Research article Pham Ngoc Thach Medical and Pharmaceutical Journal

DOI: 10.59715/pntjmp.5.2.11

Machine learning prediction of severe knee osteoarthritis from routine data: performance, interpretability, and decision curve analysis

Nguyen Trong Hien[1], Nguyen Minh Tuan[2], To Nam Kien[3]

[1]Department of Biostatistics and Informatics, Faculty of Public Health, Pham Ngoc Thach University of Medicine, Ho Chi Minh city.
[2]Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Ho Chi Minh city. (minhtuan@ptit.edu.vn)
[3]Department of Rheumatology, Thong Nhat Hospital, Ho Chi Minh city.

Abstract

Background: Early identification of severe knee osteoarthritis (KOA) usually relies on X-ray imaging, but this is not always possible in healthcare facilities with limited resources. Estimating KOA severity from routine clinical and biochemical data may help clinicians decide when to request imaging and when to refer patients.

Methods: A retrospective study was conducted on 189 patients with KOA whose severity was classified according to the Kellgren–Lawrence scale. Four machine learning models (i.e., logistic regression, Random Forest, XGBoost and SVM) were developed based on 15 routine variables. SHAP analysis was used to interpret the model and to select a reduced set of 10 variables. Performance was evaluated using five-fold stratified cross-validation and an independent test set of 20%. Decision curve analysis (DCA) was used to evaluate clinical benefit.

Results: The models showed moderate discriminative ability in cross validation (AUPRC 0.44 to 0.52; ROC AUC 0.54 to 0.62). The SHAP optimised logistic regression model achieved the best performance on the test set (AUPRC 0.70; ROC AUC 0.81). Important variables included serum uric acid, BMI, age, and renal function. Decision curve analysis showed a positive net benefit across clinically relevant risk thresholds.

Conclusion: A simple and interpretable logistic regression model based on routine data may help predict severe KOA and prioritise X-ray indication. Further validation studies with large multicenter samples are needed.

Keywords: knee osteoarthritis; Kellgren–Lawrence; machine learning; precision–recall curve; SHAP; decision curve analysis; serum uric acid.

Received: 26/11/2025
Revised: 10/12/2025
Accepted: 20/04/2026

Author contact:
Nguyen Trong Hien
Email: hiennt@pnt.edu.vn
Phone: +84 939205330

1. INTRODUCTION

Knee osteoarthritis (KOA) is one of the leading causes of pain, disability and movement limitation in older adults. Its prevalence increases rapidly with age and imposes a considerable economic burden on both patients and healthcare systems [1, 2]. The extent of joint damage is usually assessed on knee X-ray images using the Kellgren–Lawrence (KL) grading scale [3]. However, KL grading mainly reflects moderate and severe structural damage and is less sensitive to early changes in cartilage and synovium [4, 5]. KL grading shows only a limited concordance between radiographic assessment and patient reported symptoms [6–8]. This has encouraged the incorporation of clinical and biochemical indices alongside X-ray imaging to achieve a better description of KOA. Serum uric acid (SUA) is a marker of particular interest in this context.

Experimental data show that elevated SUA and monosodium urate crystals can activate the NLRP3 inflammasome and increase the production of interleukin 1β. This promotes synovial inflammation and cartilage damage [9, 10]. Several observational studies have shown that individuals with higher SUA levels often have more severe radiographic knee osteoarthritis, even in populations without gout [11–14]. These data support the concept of metabolic osteoarthritis. This concept proposes that systemic metabolic factors act together with mechanical loading at the joint to increase the extent of joint damage.

In recent years, machine learning (ML) and artificial intelligence (AI) have advanced rapidly, offering new ways to predict the risk, progression and severity of knee osteoarthritis. A systematic review of KOA prediction models showed that ML algorithms such as Random Forest and Support Vector Machine often perform better than traditional statistical models [1]. Deep learning models that use imaging or gait data can also classify the disease with very high accuracy [2, 15, 16].

However, the need for radiographic images or dedicated equipment makes these models difficult to apply widely and hard to interpret at the level of individual patients. This has created a demand for simple, low cost and interpretable prediction models. These models rely mainly on routine information such as demographic characteristics, comorbidities, symptoms and standard biochemical tests [16]. Such models can estimate KOA severity at the primary care level or in settings with limited access to imaging. At the same time, they provide insight into the metabolic and mechanical factors that contribute to severe radiographic damage significantly.

In practice, collected data contain fewer severe KOA cases than mild cases. Therefore, the precision–recall curve and the area under this curve (AUPRC) reflect the ability of the model to identify the rare class (severe KOA) better than the ROC curve [17]. In addition, decision curve analysis (DCA) links predicted risk with net clinical benefit across different decision thresholds.

The interpretability of predictor variables is also crucial for the model to be accepted in clinical practice. SHapley Additive exPlanations (SHAP) provide a unified and model agnostic analytical framework. SHAP allows each prediction to be decomposed into the sum of contributions from each variable, yielding both global importance rankings and explanations at the level of individual patients [18]. Recent studies have compared SHAP based rankings with traditional importance measures and have highlighted the value of SHAP for variable selection and model simplification [19]. When combined with decision tree models and other ML algorithms, SHAP can turn black box models into tools with clear clinical meaning [18].

Building on our previous study of the association between SUA and KOA severity on radiographs [20], we developed and evaluated machine learning models to predict severe KOA (KL 3–4) from routine clinical and biochemical data, with SUA as the central variable. We compared performance using AUPRC among four ML algorithms, namely logistic regression, Random Forest, XGBoost and SVM. Next, we used SHAP to obtain global and local explanations and to propose an optimal reduced set of variables based on SHAP [18, 19]. Finally, we applied decision curve analysis to assess the clinical application potential of the best performing model.

2. SUBJECTS AND METHODS

2.1. Study design

We carried out a cross-sectional study in patients with KOA identified on X-ray at Nguyen Tri Phuong Hospital. Clinical data, routine laboratory results, knee radiographs and KL grading were extracted from the hospital information system and the imaging archive for the period from December 2021 to July 2022.

2.2. Study participants and sample size

Patients aged 18 years or older were diagnosed with KOA by rheumatologists or orthopaedic surgeons and had at least one knee X-ray of sufficient quality for KL grading. For each patient, an index knee with KL grade from 2 to 4 was selected according to a predefined rule, usually the more painful or more severely damaged side. We excluded patients (1) with other joint diseases or secondary causes of osteoarthritis such as inflammatory arthritis, gout, septic arthritis, deformity after major trauma, prior knee replacement or corrective surgery, or congenital malformations; (2) with acute conditions or taking medications that could markedly alter the biochemical indices of interest; (3) with incomplete records or unreadable X-ray images, missing KL grading, or missing essential clinical or laboratory variables.

After applying the exclusion criteria, we recorded data from 189 patients, including 115 with mild KOA (KL 2) and 74 with severe KOA (KL 3 to 4).

2.3. Study variables

The primary outcome variable was the severity of KOA on radiographs of the index knee, dichotomised into KL 2 (coded as 0, mild KOA) and KL 3–4 (coded as 1, severe KOA). KL grading was performed on standard anteroposterior and lateral knee radiographs by experienced physicians who were blinded to laboratory results, according to published criteria [3].

The main predictive variables were selected based on the hospital medical records and previous studies on knee osteoarthritis (KOA) [4, 5, 10–12, 14, 16]. The demographic variables included age (years), sex, and body mass index (BMI, kg/m^2). The biochemical variables included serum uric acid (SUA) levels and renal function. Serum uric acid testing was performed using an AU680 analyzer. In the dataset, this variable was coded as *uric2*. Serum creatinine and estimated glomerular filtration rate (eGFR) were recorded from routine biochemical tests. eGFR was used as the primary indicator representing renal function. Lipid parameters included total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides. The comorbidity group included hypertension, diabetes mellitus, dyslipidemia, metabolic syndrome, and osteoporosis. Knee pain (*oakn1*) was defined based on the presence of typical pain symptoms related to knee osteoarthritis. Prolonged standing at work (*habit3*) was defined as standing for ≥4 hours per day. Frequent heavy lifting (*habit1*) was defined as lifting objects weighing ≥10 kg multiple times per day. Information on intra-articular interventions was coded into binary variables, including a history of corticosteroid injections (*cor.tiem*) and a history of hyaluronic acid injections (*hyal*) into the knee joint.

2.4. Data processing

Continuous variables with a low proportion of missing values (below about 5–10%) were imputed with the median; binary or categorical variables were imputed with the most frequent value. Variables with a high proportion of missing values (above about 20%) were excluded from the analysis, in line with recommendations for handling missing data in medical research [21, 22].

Continuous variables were standardised using the Z score. Categorical variables were coded as binary indicators with clinically meaningful reference groups. All processing was performed in Python in Google Colab using the scikit learn library [23] and the XGBoost library [24].

2.5. Variable selection

All 15 routinely collected variables were initially included in a baseline Random Forest model. SHAP values for the severe KOA class were then computed to identify the predictors contributing most to model discrimination [18, 25]. Following the SHAP importance ranking and incorporating clinical reasoning and reviewer recommendations, a refined set of ten variables was selected. Priority was given to demographic factors, SUA, renal function, and occupational mechanical load on the knee, whereas creatinine and detailed lipid parameters were excluded to avoid redundancy and limited clinical justification in prediction. The final predictor set comprised age, sex, BMI, SUA, eGFR, occupational category (ocupat), heavy lifting (habit1), prolonged standing (habit3), prior corticosteroid injection (cor.tiem), and prior hyaluronic acid injection (hyal). All four machine learning algorithms were subsequently retrained using these ten variables.

2.6. Machine learning models

2.6.1. Training scheme

We formulated KOA severity prediction as a supervised binary classification task ($y_i \in \{0,1\}$). Models estimated the probability $\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i)$ from features $\mathbf{x}_i$. The training scheme of the models is illustrated in the figure 1.

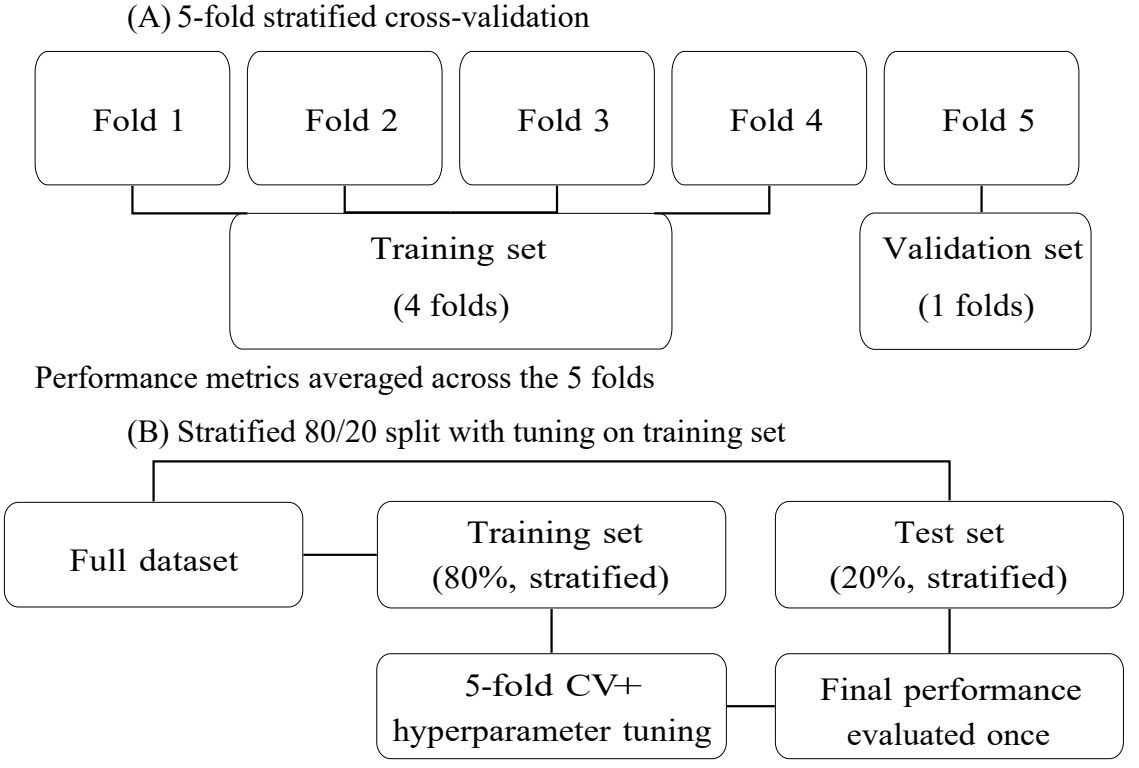


Fig. 1. Evaluation schemes: (A) 5-fold stratified cross-validation with averaged performance; (B) stratified 80/20 train–test split with cross-validation and hyperparameter tuning on the training set, and final evaluation on the held-out test set.

2.6.2. Algorithms and hyperparameter tuning

We compared four classification algorithms commonly used in medical prediction [1, 16, 24, 25], namely: (1) logistic regression (LR) with L_2 regularisation [26]; (2) Random Forest (RF), an ensemble of decision trees trained on bootstrap samples with random selection of a subset of variables at each node [25]; (3) XGBoost (XGB), a gradient boosting algorithm with a logistic loss function [24]; (4) Support Vector Machine (SVM) with a radial basis function (RBF) kernel and probability calibration.

Key hyperparameters were tuned using grid search or random search within the cross-validation framework. For LR, the strength of regularisation was scanned over a prespecified range of values. For RF, we varied the number of trees, the maximum depth and the minimum number of samples required for a split. For XGBoost, parameters such as the number of boosting rounds, learning rate, tree depth, subsampling rate and regularisation terms were tuned. For SVM, the penalty parameter C and the kernel width γ were varied. The search ranges were chosen to avoid models that were overly complex relative to the sample size.

2.7. Model evaluation

Because of the imbalance between classes (115 mild versus 74 severe), we used the AUPRC for the severe KOA class as the main performance metric [17].

We also assessed other metrics, including ROC AUC, positive predictive value, sensitivity and F1 score for the severe class at several decision thresholds. For the LR model, a sensitivity analysis was performed by varying the probability threshold from 0.20 to 0.50 and recalculating accuracy, sensitivity, specificity and positive predictive value based on cross validation predictions in the full sample.

2.8. SHAP analysis

To examine the interpretability of the model, we applied SHAP to the RF model trained on the full set of 15 variables. SHAP values were computed for all patients and all variables. The mean SHAP value (in absolute terms) was used to rank global importance. The SHAP summary plot was used to present both the ranking and the distribution of the effect of each variable. SHAP dependence plots for several important variables (SUA, BMI, age and eGFR) were used to explore non linear relationships and potential interactions. For several typical severe and mild cases, SHAP force plots illustrated how each variable pushed the predicted risk towards or away from the severe KOA group. The SHAP based ranking also provided the basis for constructing the optimal set of ten variables used in the reduced models [18, 19].

2.9. Subgroup and robustness analyses

We assessed the stability of model performance in different patient subgroups. Using predicted probabilities from cross validation of the four main algorithms (trained on ten variables), AUPRC and ROC AUC were recalculated in strata defined by age (<65 and ≥65 years), BMI (<25 and ≥25 kg/m^2), sex (male/female) and presence or absence of metabolic syndromes. These analyses were descriptive and aimed to assess whether discriminative ability was consistent across subgroups.

2.10. Decision curve analysis

To assess the clinical relevance of the models, we performed decision curve analysis on the test set. For each model and each probability threshold p_t within a clinically reasonable range, the net benefit was calculated as

$$Net\ benefit = \frac{TP}{N} - \frac{FP}{N} \times \frac{P_t}{1 - p_t}$$

where TP and FP are the numbers of true positive and false positive cases, respectively, and *N* is the total number of patients in the test set. The net benefit curves of the four models were compared with two default strategies: no imaging or intervention for any patient ("treat none") and imaging or intervention for all patients ("treat all"). We identified ranges of probability thresholds at which model-based decisions provided a higher net benefit.

3. RESULTS

3.1. Baseline characteristics

Among the 189 patients with radiographic knee osteoarthritis included in the analysis, 115 (60.8%) had mild disease (KL grade 2) and 74 (39.2%) had severe disease (KL grades 3–4). The overall mean age was 64.0 ± 9.0 years, the mean body mass index (BMI) was 24.7 ± 3.6 kg/m^2, and the mean serum uric acid (SUA) level was 5.8 ± 1.6 mg/dL. Approximately three quarters of the cohort were women.

Baseline characteristics by KOA severity are summarised in Table 1. Compared with patients with severe KOA, those with mild KOA tended to have a higher BMI (25.35 ± 3.60 vs 24.30 ± 3.59 kg/m^2, p = 0.051), whereas the difference in mean age between the two groups was not statistically significant (62.72 ± 7.86 vs 64.83 ± 9.55 years, p = 0.100). Estimated glomerular filtration rate (eGFR) was modestly higher in the group than in the mild group (78.80 ± 12.49 vs 73.25 ± 16.13 mL/min/1.73m^2, p = 0.009). SUA levels were significantly lower in patients with mild KOA compared with those with severe disease (5.32 ± 1.25 vs 6.09 ± 1.71 mg/dL, p < 0.001).

The prevalence of major comorbidities was broadly similar between the two groups. Hypertension was present in 58.3% of patients with mild KOA and 51.4% of those with severe KOA (p = 0.434). The proportions of patients with diabetes mellitus, dyslipidaemia, metabolic syndrome and osteoporosis did not differ significantly between groups (all p > 0.05). Likewise, the distribution of sex, occupational or mechanical load–related factors (frequent heavy lifting and prolonged standing at work) and prior intra-articular corticosteroid or hyaluronic acid injections was comparable between patients with mild and severe KOA (all p > 0.05).

Table 1. Baseline characteristics of patients according to KOA severity (KL 2 vs KL 3–4).

Variable	Mild KOA (KL 2)	Severe KOA (KL 3–4)	p-value
Age (years)	62.72 ± 7.86	64.83 ± 9.55	0.100
Body mass index (kg/m^2)	25.35 ± 3.60	24.30 ± 3.59	0.051
Serum uric acid (mg/dL)	5.32 ± 1.25	6.09 ± 1.71	< 0.001
eGFR (mL/min/1.73m^2)	78.80 ± 12.49	73.25 ± 16.13	0.009
Hypertension	67/115 (58.3%)	38/74 (51.4%)	0.434
Diabetes mellitus	24/115 (20.9%)	18/74 (24.3%)	0.705
Dyslipidaemia	62/115 (53.9%)	42/74 (56.8%)	0.815
Metabolic syndrome	55/115 (47.8%)	32/74 (43.2%)	0.640
Osteoporosis	20/115 (17.4%)	14/74 (18.9%)	0.942
Female sex	89/115 (77.4%)	56/74 (75.7%)	0.923
Frequent heavy lifting	37/115 (32.2%)	20/74 (27.0%)	0.555
Prolonged standing at work	37/115 (32.2%)	16/74 (21.6%)	0.158
Prior intra-articular corticosteroid injection	9/115 (7.8%)	7/74 (9.5%)	0.900
Prior intra-articular hyaluronic acid injection	10/115 (8.7%)	4/74 (5.4%)	0.571

Data are presented as mean ± SD or n/N (%). p-values are from Welch's two-sample t-test for continuous variables and chi-square test (or Fisher's exact test when appropriate) for categorical variables.

3.2. Cross-validation performance

Across models, the mean AUPRC for the severe class ranged from 0.44 to 0.52, while the ROC AUC ranged from 0.54 to 0.62 (Table 2). This reflects moderate discriminative ability in a small and imbalanced data set.

Among the algorithms, LR showed the best balance between discrimination and classification performance, with an AUPRC of 0.524 (95% CI 0.456–0.592) and a ROC AUC of 0.619 (95% CI 0.565–0.672). RF and XGB yielded similar cross validation results, with AUPRC around 0.47 and ROC AUC around 0.58. The SVM model had the lowest AUPRC (0.437; 95% CI 0.350–0.524) and ROC AUC (0.542; 95% CI 0.423–0.661).

To formally compare performance between models across folds, we used the Friedman test and pairwise Wilcoxon signed rank tests with Holm correction, based on AUPRC and ROC AUC in each fold. The results showed no statistically significant differences between models ($p > 0.05$).

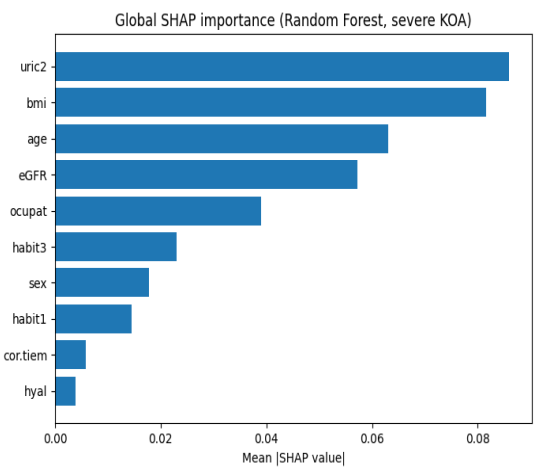
In the test set, the models showed discrimination similar to that seen in cross validation. With the SHAP based set of ten variables, LR achieved the best overall performance with an AUPRC of 0.70, a ROC AUC of 0.81, an F1 score of 0.76 and an accuracy of 0.82 for the severe KOA class (KL 3–4). XGBoost also showed acceptable performance (AUPRC 0.63, ROC AUC 0.70), whereas Random Forest had more modest discriminative ability (AUPRC 0.52, ROC AUC 0.67). In contrast, the SVM model performed poorly on the test set and did not correctly identify any severe cases at a probability threshold of 0.5, despite an overall accuracy of 0.61.

Table 2. Cross validation performance of machine learning models for predicting severe KOA (KL 3–4) from routine clinical and biochemical variables. Values are means and 95% confidence intervals (CI) across five-fold stratified cross validation.

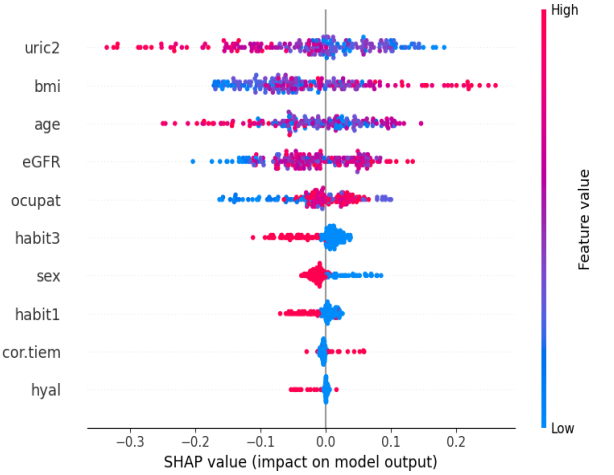
Model	AUPRC	ROC-AUC	Precision	Recall	F1-score	Accuracy
LR	0.524 (0.456–0.592)	0.619 (0.565–0.672)	0.467 (0.388–0.546)	0.567 (0.457–0.676)	0.510 (0.421–0.600)	0.577 (0.501–0.653)
RF	0.475 (0.407–0.543)	0.576 (0.529–0.624)	0.427 (0.339–0.515)	0.324 (0.262–0.385)	0.362 (0.309–0.415)	0.555 (0.505–0.606)
XGB	0.474 (0.405–0.543)	0.579 (0.535–0.623)	0.445 (0.347–0.543)	0.378 (0.282–0.474)	0.406 (0.310–0.501)	0.571 (0.505–0.637)
SVM	0.437 (0.350–0.524)	0.542 (0.423–0.661)	0.067 (0.000–0.197)	0.014 (0.000–0.042)	0.024 (0.000–0.070)	0.603 (0.599–0.607)

3.3. Global SHAP analysis and SHAP optimised model

We trained an RF model and computed SHAP values for the severe KOA class (KL 3–4). The resulting SHAP tensor contained one value for each patient and each variable. The corresponding global importance ranking is shown in Figure 2.



(a) Mean absolute SHAP values of all pre-impordictor variables



(b) SHAP summary plot for the most important variables.

Fig. 2. Global SHAP importance and distribution of SHAP values for the RF model predicting severe KOA (KL 3–4).

Based on the global SHAP ranking (Figure 2b), we identified an optimal set of ten variables in the following order: uric2, BMI, age, eGFR, ocupat, habit3, sex, habit1, cor.tiem and hyal. The models were retrained using only these ten variables and evaluated with five-fold stratified cross validation, with results presented in Table 3.

Table 3. Comparison of cross-validation performance between the full model and the SHAP optimised model.

Model	AUPRC (full)	AUPRC (SHAP)	ROC AUC (full)	ROC AUC (SHAP)
LR	0.569 (0.527–0.612)	0.533 (0.496–0.569)	0.642 (0.604–0.680)	0.605 (0.585–0.625)
RF	0.491 (0.426–0.556)	0.505 (0.453–0.556)	0.596 (0.534–0.659)	0.619 (0.584–0.654)
XGB	0.489 (0.416–0.562)	0.499 (0.448–0.549)	0.608 (0.556–0.659)	0.604 (0.558–0.649)
SVM	0.447 (0.355–0.539)	0.494 (0.438–0.550)	0.529 (0.385–0.672)	0.562 (0.508–0.615)

After SHAP-based variable reduction, discriminative ability was largely preserved across models. While RF and SVM showed modest improvements, LR and XGB exhibited slightly reduced AUPRC and ROC AUC, reflecting a trade-off between model simplicity and performance.

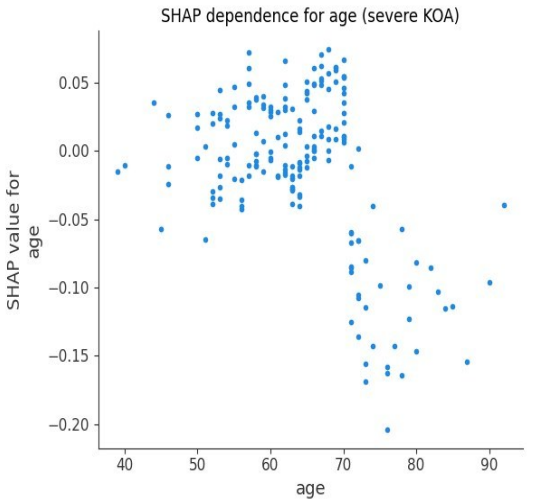
The Friedman test indicated overall differences between the four SHAP optimised models for both AUPRC ($\chi^2(3) = 9.86$, $p = 0.0198$) and ROC AUC ($\chi^2(3) = 15.0$, $p = 0.0018$). However, pairwise Wilcoxon signed rank tests with Holm correction did not detect any statistically significant differences between individual pairs of models (adjusted $p \geq 0{,}05$). Based on Table 3, LR tended to outperform RF and SVM, especially in terms of ROC AUC. The small number of folds limited statistical power, so these differences did not reach statistical significance.

3.4. SHAP dependence plots and case level explanations

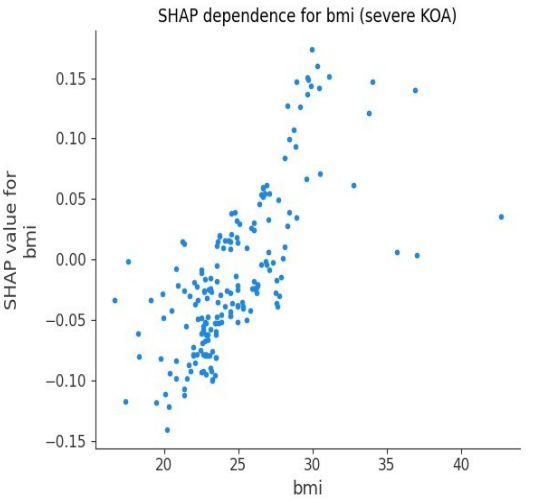
To further describe the nonlinear effects of key variables in the RF model, we computed SHAP values for all 189 patients with the full set of 15 routine clinical and biochemical variables. The SHAP dependence plots for age, BMI, eGFR and serum uric acid (uric2) are shown in Figure 3.

BMI showed a clear increasing trend, with higher SHAP values at higher BMI, consistent with a dose–response relationship between excess body weight and the risk of severe KOA. Age and eGFR exhibited more complex nonlinear patterns, with substantial vertical dispersion, suggesting interactions with other clinical and biochemical variables. Serum uric acid showed a non-linear and heterogeneous association with severe KOA in the RF model. While SUA ranked highly in SHAP importance, higher SUA values were not consistently associated with increased risk across the entire range.

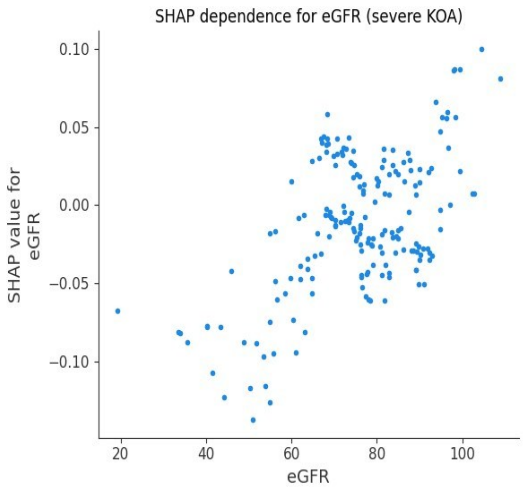
We also generated case level SHAP explanations to illustrate how the model combines predictors for individual patients (Figure 4).



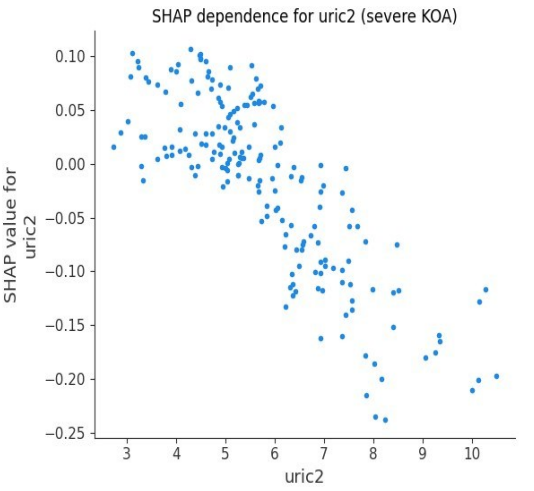
(a) Age



(b) BMI



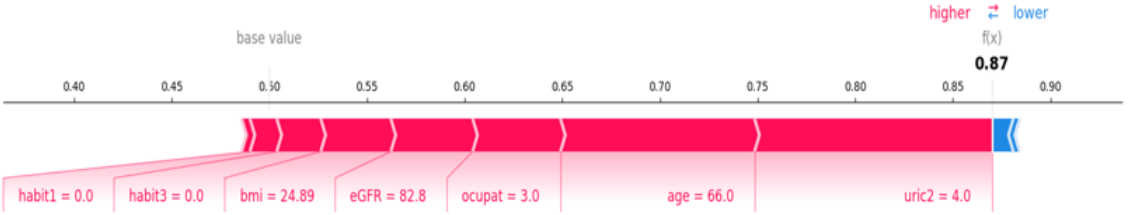
(c) eGFR



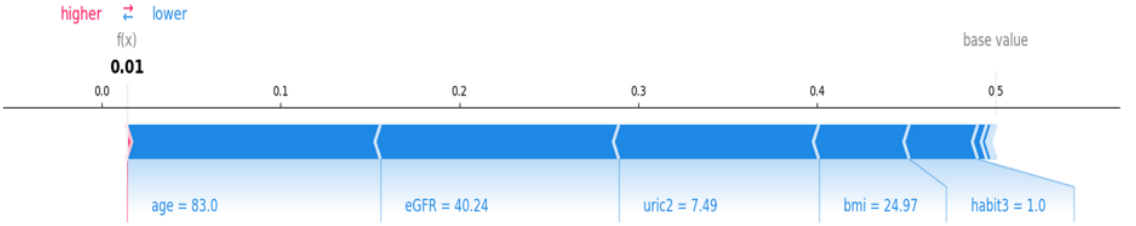
(d) Serum uric acid (uric2)

Fig. 3: SHAP dependence plots for the main predictors of severe KOA (KL 3–4) in the RF model. The horizontal axis shows the observed value of the variable, and the vertical axis shows the corresponding SHAP value for the severe KOA class; positive SHAP values indicate an increased predicted probability of severe KOA.

In the correctly classified patient with the highest predicted probability of severe KOA (p_{severe} = 0,870; Figure 4a), the five strongest positive contributors were elevated serum uric acid, older age, occupational load (ocupat), lower eGFR and higher BMI. Each of these variables had a positive SHAP value, collectively shifting the prediction strongly towards the severe KOA class. In contrast, in the correctly classified mild KOA case with the lowest predicted risk (p_{severe} = 0,010; Figure 4b), the dominant negative contributors were younger age, higher eGFR, lower SUA, lower BMI and limited standing at work (habit3). All had negative SHAP values and pulled the prediction away from the severe class.



(a) Severe KOA case with the highest predicted risk



(b) Mild KOA case with the lowest predicted risk

Fig. 4. Case level SHAP force plots illustrating individual predictions of the RF model. Positive contributions (red) push the prediction towards severe KOA, whereas negative contributions (blue) push the prediction towards mild KOA.

3.5. Decision curve analysis

The results of decision curve analysis for the four models are shown in Figure 5. Selected net benefit values are summarised in Table 4.

Table 4. Decision curves for predicting severe KOA (KL 3–4) on the test set. Net benefit (NB) is reported at four probability thresholds (pt).

Model	NB tại p_t=0.20	p_t=0.30	p_t=0.40	p_t=0.50
LR	0.263	0.192	0.114	0.132
RF	0.296	0.154	0.026	0.026
XGB	0.132	0.090	0.070	0.026
SVM	0.243	0.135	-0.123	0.000
Treat all	0.243	0.135	-0.009	-0.211
Treat none	0.000	0.000	0.000	0.000

Within the range of clinically relevant probability thresholds (approximately 10–30%), all machine learning models yielded higher net benefit than either default strategy of no imaging or treatment for any patient (“treat none”) or imaging or treatment for all patients (“treat all”). In this range, RF and LR achieved the highest net benefit, whereas XGB showed a smaller but still positive improvement over the default strategies. By contrast, SVM performed similarly to the “treat all” strategy at low thresholds and lost net benefit at higher thresholds.

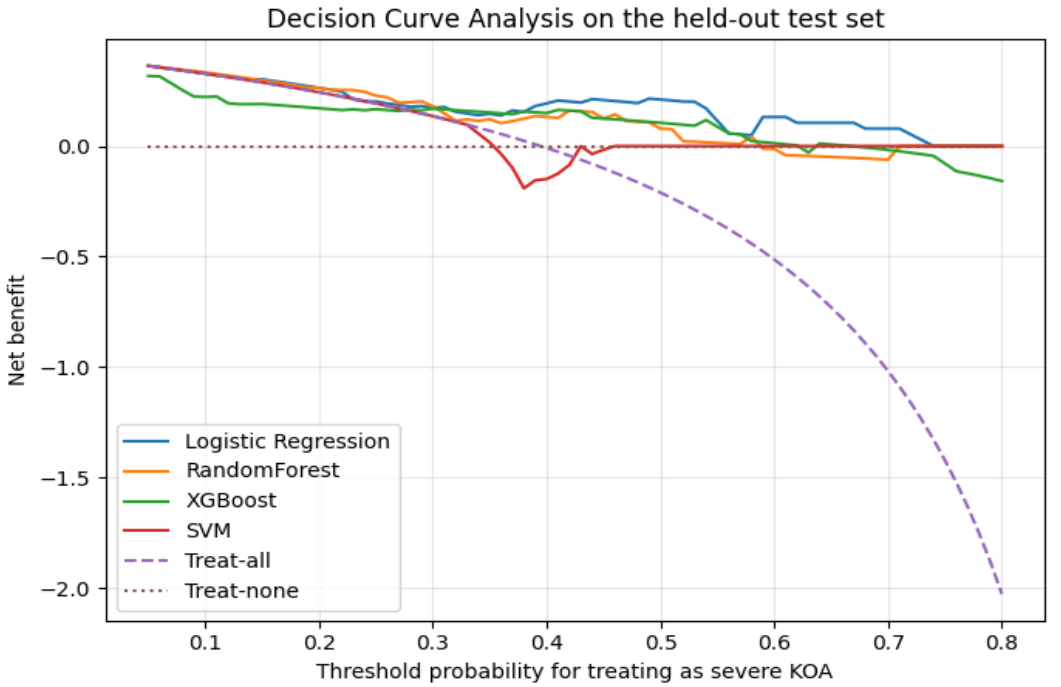


Fig. 5. Decision curves on the test set comparing the net benefit of four machine learning models predicting severe KOA (KL 3–4), together with the two strategies "treat all" and "treat none".

At a probability threshold of 20%, which could be chosen as a decision threshold for ordering additional X-ray imaging or adopting more intensive management, the RF model provided the highest net benefit (0.296), slightly higher than LR (0.263) and the "treat all" strategy (0.243). This difference corresponds to about five additional severe KOA cases correctly identified per 100 patients, without increasing unnecessary interventions in the low-risk group. At higher thresholds (30–50%), LR maintained the most stable positive net benefit, whereas the net benefit of SVM became neutral or negative. Overall, these findings indicate that within this range of risk thresholds, using RF or LR to support decisions would yield greater clinical benefit than imaging or treating all patients or none.

3.6. Subgroup analysis and model robustness

We further assessed model performance in patient subgroups and examined the robustness of LR across different decision thresholds. For each model, predicted probabilities from five-fold stratified cross-validation in the full sample were reused to calculate AUPRC and ROC AUC in subgroups defined by age, BMI, sex and metabolic syndrome (Table 5).

Table 5. Cross-validation performance (AUPRC and ROC AUC) of the four models for predicting severe KOA (KL 3–4) in predefined clinical subgroups.

Subgroup	Model	*N*	Severe *n*	AUPRC	ROC-AUC
Age <65 years	Logistic Regression	100	40	0.562	0.627
	Random Forest	100	40	0.391	0.487
	SVM	100	40	0.457	0.542
	XGBoost	100	40	0.417	0.509
Age ≥65 years	Logistic Regression	89	34	0.549	0.698
	Random Forest	89	34	0.478	0.679
	SVM	89	34	0.402	0.537
	XGBoost	89	34	0.504	0.695
BMI <25 kg/m^2	Logistic Regression	120	41	0.584	0.682
	Random Forest	120	41	0.351	0.531
	SVM	120	41	0.355	0.513
	XGBoost	120	41	0.417	0.596
BMI ≥25 kg/m^2	Logistic Regression	69	33	0.508	0.553
	Random Forest	69	33	0.487	0.567
	SVM	69	33	0.521	0.494
	XGBoost	69	33	0.493	0.554
Female	Logistic Regression	145	56	0.500	0.615
	Random Forest	145	56	0.414	0.581
	SVM	145	56	0.417	0.515
	XGBoost	145	56	0.458	0.610
Male	Logistic Regression	44	18	0.647	0.750
	Random Forest	44	18	0.499	0.622
	SVM	44	18	0.516	0.645
	XGBoost	44	18	0.488	0.643
Metabolic syndrome: no	Logistic Regression	102	42	0.748	0.772
	Random Forest	102	42	0.479	0.599
	SVM	102	42	0.567	0.640
	XGBoost	102	42	0.521	0.646
Metabolic syndrome: yes	Logistic Regression	87	32	0.414	0.527
	Random Forest	87	32	0.383	0.552
	SVM	87	32	0.348	0.430
	XGBoost	87	32	0.403	0.556
Overall	Logistic Regression	189	74	0.541	0.652
	Random Forest	189	74	0.418	0.575
	SVM	189	74	0.425	0.534
	XGBoost	189	74	0.450	0.603

In most subgroups, LR achieved the highest or near highest AUPRC and ROC AUC. In the full sample, LR yielded an AUPRC of 0.541 and a ROC AUC of 0.652, compared with 0.418 and 0.575 for RF, 0.450 and 0.603 for XGB, and 0.425 and 0.534 for SVM. Discrimination was generally similar in the groups aged <65 and ≥65 years, with slightly higher ROC AUC in the older group (0.698 versus 0.627 for LR). Performance was also comparable across BMI strata, although SVM yielded a slightly higher AUPRC than LR in the group with BMI ≥ 25 kg/m^2 (0.521 versus 0.508). Sex specific analysis showed better discrimination in men than in women (ROC AUC 0.750 versus 0.615 for LR). Notably, the models performed best in patients without metabolic syndrome (LR AUPRC 0.748 and ROC AUC 0.772), whereas performance was lower in those with metabolic syndrome (AUPRC 0.414, ROC AUC 0.527).

To assess robustness across classification thresholds, we varied the probability threshold of LR from 0.20 to 0.50 and recalculated accuracy, sensitivity, specificity and positive predictive value (PPV) (Table 6). As the threshold increased from 0.20 to 0.50, sensitivity decreased from 0.959 to 0.649, whereas specificity increased from 0.130 to 0.583. This illustrates the familiar trade-off between detecting more severe cases and reducing false positives. Overall accuracy increased from 0.455 to 0.608 across this range, and PPV rose from 0.415 at a threshold of 0.20 to 0.500 at a threshold of 0.50, indicating that higher thresholds identify severe KOA more precisely but miss a larger proportion of true cases.

Table 6. Robustness of LR performance across decision thresholds in the full sample (cross-validation predictions, N = 189).

Threshold	Accuracy	Sensitivity	Specificity	PPV
0.20	0.455	0.959	0.130	0.415
0.30	0.513	0.878	0.278	0.439
0.40	0.577	0.784	0.443	0.475
0.50	0.608	0.649	0.583	0.500

4. DISCUSSION

4.1. Summary of key findings

In this study, we showed that machine learning models built entirely from routine clinical and biochemical variables can distinguish severe KOA on radiographs (KL 3–4) from mild KOA (KL 2) with moderate performance. The main result is that the LR model based entirely on routine data achieved AUPRC = 0.70 and ROC AUC = 0.81 on the independent test set. Although simple, this model was comparable to or better than more complex decision tree models and SVM.

SHAP analysis at both the global and individual levels identified a compact subset of predictors with large effects (serum uric acid, BMI, age, renal function markers (creatinine/eGFR), lipid profiles and prolonged standing at work). When the model was restricted to these ten variables, discriminative ability in cross-validation was preserved and, in some cases, slightly improved compared with the full 15 variable model. DCA showed that, within a clinically reasonable range of probability thresholds, using the model to guide decisions on imaging or referral could provide a net clinical benefit compared with the two strategies “treat all” and “treat none”.

4.2. Clinical implications

The prominent position of age, renal function markers (eGFR), lipid profiles and prolonged standing at work indices in the SHAP ranking is consistent with the current view that osteoarthritis is a metabolic inflammatory disease rather than merely mechanical wear and tear [4, 5]. Experimental and clinical data have shown an association between urate, inflammasome activation, synovial inflammation and radiographic progression in KOA [9–12, 14]. In this study, serum uric acid was one of the variables with the largest contribution in the multivariable prediction model. consistent in highlighting SUA as a relevant factor, although the direction and shape of the association differed across analytical approaches [20].

4.3. Methodological implications

From a methodological perspective, this study emphasises the importance of using AUPRC as the main metric in classification problems with imbalanced classes, where a single ROC AUC value may be overly optimistic [17]. By deliberately optimising and reporting AUPRC for the severe KOA group, we aligned model evaluation with the clinical objective of identifying a small group of high-risk patients.

The study also shows that SHAP, which was developed as a model explanation tool independent of the underlying algorithm, can be used for feature selection and model simplification. The SHAP based ranking in our study produced a set of ten predictors that was both clinically coherent and statistically efficient, and the SHAP optimised models maintained or slightly improved discriminative ability compared with the full model, in line with recent reports on SHAP based variable selection in KOA and other fields [18, 19]. Incorporating decision curve analysis into the evaluation framework also helps link predictive performance with net clinical benefit, complementing previous KOA prediction studies that have mainly focused on ROC based metrics or image derived features [1, 2, 15, 16].

4.4. Limitations

This study was conducted at a single tertiary hospital, with a modest sample size and a limited number of severe cases, leading to wide confidence intervals and a risk of overestimating performance on the test set. The cross-sectional design does not allow assessment of longitudinal risk, radiographic progression or discordance between symptoms and structural damage [6–8]. Some risk factors for knee osteoarthritis were not fully collected or standardised in the routine medical records, including occupational factors, dietary factors, inflammatory markers, and SNP alleles. The outcome was dichotomised (KL 2 versus KL 3–4) instead of modelling the full spectrum of the KL grading scale [3].

4.5. Future directions

In the future, this ten variable logistic regression model needs to be validated in larger, multicentre and more diverse populations, including community-based populations. At the same time, the analytical framework should be extended to multi class prediction or prediction along the KL grading scale, as well as longitudinal prediction of structural progression [1, 15].

5. CONCLUSION

Based on routine clinical and biochemical data collected at a single tertiary hospital, we built and internally validated several machine learning models to distinguish severe KOA (KL 3–4) from mild KOA (KL 2). In cross - validation, all models achieved moderate discrimination but were clearly better than chance in the setting of class imbalance.

The main result is the ten variable

logistic regression model optimised by SHAP, which achieved an AUPRC of about 0.70 and a ROC AUC of 0.81 for severe KOA on a small test set. This level of performance, together with the transparency of LR and the additional explanations provided by SHAP, suggests that a simple and interpretable model based entirely on routine data can provide an estimate of KOA severity. Decision curve analysis also suggests that using this model to guide imaging or referral decisions may yield a net benefit compared with treating all patients or treating none within a reasonable range of decision thresholds.

These findings support the feasibility of low cost prediction tools with few variables for risk stratification and KOA screening, especially in settings with limited access to imaging or specialists.

REFERENCES

1. Ramazanian, A., Zargaran, M., Moinfar, Z., et al. (2023). Review of KOA prediction models. Journal of Biomedical Informatics, 125, 103976.
2. Miraj, M., Althagafi, A., Khan, M., et al. (2024). AI and machine learning in early diagnosis of KOA: A review. Frontiers in Public Health, 12, 123456.
3. Schiphof, D., de Klerk, B. M., Kerkhof, H. J., et al. (2011). Kellgren–Lawrence criteria and KOA diagnosis. Annals of the Rheumatic Diseases, 70(8), 1422–1427.
4. Goldring, M. B., & Otero, M. (2011). Inflammation in osteoarthritis. Current Opinion in Rheumatology, 23(5), 471–478.
5. Berenbaum, F. (2013). Osteoarthritis as an inflammatory disease (osteoarthritis is not osteoarthrosis!). Osteoarthritis and Cartilage, 21(1), 16–21.
6. Toivanen, A. T., Arokoski, J. P., Manninen, P. S., et al. (2007). Agreement between clinical and radiological KOA diagnosis. Scandinavian Journal of Rheumatology, 36(1), 58–63.
7. Neogi, T., Felson, D., Niu, J., et al. (2009). Radiographic features and pain in KOA. BMJ, 339, b2844.
8. Cubukcu, D., Sarsan, A., & Alkan, H. (2012). Relationships between pain, function and radiographic findings in osteoarthritis of the knee: A cross-sectional study. Arthritis, 2012, 984060.
9. Denoble, A. E., Huffman, K. M., Stabler, T. V., et al. (2011). Uric acid is a danger signal of increasing risk for osteoarthritis through inflammasome activation. Proceedings of the National Academy of Sciences, 108(5), 2088–2093.
10. Neogi, T., Krasnokutsky, S., & Pillinger, M. H. (2019). Urate and OA: Reciprocal links. Joint Bone Spine, 86(5), 576–582.
11. Ding, X., Zeng, C., Wei, J., et al. (2016). Serum uric acid and hyperuricemia in knee osteoarthritis. Rheumatology International, 36(4), 567–573.
12. Krasnokutsky, S., Oshinsky, C., Attur, M., et al. (2017). Serum urate and joint space narrowing in non-gout knee osteoarthritis. Arthritis & Rheumatology, 69(6), 1213–1220.
13. Srivastava, S. R., Srivastava, R., Sharma, A. C., et al. (2018). SUA influence on KOA: Revisit reference range? Osteoarthritis and Cartilage, 26(Suppl.), S224– S225.
14. Bassiouni, S., Aly, S., Abdelrahman, A., et al. (2021). Association of serum uric acid with radiological severity in knee osteoarthritis. Egyptian Rheumatology and Rehabilitation, 48(1), 6.
15. Tiulpin, A., Thevenot, J., Rahtu, E., et al. (2019). Multimodal ML for KOA

progression. Scientific Reports, 9, 1727.
16. Chen, Y., Zhang, J., Liu, X., et al. (2025). Machine learning-based prediction of knee osteoarthritis using clinical and biochemical data. Scientific Reports, 15(2), 2101.
17. Saito, T., & Rehmsmeier, M. (2015). Precision–recall versus ROC analysis in imbalanced datasets. PLOS ONE, 10(3), e0118432.
18. Fan, Z., Song, W., Ke, Y., et al. (2024). XGBoost–SHAP-based interpretable diagnostic framework for knee osteoarthritis: A population-based retrospective cohort study. Arthritis Research & Therapy, 26(1), 213.
19. Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. Journal of Big Data, 11, 44.
20. Kien, T. N. (2022). Association between serum uric acid levels and the severity of knee osteoarthritis at Nguyen Tri Phuong Hospital. Residency thesis, Pham Ngoc Thach University of Medicine.
21. Dong, Y., & Peng, C. J. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 222.
22. Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–A practical guide with flowcharts. BMC Medical Research Methodology, 17(1), 162.
23. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
24. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 785–794).
25. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
26. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.