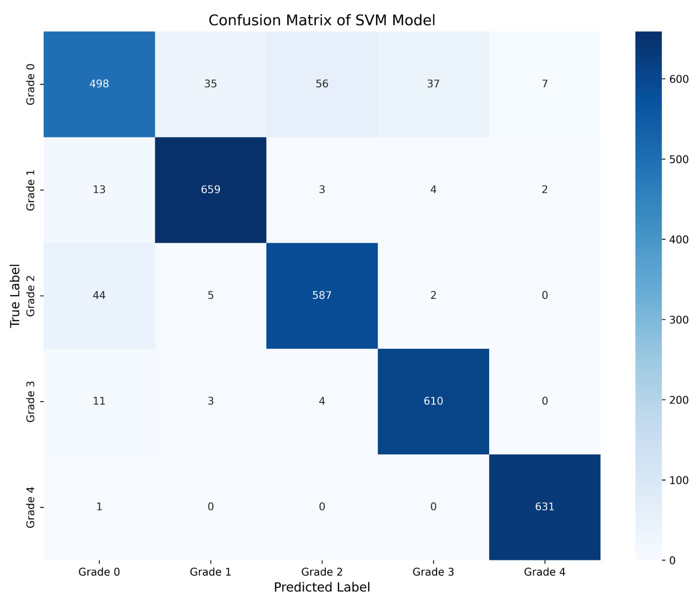


Figure 4. Confusion matrix of the support vector machine (SVM) model for classifying different Kellgren-Lawrence (KL) grades of knee osteoarthritis. The color gradient represents the number of samples, with darker blue indicating a higher count of samples in that cell.

Table 3. Recall (sensitivity) of the support vector machine (SVM) model for each Kellgren-Lawrence (KL) grade of knee osteoarthritis

KL Grade	Recall (Sensitivity)
0	0.7867
1	0.9677
2	0.9201
3	0.9713
4	0.9984

(56 cases). These misclassifications, considered "conservative errors" in clinical practice, can be corrected through further imaging reviews without leading to serious diagnostic risks. Overall, the balancing strategy significantly improves the detection capability for minority classes, especially Grade 4, without inflating overall performance metrics, aligning with clinical diagnosis requirements.

Results of model interpretation

A total of 18 features were ultimately selected for KL grading diagnosis of knee osteoarthritis. Features were extracted using different processing methods and statistical combinations, as well as from diverse data sources. **Table 4** summarizes various key features used in image modeling and analysis, including first-order statistical features (e.g., Feature 04, 18) and features derived from mathematical transformations such as mean absolute deviation, robust mean absolute deviation, entropy, and kurtosis (e.g., Feature 08, 11, 15).

SHAP-based local interpretability analysis

SHAP was used to interpret feature contributions and model decisions. SHAP, based on Shapley values from cooperative game theory,

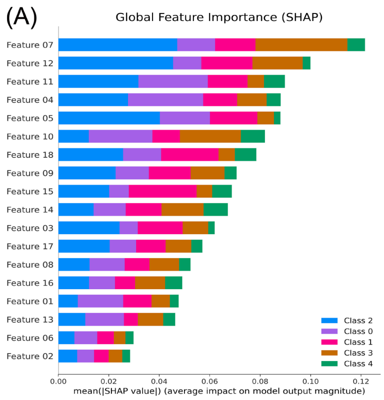
Table 4. Abbreviated feature identifiers and their corresponding definitions

Features	Corresponding ID
square_firstorder_InterquartileRange	Feature 01
squareroot_firstorder_InterquartileRange	Feature 02
square_firstorder_MeanAbsoluteDeviation	Feature 03
diagnostics_Image-original_Minimum	Feature 04
wavelet-LL_firstorder_InterquartileRange	Feature 05
original_firstorder_InterquartileRange	Feature 06
square_firstorder_Entropy	Feature 07
gradient_firstorder_Skewness	Feature 08
original_firstorder_Uniformity	Feature 09
wavelet-LL_glrlm_RunEntropy	Feature 10
square_firstorder_RobustMeanAbsoluteDeviation	Feature 11
wavelet-LL_firstorder_MeanAbsoluteDeviation	Feature 12
wavelet-LL_firstorder_RobustMeanAbsoluteDeviation	Feature 13
square_firstorder_Kurtosis	Feature 14
squareroot_firstorder_RobustMeanAbsoluteDeviation	Feature 15
exponential_firstorder_InterquartileRange	Feature 16
exponential_firstorder_RobustMeanAbsoluteDeviation	Feature 17
original_firstorder_RobustMeanAbsoluteDeviation	Feature 18

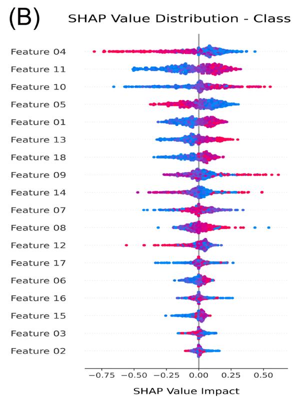
Note: Interquartile Range (IQR) represents the distribution range of the middle 50% of pixel values.

calculates each feature's contribution to predictions, offering an intuitive understanding of how features affect model outputs. **Figure 5** presents the feature importance distribution for the SVM model in the grading task. Overall, Feature 07 and Feature 12 significantly impacted model's predictive performance. A high SHAP value of Feature 07 indicates its core role in assessing lesion severity. Feature 07, a metric quantifying the disorder of image gray value distribution, directly links to key radiographic signs of knee X-rays across KL grades. For KL Grade 0-1, low entropy aligns with the uniform gray distribution of normal joint spaces or mild lip-like osteophytes in Grade 1, as bone tissue maintains high-density uniformity with minimal gray fluctuations. For Grades 2-3, significantly increased entropy correlates with density heterogeneity from definite osteophytes and patchy subchondral sclerosis—osteophyte proliferation enhances local gray-level fluctuations, while contrast between sclerotic and normal bone further boosts entropy. For Grade 4, peak entropy matches the complex structure of large osteophytes and marked joint space narrowing; interleaved bone destruction-repair regions cause the highest gray distribution disorder, consistent with the clinical radiographic feature of 'density inhomogeneity in severe lesions'. This mapping confirms entropy's clinical value as a texture metric and provides an interpretable radiological basis for the model's decisions. Subplots (B-F) display the direction and intensity of feature contributions through swarm plots. The x-axis of the SHAP summary plot represents SHAP values, where positive values indicate that the feature drives predictions toward a specific class, while negative values suggest the opposite. For class 0, Feature 04 significantly impacts model predictions, and a broader distribution of feature values may lead to variations in model predictions. In class 1, Feature 15 and Feature 18 negatively impact model predictions. For class 2, Feature 07 and Feature 12 show significant positive impacts. In class 3, Feature 07 has a strong positive influence. For class 4, Feature 14 positively affects predictions but has a wide distribution, indicating variable impacts on the model output.

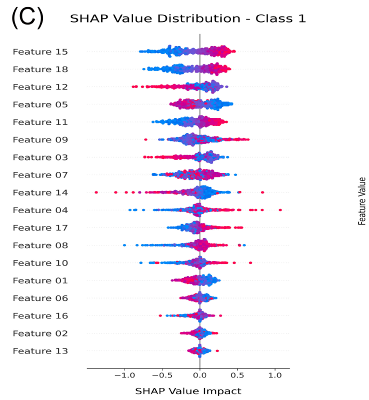
This study employed random selection of 10 samples from each class to generate the SHAP decision plot presented in **Figure 6**, illustrating the feature influences on the SVM model across different classes. The horizontal axis represents the SHAP values, where positive values denote that a feature propels the prediction towards the positive class, while negative values indicate a propensity towards the negative class. Features are ranked along the vertical axis in descending order of their importance to the model's output. It can be observed that



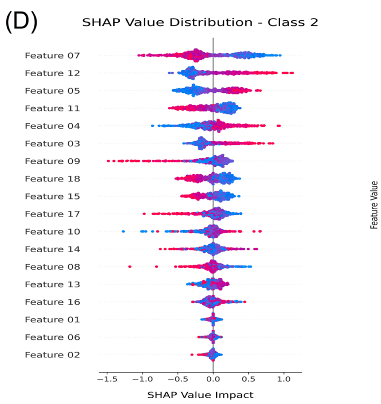
Analysis of the global importance of each feature on the SVM



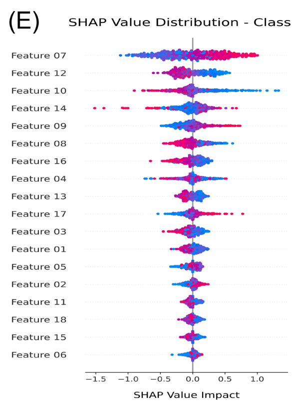
SHAP summary of the effect of each feature on the output of the SVM under class 0



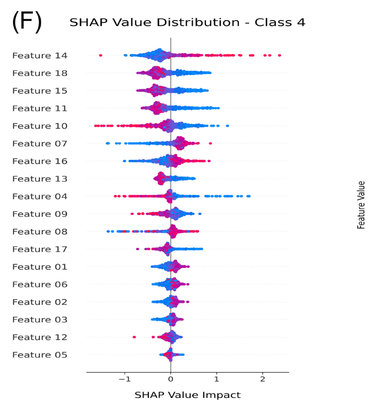
SHAP summary of the effect of each feature on the output of the SVM under class 1



SHAP summary of the effect of each feature on the output of the SVM under class 2



SHAP summary of the effect of each feature on the output of the SVM under class 3



SHAP summary of the effect of each feature on the output of the SVM under class 4

Figure 5. Summary of global feature importance and class-specific feature impact analysis of the support vector machine (SVM) model using SHapley Additive exPlanations (SHAP). (A) Ranking of features by their average SHAP values across all categories; (B-F) Feature's impact plots for classes 0-4, showing each feature's influence on the SVM output. Each point represents an individual sample's SHAP value, with colors indicating feature magnitude.

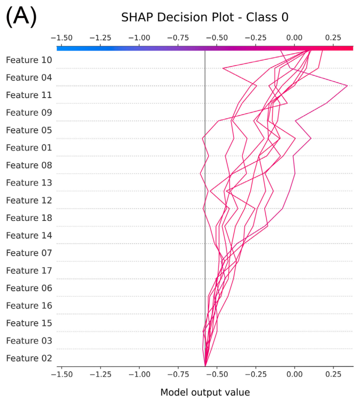
within Class 0, Feature 04 exerts a positive contribution to the prediction outcome. Features in Class 1 exhibited greater variability in their SHAP values, suggesting a potential interdependence between Feature 12 and Feature 04. Feature 12 in Class 2 demonstrated a distinct negative influence on the model's predictive results. In Class 3, Feature 07 and Feature 10 exhibited analogous patterns, with higher values associated with positive contributions and lower values correlated with negative contributions. Finally, elevated values of Feature 14 and Feature 18 in Class 4 were found to positively impact the model's prediction results.

LIME-based local interpretability analysis

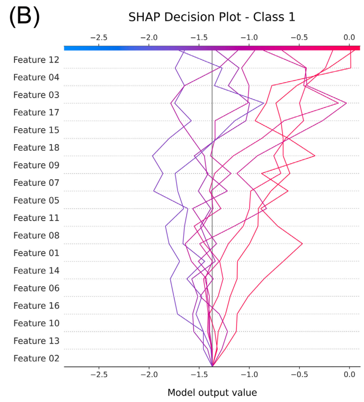
In this study, LIME was used to visualize the feature importance of the SVM model from the perspective of individual prediction samples. As shown in **Figure 7**, features such as Feature 03 and Feature 09 exerted significant positive impacts on positive predictions, indicating a strong positive correlation with the target variable. Feature 16 exerted a notable negative influence, suppressing model predictions. Overall, features with positive impacts predominate.

Discussion

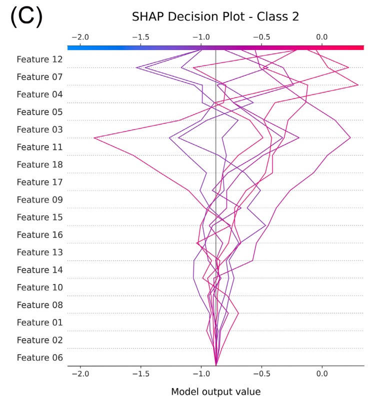
This study utilized a portion of knee X-ray images from the OAI organization. After inversion processing and 2D radiomics feature extraction, 18 key features were selected using variance thresholding and ANOVA. To address data imbalance, SMOTE sampling and class weight balancing were applied. In the model evaluation, eight machine learning models were compared, with the SVM model performing the best.



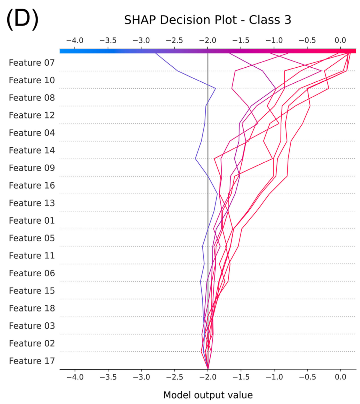
SHAP decision plot of the effect of each feature on the SVM under class 0



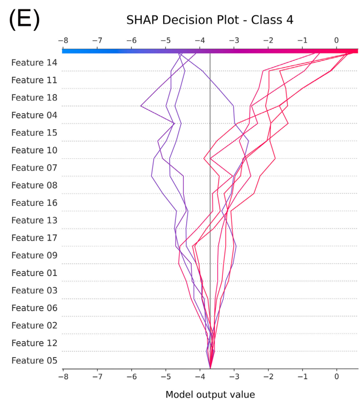
SHAP decision plot of the effect of each feature on the SVM under class 1



SHAP decision plot of the effect of each feature on the SVM under class 2



SHAP decision plot of the effect of each feature on the SVM under class 3



SHAP decision plot of the effect of each feature on the SVM under class 4

Figure 6. SHapley Additive exPlanations (SHAP)-based decision plot illustrating feature influence on the support vector machine (SVM) model. (A-E) Visualization of the decision-making process of the SVM model for Classes 0 to 4. The influence of each feature is represented by line segments, with the distance from the central axis being proportional to the magnitude of its impact on the model output. Feature values are color-coded, with blue indicating lower values and red indicating higher values.

The model's decision-making process was analyzed using SHAP and LIME to identify key features. This work provides a reliable reference for clinical practice, enhancing the model's credibility and practicality.

Prior studies mainly focused on the binary classification of knee OA, aiming only to determine disease presence. For instance, study of G.B. Joseph et al. achieved an AUC of 0.772, and Xu Wang et al. obtained an AUC of 96.5% and an F1 score of 93.5% [18, 19]. Despite technical advancements in these studies, they fell short in model interpretability and lacked in-depth analysis of feature contributions to predictions. This study advances to a multi-classification framework using the KL scale and various ML models, enabling more precise OA severity prediction. Additionally, SHAP and LIME were employed to enhance the interpretability of the models. As shown in **Table 4** and **Figure 5A**, Feature 07, by quantifying image texture complexity, helps the model detect subtle changes in diseased areas [20, 21]. This is crucial for early diagnosis and monitoring disease progression. Feature 12 reflects the uniformity of low-frequency image components, providing valuable information on joint structural integrity and aiding in distinguishing between normal and pathological joints. **Figures 5** and **6** reveal the impact of features on different KL grades, underscoring their role in model predictions. In classes 0 and 1, Feature 04 helps standardize image processing and distinguish pathological from non-pathological low signals by assessing the discreteness of the lowest signal intensity areas in the patellar image, thus improving grading accuracy. In classes 1 and 2, Feature 12 shows high importance with significant fluctuations, aligning with the global feature impor-

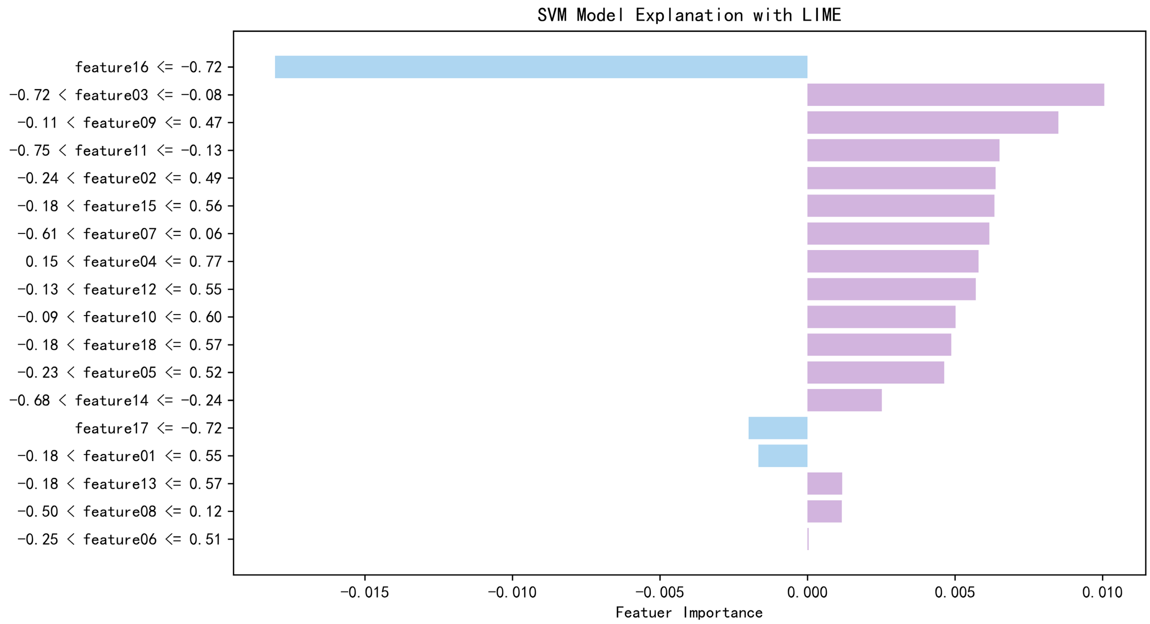


Figure 7. Local feature importance for individual samples interpreted by Local Interpretable Model-agnostic Explanations (LIME). The figure illustrates the influence of each feature on model predictions, ranked by importance. The length of bars represents feature importance, with negative values indicating a negative impact on predictions and positive values indicating a positive impact. The inequality next to the feature name shows the feature value's threshold range.

tance analysis. In class 3, Feature 07 exhibits a clear directional impact on predictions. The varying SHAP values of these features influence model predictions differently, highlighting the transparent and visual nature of the decision-making process. A transparent decision-making process enhances clinicians' trust in model outputs and improves diagnostic precision. LIME further validates the dynamic role of features across different samples, providing clinicians with a clear basis for decision-making.

Nevertheless, this study still has several limitations. X-ray image data were exclusively sourced from the OAI database. The absence of external data from other institutions or imaging modalities might restrict the model's generalizability across diverse clinical scenarios. Additionally, the dataset suffered from significant class imbalance. Future work will incorporate multi-modal data from multiple institutions to evaluate the model's generalizability and combine radiomics with deep learning features to enhance sensitivity in early diagnosis.

Conclusion

This study investigated the five-class classification of knee osteoarthritis using multiple machine learning models and demonstrated that SVM exceled in classification accuracy. The model's decision-making process was further elucidated through SHAP and LIME analyses, providing valuable insights into how key features influenced predictions.

Author contributions: He Ren, Yina Zhang, Yutong Xie contributed to the conception of the study. Anqi Wu, Xianglun Kong, Chenxiao Bai performed the experiment and contributed significantly to analysis and manuscript preparation. Miao Yu, Yimeng Wang helped perform the analysis with constructive discussions. Ping Li oversaw the project, offered key revisions to the manuscript, and gave approval for the final version.

References

[1] Hall M, van der Esch M, Hinman RS, et al. How does hip osteoarthritis differ from knee osteoarthritis? Osteoarthritis Cartilage 2022;30(1):32-41.
[2] Gelber AC. Knee Osteoarthritis. Ann Intern Med 2024;177(9):ITC129-ITC144.
[3] Kloppenburg M, Namane M, Cicuttini F. Osteoarthritis. Lancet 2025;405(10472):71-85.

[4] Di Matteo A, Bathon JM, Emery P. Rheumatoid arthritis. Lancet 2023;402(10416):2019-2033.
[5] Deng P, Wang S, Sun X, et al. Global Trends in Research of Gouty Arthritis Over Past Decade: A Bibliometric Analysis. Front Immunol 2022;13:910400.
[6] Park SY, Wooden TK, Pekas EJ, et al. Effects of passive and active leg movements to interrupt sitting in mild hypercapnia on cardiovascular function in healthy adults. J Appl Physiol (1985) 2022;132(3):874-887.
[7] Harden JE, David Branch J, Reynolds LJ. Impacts of Physical Inactivity Models on Endothelial Function: A Systematic Review. Sports Med 2025;55(8):1937-1952.
[8] Kaya AT, Akman B. Relationship of the Novel Scoring System for Lower Extremity Venous Thrombosis with Pulmonary Embolism. Acad Radiol 2024;31(9):3811-3824.
[9] Baranga L, Khanuja S, Scott JA, et al. In Situ Pulmonary Arterial Thrombosis: Literature Review and Clinical Significance of a Distinct Entity. AJR Am J Roentgenol 2023;221(1):57-68.
[10] Kahn Susan R, de Wit K. Pulmonary Embolism. N Engl J Med 2022;387(1):45-57.
[11] Ou J, Zhang J, Alswadeh M, et al. Advancing osteoarthritis research: the role of AI in clinical, imaging and omics fields. Bone Res 2025;13(1):48.
[12] Li X, Roemer FW, Cicuttini F, et al. Early knee OA definition-what do we know at this stage? An imaging perspective. Ther Adv Musculoskelet Dis 2023;15:1759720X231158204.
[13] Ehmig J, Engel G, Lotz J, et al. MR-Imaging in Osteoarthritis: Current Standard of Practice and Future Outlook. Diagnostics (Basel) 2023; 13(15):2586.
[14] Li D, Lu H, Wu J, et al. Development of machine learning models for predicting depressive symptoms in knee osteoarthritis patients. Sci Rep 2024;14(1):28603.
[15] Pabinger C, Lothaller H, Kobinia GS. Intra-articular injection of bone marrow aspirate concentrate (mesenchymal stem cells) in KL grade III and IV knee osteoarthritis: 4 year results of 37 knees. Sci Rep 2024;14(1):2665.
[16] Farooq MU, Ullah Z, Khan A, et al. DC-AAE: Dual channel adversarial autoencoder with multi-task learning for KL-grade classification in knee radiographs. Comput Biol Med 2023; 167:107570.
[17] Li G, Li S, Xie J, et al. Identifying changes in dynamic plantar pressure associated with radiological knee osteoarthritis based on machine learning and wearable devices. J Neuroeng Rehabil 2024;21(1):45.
[18] Joseph GB, McCulloch CE, Nevitt MC, et al. Machine learning to predict incident radiographic knee osteoarthritis over 8 Years using combined MR imaging features, demographics, and clinical factors: data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2022;30(2):270-279.
[19] Wang X, Wang T, Su Z. Attention-enhanced deep learning and machine learning framework for knee osteoarthritis severity detection in football players using X-ray images. Journal of Radiation Research and Applied Sciences 2025;18(2):101428.
[20] Spotorno N, Strandberg O, Stomrud E, et al. Diffusion MRI tracks cortical microstructural changes during the early stages of Alzheimer's disease. Brain 2024;147(3):961-969.
[21] Kelly BS, Mathur P, McGuinness G, et al. A Radiomic "Warning Sign" of Progression on Brain MRI in Individuals with MS. AJNR Am J Neuroradiol 2024;45(2):236-243.