Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identification of Knee Osteoarthritis–Related Markers and Development of an Interpretable Random Forest Model Using SHAP

Songyao Qi
Guangzhou University of Chinese Medicine
Zhuoyan Chen
Guangzhou University of Chinese Medicine
Ziwen Xia
Guangzhou University of Chinese Medicine
Zilin Chen
Guangzhou University of Chinese Medicine
Zixuan Chen
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine
Feng Lin
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine
Jing Su
sujing2013@gzucm.edu.cn

The Third Affiliated Hospital of Guangzhou University of Chinese Medicine

Research Article

Keywords: knee osteoarthritis, machine learning, random forest, Shapley Additive Explanations, risk stratification, predictive markers, optimal cut-off value

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11321236/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Identification of Knee Osteoarthritis–Related Markers and Development of an Interpretable Random Forest Model Using SHAP

Songyao Qi[2#]，Zhuoyan Chen[2#]，Ziwen Xia[2]，Zilin Chen[2]，

Zixuan Chen[1]，Feng Lin[1]，Jing Su[1*]

1 The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, 510378, China;

2 Guangzhou University of Chinese Medicine, Guangzhou, 510006, China;

#：These authors contributed equally to this work and share first authorship，

*Corresponding author:

Jing Su,The Third Affiliated Hospital of Guangzhou University of Chinese Medicine,Guangzhou, China,Email:sujing2013@gzucm.edu.cn

Abstract

Background

Knee osteoarthritis (KOA) is a common chronic degenerative joint disorder associated with substantial impairment of quality of life and increasing public health burden. Early identification of individuals at elevated risk may facilitate timely risk stratification and intervention. This study aimed to identify key KOA-related markers using machine-learning approaches, interpret model predictions using Shapley Additive Explanations (SHAP), and determine clinically relevant cut-off values for early screening.

Methods

This cross-sectional study included 1,243 participants. Candidate features were selected using three complementary approaches: least absolute shrinkage and selection operator (LASSO) regression, extreme gradient boosting (XGBoost), and Boruta. Nine machine-learning algorithms were developed and compared, including logistic regression, XGBoost, random forest, LightGBM, and support vector machine. The optimal model was interpreted using SHAP to quantify feature contributions. Model performance was evaluated using multiple metrics, including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, F1 score, and Cohen’s kappa. Univariable ROC analyses were performed to estimate optimal cut-off values for candidate markers.

Results

Random forest achieved the best overall classification performance among the nine models. SHAP analysis identified 11 informative features for KOA prediction: age,

EO%, uric acid (UA), MONO%, hemoglobin (HGB), CO_2, MCHC, LDL, TG, Urea/Crea, and CHOL/TG, with age showing the greatest contribution. The model achieved an AUC of 0.988 (95% CI, 0.979 – 0.998), accuracy of 0.941, sensitivity of 0.934, and specificity of 0.963 in the internal validation set. In the temporal validation cohort, performance remained strong, with an AUC of 0.972, accuracy of 0.910, sensitivity of 0.900, and specificity of 0.920. Age demonstrated the highest individual discrimination, with an optimal cut-off value of 51 years.
Conclusions
n interpretable random forest model based on routinely available demographic and laboratory variables showed strong discrimination for KOA classification. The SHAP-identified features, particularly age, may provide a multidimensional basis for KOA risk stratification. Further prospective multicenter studies are required to validate its clinical applicability.

Keywords: knee osteoarthritis; machine learning; random forest; Shapley Additive Explanations; risk stratification; predictive markers; optimal cut-off value

1. Introduction

Knee osteoarthritis (KOA) is one of the most common chronic joint disorders in adults and is associated with persistent pain, restricted mobility, functional impairment, and reduced quality of life. [1]According to the latest analysis based on the Global Burden of Disease Study 2021 (GBD 2021), approximately 375 million people worldwide were living with KOA in 2021, with an estimated 30.85 million incident cases. [2]As population ageing accelerates and major risk factors such as obesity remain highly prevalent, the disease and socioeconomic burden of KOA is expected to increase further, posing a growing public health challenge worldwide .[3]

KOA is now recognized not simply as a consequence of mechanical “wear and tear,” but as a whole-joint disorder involving the articular cartilage, subchondral bone, synovium, menisci, ligaments, and periarticular soft tissues. [4, 5]Its onset and progression are influenced by multiple factors, including ageing, obesity, previous joint injury, abnormal mechanical loading, and genetic susceptibility. Metabolic dysregulation and chronic low-grade inflammation are also increasingly recognized as important systemic features of KOA. [6]Emerging evidence suggests that circulating inflammatory mediators, metabolic indices, and other peripheral blood markers may capture aspects of the inflammatory – metabolic phenotype of KOA, providing readily measurable and objective information for risk stratification.[7, 8]
Although conventional logistic regression offers clear interpretability, nonlinear functional forms and interactions generally need to be specified explicitly, which may make it less flexible for capturing complex nonlinear patterns and higher-order interactions among clinical variables. [9]Random forest (RF), an ensemble learning method based on multiple decision trees, can accommodate nonlinear relationships and complex interactions among predictors without requiring their functional forms to

be prespecified . [10, 11]The aggregation of multiple randomized trees also provides relative robustness to noise [11]. The improved flexibility of complex machine-learning models, however, often comes at the cost of interpretability. [12]Their predictions may be difficult to trace to specific features or decision pathways, potentially undermining clinician trust and limiting translation into routine clinical practice .[13]

SHapley Additive exPlanations (SHAP), grounded in Shapley values from cooperative game theory, decomposes model predictions into feature-level contributions, enabling interpretation at both the global and individual levels. [14]For tree-based algorithms such as random forests and gradient-boosted trees, TreeSHAP offers an efficient approach to quantifying these contributions and can further characterize nonlinear patterns and potential interactions underlying model predictions. [15]Such interpretability may help bridge the gap between predictive performance and clinical understanding.[16]
Against this background, we combined LASSO regression, XGBoost, and Boruta to derive a robust set of candidate features and systematically compared nine machine-learning models for KOA classification. The best-performing model was further evaluated in an independent external sample and interpreted using SHAP to identify the features contributing most strongly to its predictions. We additionally examined clinically relevant cut-off values for the leading markers. This integrated framework was designed to support early KOA screening and more refined risk stratification while improving the clinical interpretability of machine-learning predictions.[17, 18]

2.1 Study design and participants
This retrospective cross-sectional study was conducted at the Third Affiliated Hospital of Guangzhou University of Chinese Medicine and included 1,243 participants with KOA or healthy controls. The development cohort comprised 1,098 individuals enrolled between 2017 and 2019, including 271 with KOA and 827 healthy controls. A temporally distinct validation cohort included 145 individuals enrolled in 2016, of whom 75 had KOA and 70 were healthy controls.
The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Third Affiliated Hospital of Guangzhou University of Chinese Medicine (approval no. PJ-XS-20260325-009). The requirement for informed consent was waived by the ethics committee because of the retrospective nature of the study and the use of anonymized data. An overview of the study workflow is presented in Fig. 1.

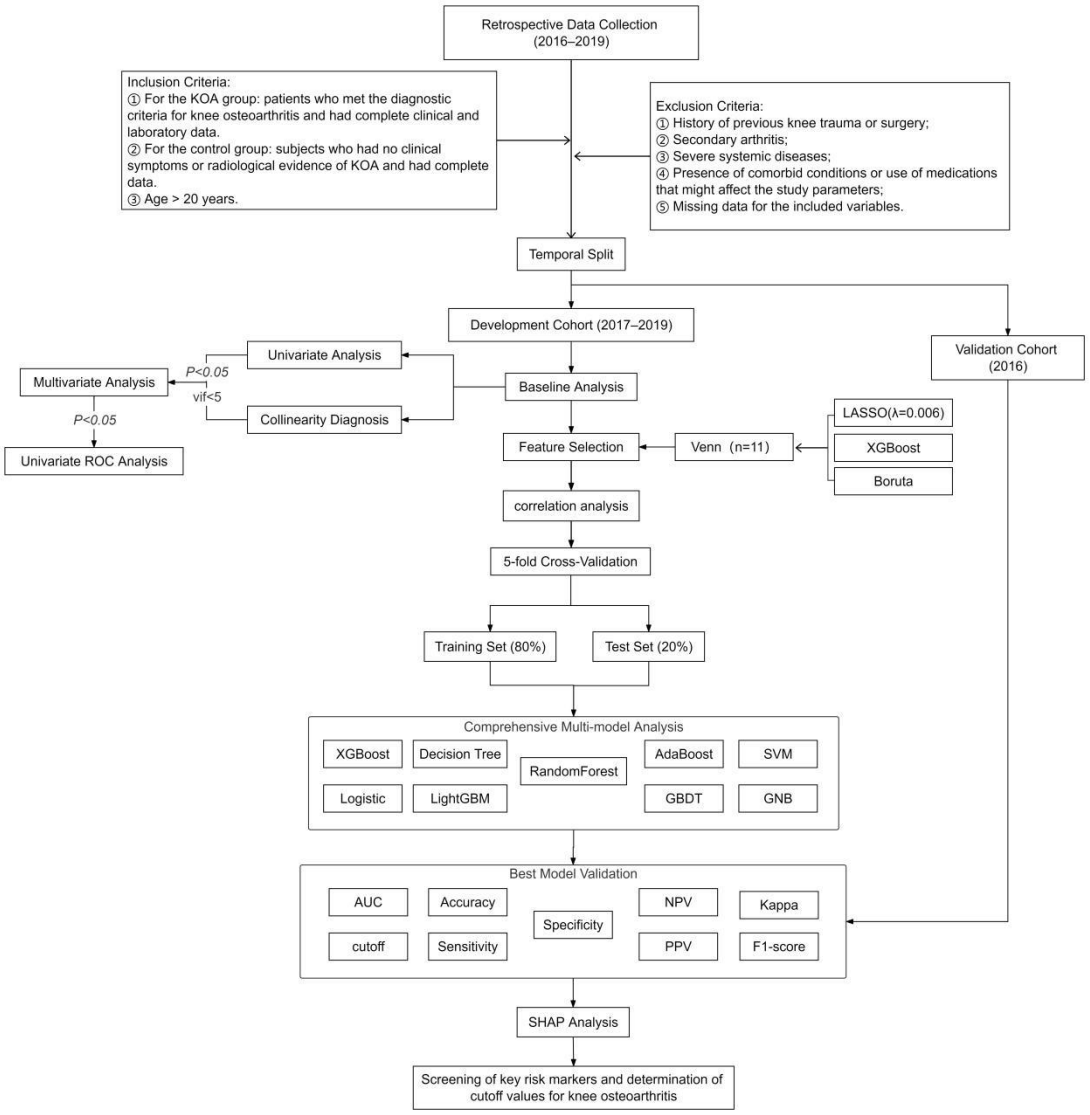

Figure 1 work flow

2.2 Inclusion and exclusion criteria

The inclusion criteria were as follows: (1) participants in the KOA group met the diagnostic criteria for KOA and had complete clinical and laboratory data; (2) healthy controls had no clinical symptoms or imaging evidence of KOA and had complete study data; and (3) all participants were older than 20 years.

The exclusion criteria were: (1) a history of knee trauma or knee surgery; (2) secondary arthritis; (3) severe systemic diseases; (4) medical conditions or use of medications considered likely to affect the study variables; and (5) missing data for any of the variables included in the analysis.

2.3 Candidate variables

A total of 52 candidate variables were collected, encompassing demographic characteristics, routine hematological and biochemical measurements, and derived ratios. These variables included sex, age, WBC, BASO, NEUT, MONO, LYMPH, EO,

BASO%, NEUT%, MONO%, LYMPH%, EO%, RBC, HGB, HCT, MCV, MCH, MCHC, RDW-CV, RDW-SD, PLT, LDL, UA, CHOL, urea, Crea, TG, CO_2, HDL, and the corresponding derived ratios.

2.4 Univariable analysis and feature selection

Multicollinearity was assessed using the variance inflation factor (VIF), and variables with a VIF $\geqslant$ 5 were excluded. Univariable analyses were then performed, and variables with $P < 0.05$ were retained. Variables retained after both multicollinearity assessment and univariable analysis were entered into a multivariable analysis. Those remaining statistically significant at $P < 0.05$ were subsequently evaluated using univariable receiver operating characteristic (ROC) analysis.

Feature selection was additionally performed using least absolute shrinkage and selection operator (LASSO) regression, extreme gradient boosting (XGBoost), and the Boruta algorithm. The features selected by all three methods were identified from their intersection, visualized using a Venn diagram. Eleven overlapping features were ultimately retained for correlation analysis, determination of clinically relevant cut-off values, and subsequent machine-learning modelling.

2.5 Machine-learning model development and evaluation

Nine classification algorithms were evaluated: extreme gradient boosting (XGBoost), logistic regression, Light Gradient Boosting Machine (LightGBM), random forest (RF), AdaBoost, decision tree, gradient boosting decision tree (GBDT), Gaussian naïve Bayes, and support vector machine (SVM). Five-fold cross-validation was used during model development and evaluation. Model performance was assessed using multiple complementary metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Cohen’s kappa.

Calibration was assessed using calibration curves and quantified with the Brier score, whereas decision curve analysis (DCA) was performed to evaluate the potential clinical net benefit of the models across a range of threshold probabilities. Additional analyses addressing class imbalance and sensitivity analyses were conducted to examine the robustness of model performance under alternative analytical conditions.

Hyperparameters for the best-performing RF model were optimized using GridSearchCV. The final parameter settings were criterion = “gini”, max_depth = None, max_features = “sqrt”, min_impurity_decrease = 0, min_samples_leaf = 1, min_samples_split = 2, and n_estimators = 100.

2.6 SHAP-based model interpretation

The best-performing RF model was interpreted using SHapley Additive exPlanations (SHAP). Global feature importance was quantified using the mean absolute SHAP value [mean(|SHAP value|)]. SHAP summary plots were used to visualize both the magnitude and direction of each feature’s contribution to model predictions across the study population. SHAP dependence plots were further generated to characterize nonlinear patterns in feature contributions, while individual force plots were used to

illustrate feature-level contributions to specific predictions.

2.7 Statistical analysis

All statistical analyses were performed using R (versions 4.2.3 and 4.3.0) and Python (version 3.11.4). Continuous variables were summarized as mean ± standard deviation (SD) or median with interquartile range (IQR), as appropriate according to their distributions. Categorical variables were summarized as frequencies and percentages. Between-group comparisons were performed using Student's t test, the Mann–Whitney U test, or the chi-square test, as appropriate. All tests were two-sided, and $P < 0.05$ was considered statistically significant.

3.Results

3.1 Development Queue Baseline Analysis

The development cohort included 1,098 participants, comprising 271 individuals with KOA and 827 healthy controls. Participants with KOA were older than healthy controls, with a median age of 66 versus 44 years. They also had higher EO%, MONO%, LDL, TG, and Urea/Crea values and lower HGB, MCHC, CO_2, and CHOL/TG values (all $P < 0.05$; Table 1).

Variables	Deficiency	Categories	Total (n=1098)	KOA group (n=271)	Healthy control group (n=827)	Statistics	p
sex ,n(%)	0	male	848(77.231)	209(77.122)	639(77.267)	0.002	0.960
		female	250(22.769)	62(22.878)	188(22.733)		
age ,median[IQR]	0		47.000[38.000,60.000]	66.000[58.000,74.000]	44.000[35.000,50.000]	20.531	<0.001
WBC ,median[IQR]	0		6.190[5.190,7.360]	6.600[5.320,7.760]	6.060[5.160,7.210]	3.484	<0.001
BASO ,median[IQR]	0		0.030[0.020,0.040]	0.030[0.020,0.050]	0.020[0.020,0.040]	7.011	<0.001
NEUT ,median[IQR]	0		3.600[2.890,4.520]	3.730[3.000,4.800]	3.530[2.850,4.400]	2.068	0.039
MONO ,median[IQR]	0		0.370[0.300,0.430]	0.400[0.400,0.600]	0.310[0.280,0.400]	13.131	<0.001

LYMPH ,median[IQR]	0	2.000[1.680,2.400]	2.000[1.600,2.400]	2.000[1.700,2.380]	-0.521	0.602
EO ,median[IQR]	0	0.005[0.003,0.008]	0.010[0.006,0.016]	0.005[0.003,0.007]	14.120	<0.001
BASO percent ,median[IQR]	0	0.400[0.300,0.600]	0.500[0.400,0.700]	0.400[0.300,0.600]	5.844	<0.001
NEUT percent ,mean(±SD)	0	58.882±8.138	58.405±9.032	59.039±7.816	-1.033	0.302
MONO percent ,median[IQR]	0	5.800[4.900,6.800]	7.100[6.000,8.400]	5.400[4.700,6.300]	13.751	<0.001
LYMPH percent ,mean(±SD)	0	33.020±7.886	31.211±8.774	33.613±7.479	-4.044	<0.001
EO percent ,median[IQR]	0	0.090[0.060,0.120]	0.150[0.100,0.210]	0.080[0.060,0.100]	16.005	<0.001
RBC ,median[IQR]	0	4.590[4.290,4.930]	4.330[4.060,4.780]	4.650[4.390,4.960]	-8.401	<0.001
HGB ,median[IQR]	0	134.000[125.000,142.000]	125.000[116.000,135.000]	136.000[129.000,145.000]	-10.613	<0.001
HCT ,median[IQR]	0	0.400[0.390,0.430]	0.400[0.370,0.420]	0.410[0.400,0.435]	-7.211	<0.001
MCV ,median[IQR]	0	90.498[86.768,93.525]	92.105[88.679,94.737]	90.038[86.316,92.975]	5.203	<0.001
MCH ,median[IQR]	0	29.551[28.415,30.476]	29.435[28.362,30.465]	29.581[28.421,30.485]	-1.030	0.303
MCHC ,median[IQR]	0	326.531[317.500,334.783]	320.000[310.256,326.667]	328.889[320.000,335.897]	-9.482	<0.001
RDWCV ,median[IQR]	0	12.600[12.000,13.200]	13.100[12.600,13.800]	12.400[11.900,13.000]	10.391	<0.001
RDWSD ,median[IQR]	0	41.300[39.500,43	43.300[41.300,4	40.800[39.200,4	9.703	<0.001

		.600]	5.000]	2.800]		
PLT ,median[IQR]	0	245.000[211.000, 284.000]	242.000[203.000,278.000]	247.000[214.000,285.000]	-2.013	0.044
LDL ,median[IQR]	0	3.010[2.460,3.570]	3.310[2.640,3.990]	2.930[2.430,3.460]	5.272	<0.001
UA ,median[IQR]	0	303.300[247.000, 359.000]	319.500[228.300,391.600]	302.000[250.000,348.000]	1.784	0.074
CHOL ,median[IQR]	0	5.070[4.400,5.780]	5.260[4.480,5.980]	5.000[4.380,5.710]	2.260	0.024
Urea ,median[IQR]	0	4.600[3.800,5.500]	5.400[4.400,6.400]	4.500[3.800,5.200]	9.452	<0.001
Crea ,median[IQR]	0	60.950[53.190,71.960]	65.000[55.000,77.250]	59.810[52.900,69.620]	4.741	<0.001
TG ,median[IQR]	0	1.140[0.780,1.700]	1.510[1.100,2.120]	1.020[0.740,1.570]	8.501	<0.001
CO2 ,median[IQR]	0	25.300[23.000,27.800]	23.500[21.900,25.900]	26.000[23.600,28.200]	-8.786	<0.001
HDL ,median[IQR]	0	1.280[1.090,1.530]	1.340[1.130,1.540]	1.260[1.080,1.530]	1.516	0.130
WBC/PLT ,median[IQR]	0	0.025[0.021,0.031]	0.027[0.023,0.033]	0.025[0.021,0.030]	5.017	<0.001
WBC/RBC ,median[IQR]	0	1.344[1.127,1.610]	1.494[1.228,1.774]	1.303[1.112,1.557]	6.695	<0.001
WBC/HGB ,median[IQR]	0	0.046[0.039,0.055]	0.051[0.043,0.060]	0.045[0.038,0.053]	7.704	<0.001
HGB/RBC ,median[IQR]	0	29.551[28.415,30.476]	29.435[28.362,30.465]	29.581[28.421,30.485]	-1.030	0.303
HGB/PLT ,median[IQR]	0	0.545[0.464,0.653]	0.533[0.427,0.649]	0.551[0.472,0.653]	-2.443	0.015

PLT/RBC ,median[IQR]	0	53.444[45.074,62.680]	54.691[45.074,65.842]	53.360[45.143,61.384]	1.940	0.052
(PLR)PLT/LYMPH ,median[IQR]	0	118.750[98.824,150.588]	116.111[97.083,154.615]	120.000[99.600,149.444]	-0.838	0.402
(NLR)NEUT/LYMPH ,median[IQR]	0	1.781[1.417,2.333]	1.857[1.429,2.474]	1.765[1.411,2.264]	2.185	0.029
(SII)NLR/PLT ,median[IQR]	0	446.351[319.846,598.179]	453.650[327.158,640.286]	445.684[319.846,582.492]	0.871	0.384
LYMPH/MONO ,median[IQR]	0	5.667[4.400,7.138]	4.333[3.286,5.500]	6.081[5.000,7.500]	-12.504	<0.001
NEUT/MONO ,median[IQR]	0	10.310[8.250,12.667]	8.250[6.571,10.500]	11.000[9.000,13.389]	-11.566	<0.001
CHOL/TG ,median[IQR]	0	4.431[3.061,6.340]	3.515[2.455,4.860]	4.883[3.303,6.723]	-8.227	<0.001
LDL/HDL ,median[IQR]	0	2.412[1.875,2.945]	2.553[2.106,2.978]	2.343[1.801,2.931]	3.794	<0.001
MONO/HDL ,median[IQR]	0	0.273[0.208,0.373]	0.352[0.267,0.458]	0.254[0.196,0.339]	10.203	<0.001
FHDL ,median[IQR]	0	3.750[3.140,4.390]	3.890[3.230,4.530]	3.700[3.130,4.330]	2.180	0.029
CHOL/HDL ,median[IQR]	0	3.950[3.417,4.541]	3.934[3.552,4.366]	3.957[3.346,4.602]	0.336	0.737
TG/HDL ,median[IQR]	0	0.915[0.585,1.411]	1.147[0.825,1.714]	0.809[0.532,1.319]	7.427	<0.001
Urea/Crea ,median[IQR]	0	0.075[0.062,0.091]	0.081[0.069,0.097]	0.073[0.060,0.088]	5.354	<0.001
UA/Crea ,median[IQR]	0	4.795[4.014,5.715]	4.715[3.508,5.980]	4.817[4.101,5.649]	-1.608	0.108
UA/HDL ,median[IQR	0	228.319[171.069,	231.858[142.16	227.813[175.401	-0.902	0.367

]		301.667]	2,324.065]	,297.273]		
UA/LYMPH	,median[I	0	146.842[116.786,	155.625[105.29	145.755[117.929	1.085	0.278
QR]			186.458]	4,204.091]	,181.364]		
NLR/PLR	,median[IQ	0	0.015[0.012,0.01	0.016[0.012,0.0	0.014[0.012,0.01	3.430	<0.001
R]			9]	20]	8]		

Table 1: Comparison of demographic and laboratory characteristics between the KOA group and the healthy control group. Continuous variables were expressed as mean ± standard deviation or median (interquartile range) depending on their distribution; categorical variables were expressed as number of cases (%).

After multicollinearity assessment and sequential univariable and multivariable analyses, age, CO_2, and CHOL/TG remained associated with KOA and were retained for ROC analysis (Supplementary Tables S1 – S4). As shown in Fig. 2A, the multivariable logistic regression model incorporating these three variables showed good discrimination in the development cohort, with an AUC of 0.932 (95% CI, 0.916 – 0.948). In the univariable analyses (Fig. 2B), age showed the highest discriminative ability, with an AUC of 0.915 (95% CI, 0.897 – 0.932), numerically higher than those of CO_2 and CHOL/TG. The combined model therefore yielded a higher AUC than any of the individual variables examined.

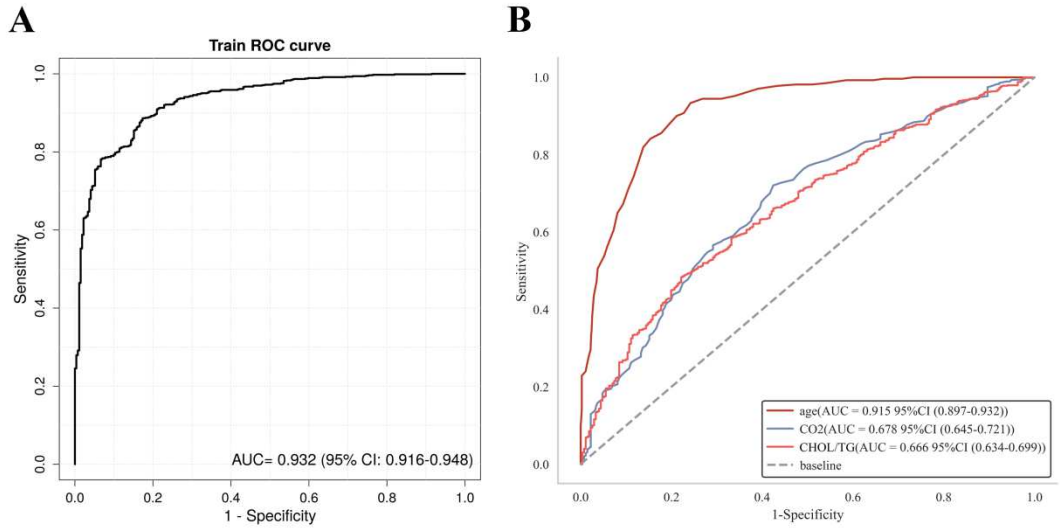

Figure 2: ROC curve and univariate ROC analysis for the multivariate logistic regression model. A: ROC curve for the combined logistic regression model including age, CO_2, and CHOL/TG; AUC = 0.932 (95% CI: 0.916 – 0.948); B: Univariate ROC curves for age, CO_2, and CHOL/TG.

3.3 Multi-method feature selection

To enhance the robustness of feature selection, three complementary approaches—LASSO regression, XGBoost, and Boruta—were applied. For LASSO regression, the cross-validation curve and coefficient path are shown in Fig. 3A and Fig. 3B, respectively, with the corresponding coefficient estimates provided in Supplementary Table S5. The cross-validation performance of XGBoost across different numbers of selected features is presented in Fig. 3C, while feature importance derived from the Boruta algorithm is shown in Fig. 3D.

LASSO, XGBoost, and Boruta selected 17, 22, and 37 features, respectively. The overlap among the three feature sets is illustrated by the Venn diagram in Fig. 3E. Eleven features were consistently selected by all three methods: TG, LDL, EO%, CO_2, HGB, CHOL/TG, MONO%, age, Urea/Crea, MCHC, and UA. These 11 features were subsequently carried forward for correlation analysis, machine-learning model development, SHAP-based interpretation, and cut-off analysis.

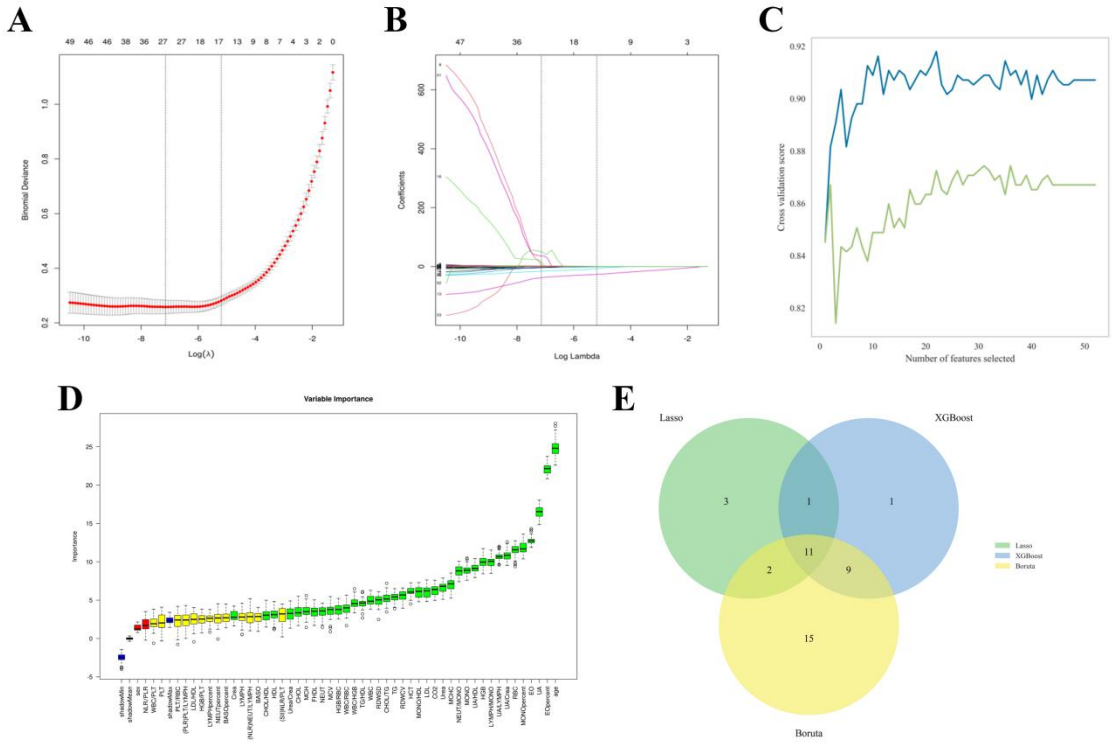

Figure 3: Multimethod feature selection. A: LASSO cross-validation curve; B: LASSO coefficient path; C: Number of selected features versus cross-validation performance in XGBoost; D: Boruta variable importance analysis; E: Venn diagram comparing the feature selection results from LASSO, XGBoost, and Boruta – the three methods share 11 overlapping features.

3.4 Correlation analysis of the selected features

Spearman correlation analysis was performed among the 11 consensus features retained for subsequent modelling. Pairwise correlations are displayed in the heatmap in Fig. 4A, while the corresponding scatterplot matrix with marginal distributions is shown in Fig. 4B. Overall, most pairwise correlations were weak to moderate, with

limited evidence of widespread redundancy among the selected features. A strong inverse correlation was observed between TG and CHOL/TG (Spearman's $\rho = -0.922$), indicating substantial overlap in the information captured by these two variables. HGB and MCHC showed a moderate positive correlation (Spearman's $\rho = 0.416$), suggesting some degree of shared information between these erythrocyte-related measures. Taken together, the correlation structure indicated that most of the 11 features provided nonredundant information, while the strong association between TG and CHOL/TG was considered when interpreting subsequent model and SHAP results.

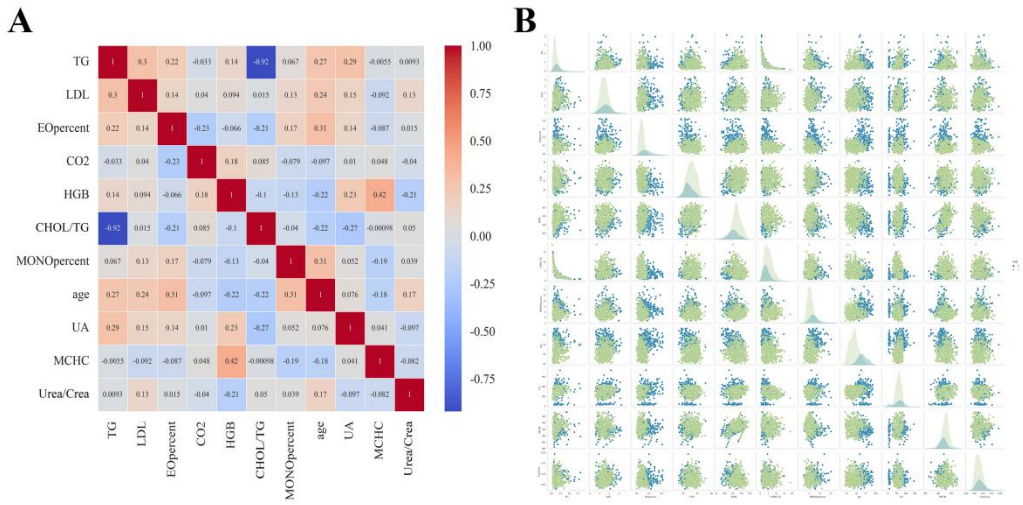

Figure 4: Correlation analysis of the 11 ultimately included candidate features. A: Spearman correlation heatmap; B: Pairwise scatter plots and marginal distributions among variables.

3.5 Comparison of classification performance across models

The classification performance of nine machine-learning algorithms was compared: XGBoost, logistic regression, LightGBM, random forest (RF), AdaBoost, decision tree, GBDT, Gaussian naïve Bayes (GNB), and support vector machine (SVM). The overall performance of the nine models in the temporal validation set is summarized in Fig. 5 and Table 2.

Decision curve analysis showed the potential net benefit of each model across a range of threshold probabilities in the temporal validation set (Fig. 5A). Calibration curves were used to examine the agreement between predicted probabilities and observed outcomes (Fig. 5B). Model discrimination, expressed as AUC with 95% confidence intervals (CIs), was compared in Fig. 5C, while ROC curves for the temporal validation set are presented in Fig. 5D. Table 2 provides the corresponding AUC, optimal probability threshold, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Cohen's kappa for each model.

In the temporal validation set, RF showed the strongest overall classification

performance, with an AUC of 0.988 (95% CI, 0.979–0.998), an accuracy of 0.941, a sensitivity of 0.934, and a specificity of 0.963. XGBoost, logistic regression, GNB, and SVM also achieved comparatively high AUCs, whereas LightGBM showed relatively lower discrimination. ROC and precision–recall (PR) curves for the development set are provided in Supplementary Fig. S1A–C, with detailed performance metrics reported in Supplementary Table S6. Pairwise comparisons of model AUCs using DeLong's test are reported in Supplementary Table S7. Considering its validation-set discrimination, overall classification performance, and suitability for subsequent model interpretation, RF was selected as the primary model for further analysis.

Table 2: Validation set results for nine machine learning models

Disaggregated model	AUC(95%CI)	cutoff(95%CI)	Accuracy(95%CI)	Sensitivity(95%CI)	Specificity(95%CI)	positive predictive value(95%CI)	Negative Predictive Value(95%CI)	F1 Score(95%CI)	Kappa(95%CI)
XGBoost	0.980 (0.965-0.994)	0.773(0.773-0.773)	0.9(0.900-0.900)	0.898(0.898-0.898)	0.907(0.907-0.907)	0.968(0.968-0.968)	0.742(0.742-0.742)	0.931(0.931-0.931)	0.749(0.749-0.749)
logistic	0.981 (0.967-0.996)	0.711(0.711-0.711)	0.918(0.918-0.918)	0.904(0.904-0.904)	0.963(0.963-0.963)	0.987(0.987-0.987)	0.765(0.765-0.765)	0.943(0.943-0.943)	0.797(0.797-0.797)
LightGBM	0.836 (0.754-0.918)	0.88(0.880-0.880)	0.818(0.818-0.818)	0.813(0.813-0.813)	0.833(0.833-0.833)	0.938(0.938-0.938)	0.592(0.592-0.592)	0.871(0.871-0.871)	0.568(0.568-0.568)
RandomForest	0.988 (0.979-0.998)	0.66(0.660-0.660)	0.941(0.941-0.941)	0.934(0.934-0.934)	0.963(0.963-0.963)	0.987(0.987-0.987)	0.825(0.825-0.825)	0.96(0.960-0.960)	0.849(0.849-0.849)
AdaBoost	0.975 (0.958-0.991)	0.51(0.510-0.510)	0.918(0.918-0.918)	0.946(0.946-0.946)	0.833(0.833-0.833)	0.946(0.946-0.946)	0.833(0.833-0.833)	0.946(0.946-0.946)	0.779(0.779-0.779)

Decisio	0.959	0.92(0.88	0.909(0.9	0.888(0.8	0.974(0.9	0.991(0.9	0.739(0.736-0.7	0.936(0.	0.778(0.7
nTree	(0.930-	8-0.953)	09-0.909)	85-0.891)	65-0.983)	87-0.994)	42)	936-0.93	77-0.780)
	0.987)							7)	
GBDT	0.976	1.0(1.000-	0.895(0.8	0.867(0.8	0.981(0.9	0.993(0.9	0.707(0.707-0.7	0.926(0.	0.75(0.75
	(0.959-	1.000)	95-0.895)	67-0.867)	81-0.981)	93-0.993)	07)	926-0.92	0-0.750)
	0.992)							6)	
GNB	0.981	0.871(0.8	0.905(0.9	0.88(0.88	0.981(0.9	0.993(0.9	0.726(0.726-0.7	0.933(0.	0.77(0.77
	(0.967-	71-0.871)	05-0.905)	0-0.880)	81-0.981)	93-0.993)	26)	933-0.93	0-0.770)
	0.995)							3)	
SVM	0.985	0.814(0.8	0.923(0.9	0.916(0.9	0.944(0.9	0.981(0.9	0.785(0.785-0.7	0.947(0.	0.805(0.8
	(0.972-	00-0.829)	23-0.923)	16-0.916)	44-0.944)	81-0.981)	85)	947-0.94	05-0.805)
	0.997)							7)	

Table 2: Comparison of the classification performance of nine machine learning models on the validation set. The table reports AUC, optimal probability threshold, accuracy, sensitivity, specificity, PPV, NPV, F1 score, and Kappa value.

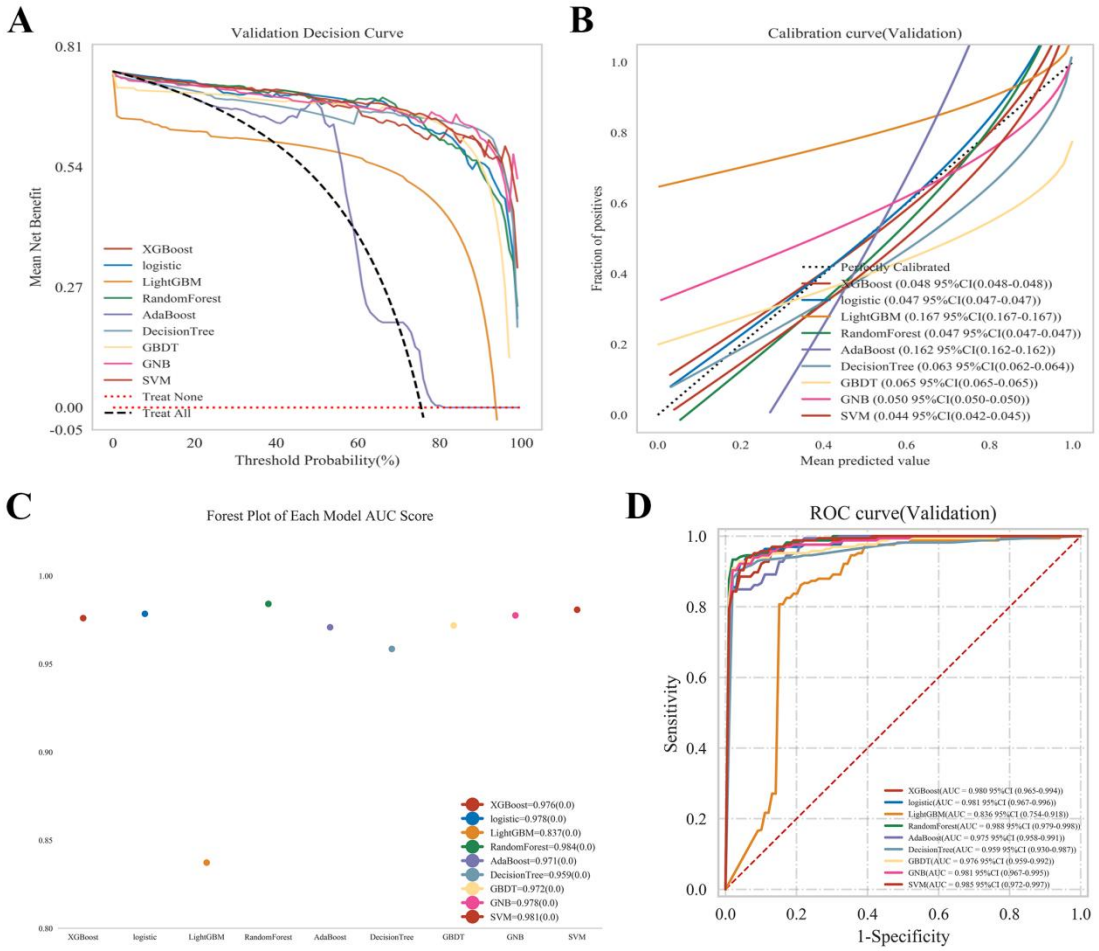

Figure 5: Performance on the validation set for nine machine learning algorithms. A: Validation set decision curve; B: Validation set calibration curve; C: AUC forest plot for each model; D: Validation set ROC curve.

3.6 Performance of the final random forest model

The final RF model incorporated 11 features: age, EO%, MONO%, HGB, MCHC, LDL, UA, TG, CO_2, CHOL/TG, and Urea/Crea. Of the 1,098 participants in the development cohort, 220 (20.04%) were randomly assigned to an internal hold-out test set. In this test set, the RF model achieved an AUC of 0.978 (95% CI, 0.962 – 0.994; Fig. 6A).

The learning curves showed that training and cross-validation performance progressively stabilized as the training sample size increased (Fig. 6B). Decision curve analysis indicated the potential net benefit of the RF model across a range of threshold probabilities in the internal test set (Fig. 6C). The calibration curve showed close agreement between predicted probabilities and observed KOA status, with a Brier score of 0.056 (95% CI, 0.040 – 0.075; Fig. 6D).

Detailed performance metrics for the training, cross-validation, and internal test sets are provided in Supplementary Tables S8 – S10. Additional evaluations of the RF model are presented in Supplementary Fig. S2, including ROC curves and confusion

matrices for the corresponding datasets (Supplementary Fig. S2A – D), the Kolmogorov – Smirnov (KS) plot for the internal test set (Supplementary Fig. S2E), and a parallel-coordinate plot of the 11 selected features (Supplementary Fig. S2F). Together, these analyses indicated that the final RF model maintained strong discrimination and consistent classification performance within the development cohort.

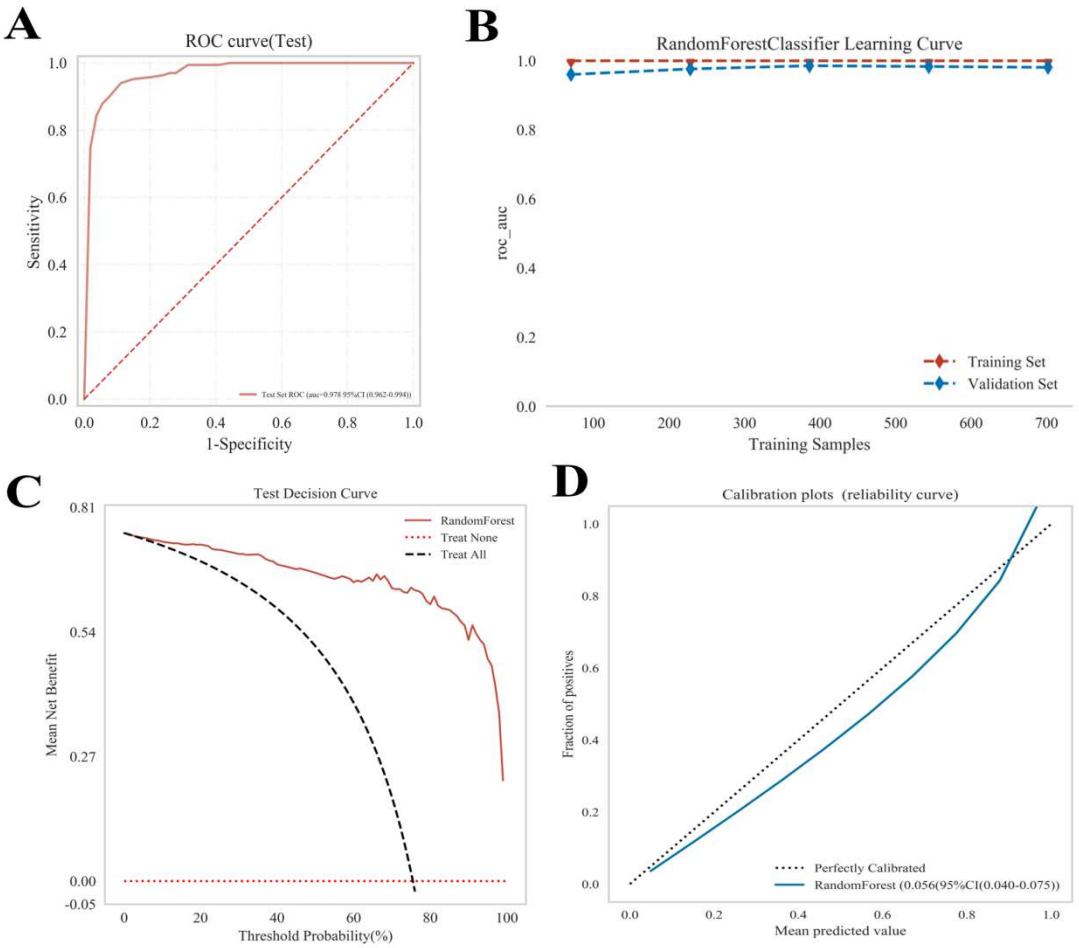

Figure 6: Internal validation performance of the final 11-variable random forest model. A: Test set ROC curve; B: Learning curve; C: Test set decision curve; D: Calibration curve

3.7 SHAP-based model interpretation

The SHAP summary plot (Fig. 7A) illustrates the distribution, magnitude, and direction of feature contributions to the RF model output. Higher age values were predominantly associated with SHAP values that shifted the model output toward the KOA class, indicating that age contributed strongly to KOA-oriented predictions. Higher EO% and MONO% values showed similar patterns, contributing predominantly toward predictions of KOA. In contrast, HGB showed an inverse pattern: higher HGB values tended to shift predictions toward the healthy-control

class, whereas lower values contributed toward KOA predictions.

Ranking by mean absolute SHAP value identified age as the most influential feature in the model, followed by EO%, UA, MONO%, HGB, CO_2, MCHC, LDL, TG, Urea/Crea, and CHOL/TG (Fig. 7B). The SHAP dependence plot for CHOL/TG suggested a nonlinear pattern in its contribution to the model output (Fig. 7C). Individual SHAP force plots further illustrated how different features contributed positively or negatively to model predictions for representative participants (Fig. 7D – G). Collectively, the SHAP analyses indicated that RF classification was driven by the combined contributions of age, inflammation-related, erythrocyte-related, and metabolic features rather than by any single variable.

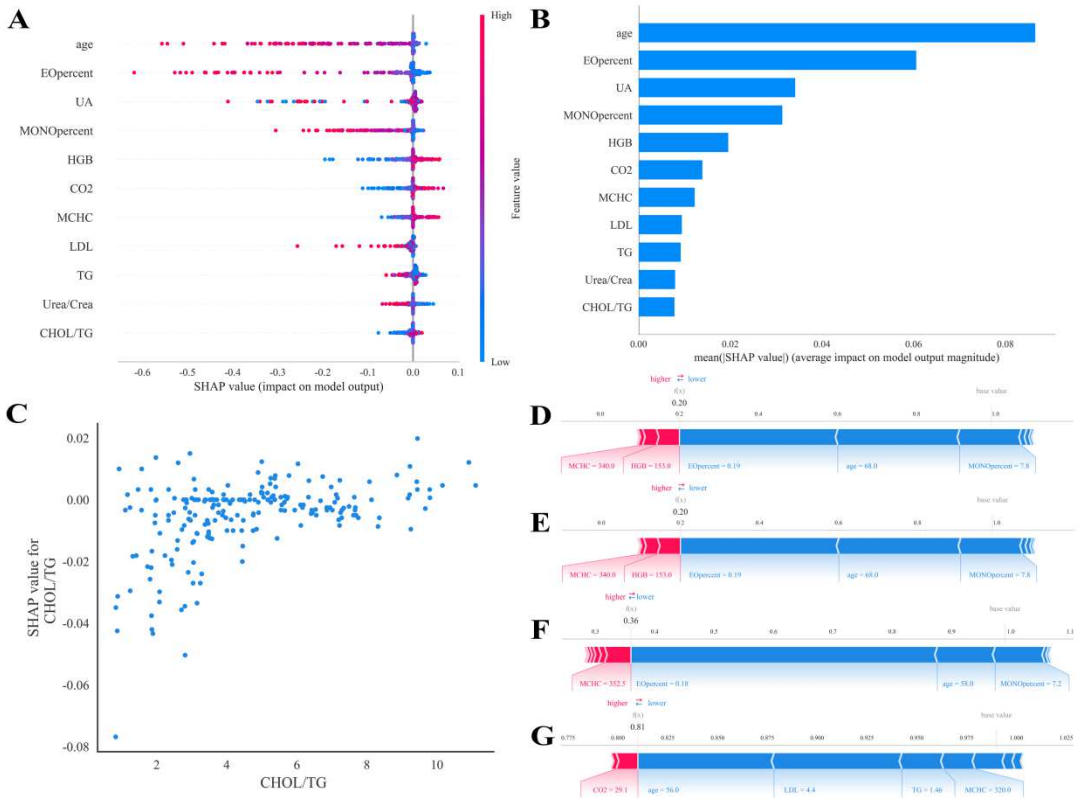

Figure 7: SHAP interpretation for the Random Forest model. A: SHAP summary plot; B: Ranking of average absolute SHAP values; C: SHAP dependence plot for CHOL/TG; D – G: SHAP force plots for representative individuals. Red indicates higher feature values; blue indicates lower feature values.

3.8 SMOTE-based class balancing analysis

The development cohort showed clear class imbalance, comprising 271 participants with KOA and 827 without KOA. In an exploratory analysis, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the class distribution, increasing the KOA class to 813 while retaining 827 non-KOA participants, resulting in a total sample of 1,640. Of these, 328 cases (20.00%) were assigned to the test set. Under the SMOTE-based analysis, the model achieved an AUC of 0.993 (95% CI,

0.990 – 0.995) in the validation set and 0.984 (95% CI, 0.972 – 0.996) in the test set (Fig. 8). As shown in Fig. 8A, the model retained strong discrimination in the test set after SMOTE processing, with an AUC of 0.984 (95% CI, 0.972 – 0.996).

The learning curves are presented in Fig. 8B. Decision curve analysis indicated a positive net benefit across relevant threshold probabilities (Fig. 8C). The calibration curve suggested good agreement between predicted probabilities and observed outcomes, with a Brier score of 0.044 (95% CI, 0.032 – 0.059; Fig. 8D). SHAP summary plots and rankings based on mean absolute SHAP values are shown in Fig. 8E and Fig. 8F, respectively. The SHAP dependence plot for Urea/Crea is presented in Fig. 8G, and representative SHAP force plots for individual predictions are displayed in Fig. 8H – K.

Additional performance metrics for the training, validation, and test sets are provided in Supplementary Tables S11 – S13. Supplementary Fig. S3 presents additional diagnostic plots, including ROC curves, confusion matrices, the Kolmogorov – Smirnov (KS) statistic plot, and the parallel-coordinate plot. Specifically, Supplementary Fig. S3A and S3B show ROC curves for the training and validation sets after SMOTE, Supplementary Fig. S3C and S3D show the corresponding confusion matrices, Supplementary Fig. S3E presents the KS plot for the test set, and Supplementary Fig. S3F shows the parallel-coordinate plot of the 11 selected features. Taken together, these findings suggested that RF performance remained robust after class balancing.

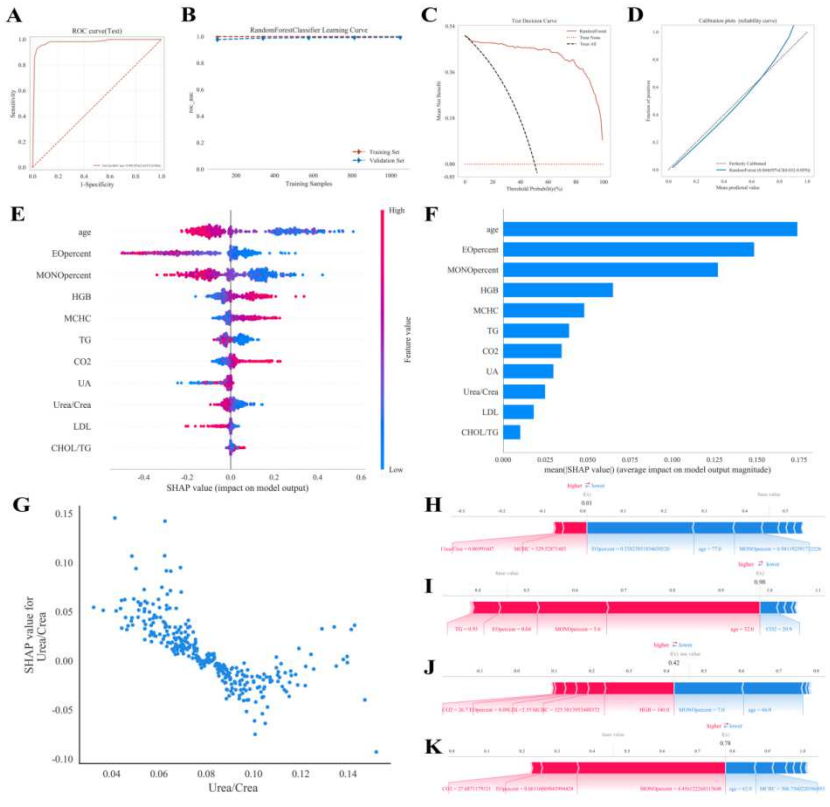

Figure 8: Random Forest sensitivity analysis after SMOTE-based class sample

balancing. A: Test set ROC curve; AUC = 0.984 (95% CI: 0.972 – 0.996); B: Learning curve; C: Decision curve; D: Calibration curve; E: SHAP summary plot; F: Ranking of average absolute SHAP values; G: SHAP dependence plot for Urea/Crea; H – K: SHAP explanations for representative individuals.

3.9 Sensitivity analysis excluding age

Because baseline analyses showed a marked between-group difference in age, a sensitivity analysis was performed by excluding age from the RF model. In the test set, the age-excluded model achieved an AUC of 0.943 (95% CI, 0.911 – 0.976; Fig. 9A), indicating that although age contributed important discriminatory information, it did not fully account for the model’s classification performance. Learning-curve analysis suggested stable performance with increasing sample size (Fig. 9B). Decision curve analysis showed positive net benefit across relevant threshold probabilities (Fig. 9C), and the calibration curve indicated reasonable agreement between predicted probabilities and observed outcomes, with a Brier score of 0.076 (95% CI, 0.053 – 0.103; Fig. 9D). SHAP-based analyses for the age-excluded model are shown in Fig. 9E – K. Additional ROC curves, confusion matrices, the KS plot, and the parallel-coordinate plot are provided in Supplementary Fig. S4, with detailed performance metrics reported in Supplementary Tables S14 – S16. Overall, these findings suggest that hematological and metabolic features continued to provide substantial discriminatory information for KOA classification even after exclusion of age.

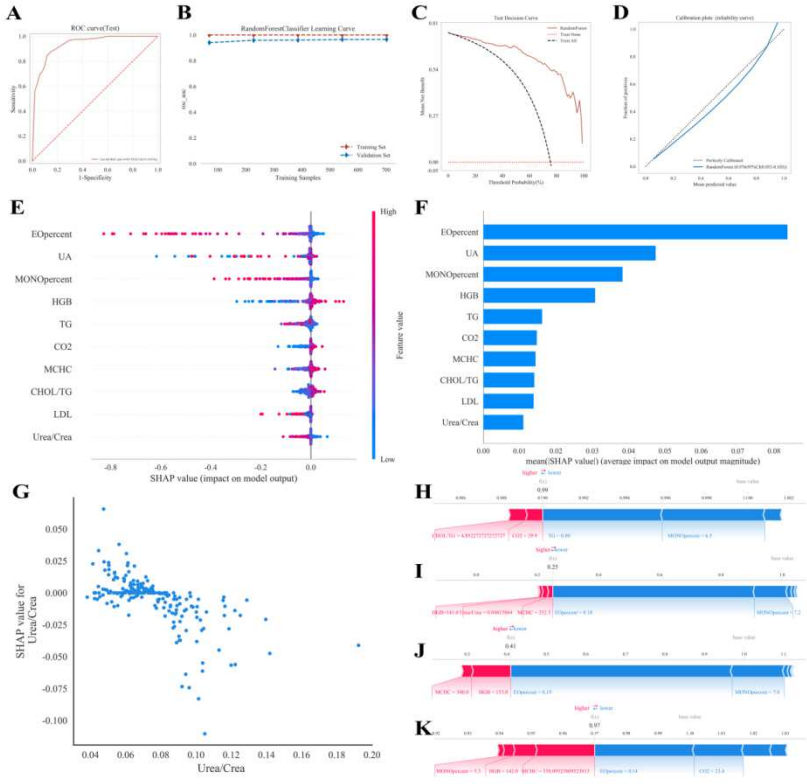

Figure 9: Random Forest sensitivity analysis for 10 variables after adjustment for age. A: Test set ROC curve; AUC = 0.943 (95% CI: 0.911 – 0.976); B: Learning curve; C: Decision curve; D: Calibration curve; E: SHAP summary plot; F: Ranking of average absolute SHAP values; G: SHAP dependence plot for Urea/Crea; H – K: SHAP explanations for representative individuals.

3.10 Independent temporal validation

The independent temporal validation cohort comprised 145 participants, including 75 individuals with KOA and 70 healthy controls. To evaluate model performance in a temporally independent sample, the final 11-feature RF model was applied to this cohort. The model maintained strong discrimination, achieving an AUC of 0.972 (95% CI, 0.951 – 0.994; Fig. 10A).

Calibration analysis showed reasonable agreement between predicted probabilities and observed KOA status, with a Brier score of 0.082 (95% CI, 0.055 – 0.115; Fig. 10B). Decision curve analysis showed a positive net benefit across threshold probabilities ranging from 0.10 to 1.00 (Fig. 10C). The confusion matrix showed that the model correctly classified 91.0% of participants in the temporal validation cohort (Fig. 10D).

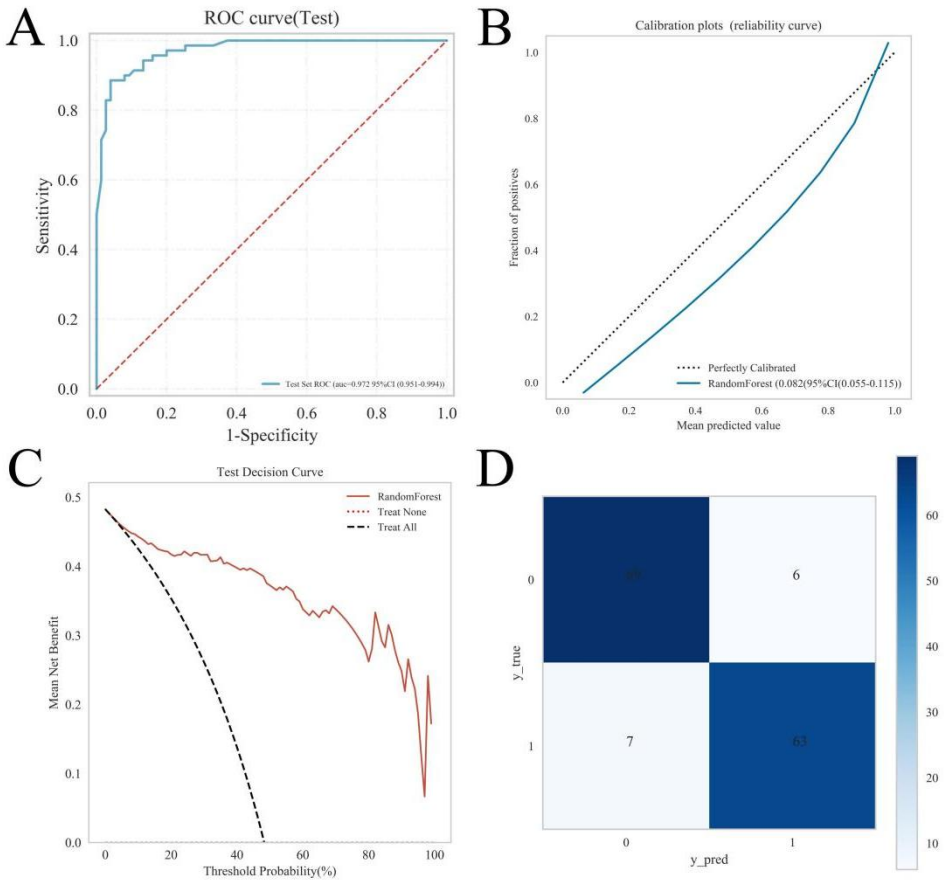

Figure 10: A: ROC curve for the external validation test set; B: Calibration curve; C: DCA curve; D: Matrix plot

At a probability threshold of 0.70, the model achieved an accuracy of 0.910, a sensitivity of 0.900, a specificity of 0.920, a positive predictive value (PPV) of 0.913, a negative predictive value (NPV) of 0.908, an F1 score of 0.906, and a Cohen’s kappa of 0.820 (Table 3). Collectively, these metrics indicated consistently strong classification performance in the temporally independent cohort. The median age was 74 years in the KOA group and 40 years in the healthy-control group. Compared with healthy controls, participants with KOA had higher MONO%, EO%, TG, and Urea/Crea values and lower HGB and CHOL/TG values (Supplementary Table S17).

Table 3: External validation performance of the Random Forest model

AUC	cutoff	Accuracy	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	F1 Score	Kappa	AUC_L	AUC_U
0.972	0.7	0.91	0.9	0.92	0.913	0.908	0.906	0.82	0.951	0.994

Table 3: Predictive performance of the final 11-variable Random Forest model in an independent external validation cohort. AUC_L and AUC_U denote the lower and upper bounds of the 95% confidence interval for the AUC, respectively.

3.11 Cut-off analysis of individual markers

In the cut-off analysis, age showed the highest discriminative performance, with an AUC of 0.915 (95% CI, 0.899–0.932). The optimal cut-off values for age and the remaining markers are summarized in Table 4.

Table 4: Threshold Analysis of Included Variables

Feature	N	AUC	sensitivity	specificity	Youden index	Optimal threshold	accuracy
age	1098	0.915(0.899-0.932)	0.934(0.818-0.955)	0.758(0.747-0.878)	0.692(0.664-0.739)	51.000(51.000-57.000)	0.827(0.791-0.867)
EOpercent	1098	0.824(0.792-0.852)	0.668(0.559-0.749)	0.839(0.767-0.936)	0.507(0.462-0.563)	0.120(0.110-0.140)	0.812(0.754-0.853)
HGB	1098	0.715(0.679-0.755)	0.735(0.650-0.831)	0.609(0.508-0.725)	0.344(0.296-0.418)	130.000(126.000-132.525)	0.711(0.651-0.761)
MCHC	1098	0.692(0.660-0.726)	0.643(0.475-0.692)	0.697(0.652-0.846)	0.341(0.289-0.400)	324.444(323.404-329.787)	0.650(0.565-0.686)
UA	1098	0.536(0.496-0.581)	0.214(0.175-0.474)	0.982(0.737-1.000)	0.196(0.164-0.249)	411.800(333.900-421.242)	0.786(0.672-0.819)

Feature	N	AUC	sensitivity	specificity	Youden index	Optimal threshold	accuracy
TG	1098	0.672(0.637-0.709)	0.760(0.631-0.800)	0.543(0.525-0.680)	0.303(0.256-0.373)	1.100(1.100-1.320)	0.617(0.581-0.671)
CO2	1098	0.678(0.637-0.717)	0.721(0.512-0.774)	0.576(0.502-0.761)	0.296(0.236-0.367)	24.000(23.548-25.952)	0.673(0.571-0.716)
CHOLTG	1098	0.666(0.637-0.694)	0.484(0.444-0.730)	0.779(0.541-0.848)	0.262(0.220-0.316)	5.053(3.499-5.252)	0.590(0.525-0.682)
UreaCrea	1098	0.608(0.572-0.647)	0.664(0.334-0.822)	0.510(0.358-0.837)	0.174(0.134-0.261)	0.074(0.065-0.094)	0.575(0.477-0.707)
MONOpercent	1098	0.778(0.747-0.808)	0.779(0.561-0.827)	0.660(0.637-0.849)	0.439(0.394-0.515)	6.000(5.900-6.700)	0.728(0.672-0.791)
LDL	1098	0.607(0.565-0.653)	0.554(0.328-0.642)	0.644(0.601-0.870)	0.198(0.152-0.278)	3.210(3.140-3.801)	0.652(0.598-0.740)

Table 4: Single-factor ROC analysis is used to evaluate the discriminative ability of the final 11 candidate variables. The table reports the AUC, sensitivity, specificity, Youden's index, optimal threshold, and accuracy; values in parentheses indicate the 95% confidence intervals (CI).

4.Discussion

This study developed an interpretable random forest‑based framework for KOA classification using routinely available demographic and laboratory variables. Three complementary feature-selection approaches converged on 11 candidate features. Among the nine machine-learning algorithms evaluated, RF showed the strongest overall classification performance, achieving an AUC of 0.988 in the model-comparison validation set. The subsequently finalized 11-feature RF model achieved AUCs of 0.983 and 0.978 in the internal validation and hold-out test sets, respectively. SHAP analysis indicated that model predictions were jointly informed by age and multiple inflammatory, hematological, and metabolic features rather than by any single variable. Discrimination remained high in the independent temporal validation cohort (AUC = 0.972), after SMOTE-based class balancing (AUC = 0.984), and after exclusion of age (AUC = 0.943). Collectively, these findings suggest that the observed classification performance was not driven solely by class imbalance or the marked age difference between groups and support the potential utility of the RF-based approach for KOA screening and risk stratification, although prospective multicenter validation remains necessary before clinical implementation.

Age emerged as the most influential feature in the RF model, ranking first in the global SHAP analysis and showing the highest univariable discrimination, with an AUC of 0.915 and an optimal cut-off value of 51 years. Participants with KOA were also markedly older than healthy controls, consistent with established epidemiological

evidence linking advancing age to KOA. Biologically, ageing may contribute to disruption of joint homeostasis through chondrocyte senescence, mitochondrial dysfunction, impaired proteostasis, chronic low-grade inflammation, and altered intercellular communication, thereby affecting cartilage, subchondral bone, synovium, and other joint tissues. [19, 20]Nevertheless, the pronounced difference in age distribution between groups raised the possibility that model discrimination might be partly age-driven. To address this concern, we repeated the RF analysis after excluding age. The age-excluded model retained good discrimination, with a test-set AUC of 0.943, indicating that substantial classification information remained in the hematological and metabolic features. Thus, age appears to be an important contributor to KOA classification, but not the sole driver of model performance, and its clinical value may be greatest when interpreted alongside other routinely available laboratory variables.

Monocyte-associated inflammatory activity may represent one component of the KOA phenotype. In our study, MONO% was higher in participants with KOA, and higher values tended to shift RF predictions toward the KOA class in SHAP analysis. This observation is biologically plausible given evidence that KOA is accompanied by chronic low-grade inflammation and synovial immune activation involving monocytes and macrophages.[6, 21]Circulating monocytes from patients with KOA have been reported to exhibit increased expression of activation markers including CD16, CCR2, and HLA-DR, with CCR2 expression also associated with BMI and pain. [22]Mechanistic studies further implicate the CCL3/CCR1 axis in recruiting circulating $CD14^{+}CD16^{-}$ monocytes to inflamed KOA synovium,[23] whereas CCL2/CCR2-mediated monocyte recruitment has been linked to macrophage accumulation, synovitis, and cartilage damage.[24]Recent peripheral-blood immunophenotyping studies have further identified systemic alterations in specific circulating monocyte subsets in KOA. [25]Together, these observations suggest that MONO% may capture aspects of systemic inflammatory and innate immune activity associated with KOA. Nevertheless, because our study was cross-sectional, it remains unclear whether altered MONO% represents a contributing factor, a consequence of established disease, or an accompanying inflammatory phenotype. Longitudinal and mechanistic studies are needed to clarify its biological and clinical significance.

EO% emerged as another potentially informative hematological feature in our model. It ranked highly in the SHAP analysis and showed a consistent directional pattern in both the development cohort and the independent temporal validation cohort. In contrast to the extensive literature on monocyte/macrophage involvement in KOA, the role of eosinophils remains comparatively understudied.[26]A recent NHANES-based analysis reported associations of peripheral eosinophil count and the eosinophil-to-lymphocyte ratio with OA status,[27]suggesting that eosinophil-related measures may carry clinically relevant information in osteoarthritis. However, this evidence does not establish a causal role for eosinophils in KOA, and the available literature is not specific to knee disease. Our findings therefore position EO% as a

candidate hematological marker for further investigation rather than as an established KOA biomarker. [28, 29]Because EO% is a relative leukocyte measure, it may also be influenced by age, immune homeostasis, metabolic status, medication use, and changes in other leukocyte subsets. KOA-specific cohort studies and mechanistic investigations are needed to determine whether EO% provides reproducible and clinically meaningful information beyond these potential sources of variation.

HGB and MCHC may capture additional information related to erythrocyte status and overall systemic health. In our study, both measures were lower in participants with KOA, and lower HGB values tended to shift RF predictions toward the KOA class in SHAP analysis. However, HGB is influenced by a wide range of factors, including age, sex, nutritional status, renal function, chronic inflammation, and comorbid disease, making it unlikely that its association with KOA reflects a disease-specific pathway alone. [30, 31]Previous work has also suggested that routine hematological and renal-function variables may contribute to modelling KOA severity. In one study of 572 patients with KOA, a LASSO- and logistic regression – based severity model retained age, MCHC, and blood urea nitrogen among its final predictors.[32, 33] Although the outcome and renal marker differed from those in our study, this finding is broadly consistent with the inclusion of MCHC and Urea/Crea in our final RF model. Notably, HGB remained significantly different between groups in the independent temporal validation cohort, whereas MCHC did not, suggesting that HGB may have a more reproducible cross-sample association. Nevertheless, its clinical relevance should be reassessed after adjustment for anemia, nutritional status, renal function, and other potential confounders.

Metabolic features also contributed substantially to the RF model, with TG, LDL, UA, CHOL/TG, Urea/Crea, and CO_2 retained among the final predictors. This finding is consistent with the broader concept that OA may have systemic metabolic components, although evidence for individual metabolic factors remains heterogeneous. [34]In particular, the association between dyslipidemia and OA has been inconsistent across studies. A systematic review and meta-analysis by Baudart and colleagues reported a relatively high prevalence of dyslipidemia among patients with OA and identified a statistical association with knee OA. [35]By contrast, another meta-analysis of metabolic syndrome and its components found positive associations of KOA with metabolic syndrome, hypertension, and hyperglycemia, but not with dyslipidemia considered in isolation. [36]These divergent findings suggest that lipid-related measures may interact with obesity, ageing, inflammatory status, and other metabolic disturbances rather than acting as independent disease-specific determinants. [7]In this context, TG, LDL, and CHOL/TG in our model are best viewed as components of a broader metabolic phenotype that contributes additional classification information.

Uric acid has biological and epidemiological support as a potentially relevant feature in OA. Denoble and colleagues reported that synovial-fluid uric acid concentrations in

knee OA were associated with IL-1β, IL-18, and disease severity, supporting a possible link between urate, inflammasome activation, and the local inflammatory milieu . [37]In a longitudinal cohort of patients with medial knee OA but without gout, higher baseline serum urate was subsequently associated with greater joint-space narrowing over 24 months. [38]However, this association has not been consistently reproduced, as a community-based longitudinal study found no positive association between serum urate and radiographic progression or tibiofemoral cartilage loss. [39]These findings suggest that the relationship between uric acid and KOA progression remains uncertain and may depend on disease phenotype or population characteristics. In our cohort, TG was higher and CHOL/TG lower among participants with KOA. Because TG is mathematically incorporated into CHOL/TG and the two variables were strongly correlated, the independent biological interpretation of CHOL/TG should be approached with caution. Interestingly, UA was not significantly different between groups in univariable analysis but was nevertheless retained by several machine-learning feature-selection approaches. This illustrates that features with limited univariable discrimination may still contribute classification information within a multivariable model through nonlinear patterns or their combined contribution with other variables.[10]

SHAP-based interpretable machine-learning frameworks have recently been applied in various clinical settings to identify important predictors and improve understanding of model decision-making. [40, 41]By combining global feature importance with summary, dependence, and individual force plots, our analysis extended interpretation beyond simple feature ranking to the direction and magnitude of feature contributions, nonlinear patterns, and individual-level prediction profiles. [15]However, SHAP values represent attribution within the fitted model rather than causal effects. [42]This distinction is particularly important when predictors are correlated, because model attribution may be shared across related features and should not be interpreted as evidence of an independent biological effect. Likewise, apparent inflection points in SHAP dependence plots should not be equated directly with clinically meaningful diagnostic thresholds. For this reason, we used a complementary framework integrating multi-method feature selection, RF modelling, SHAP-based interpretation, and ROC-based cut-off estimation, with candidate statistical cut-off values determined by maximizing the Youden index.[43]

The final RF model maintained strong discrimination in the independent temporal validation cohort, with an AUC of 0.972, a sensitivity of 0.900, and a specificity of 0.920, providing preliminary evidence of temporal robustness. Several key features, including age, EO%, MONO%, HGB, TG, CHOL/TG, and Urea/Crea, also showed broadly consistent between-group patterns across the development and temporal validation cohorts. Because the model is based largely on age and routinely available hematological and biochemical measurements, it may offer practical advantages in terms of accessibility and ease of implementation. Nevertheless, the validation cohort originated from the same institution during a different time period and therefore

represents temporal rather than fully independent external validation.The present findings consequently do not establish transportability across hospitals, geographic regions, or broader patient populations. Larger prospective multicenter studies are needed to confirm model discrimination and calibration, assess clinical net benefit, and determine whether the model provides incremental value beyond conventional clinical factors. [44]

This study has several strengths. First, three complementary feature-selection approaches—LASSO regression, XGBoost, and Boruta—were used to derive a consensus feature set, thereby reducing dependence on the results of any single selection method. Second, nine machine-learning algorithms were systematically compared, and model performance was evaluated across multiple dimensions, including discrimination, calibration, and potential clinical net benefit. Third, SHAP enabled interpretation of the RF model at both global and individual levels, extending model assessment beyond predictive performance to transparency of the prediction process. Fourth, robustness was examined through temporal validation, SMOTE-based class balancing, and sensitivity analysis excluding age. Finally, ROC-based analyses provided candidate cut-off values for the 11 selected features, facilitating their further evaluation in clinical and prospective studies.

Despite these strengths, several limitations should be acknowledged. First, the cross-sectional design precludes determination of temporal relationships and causal inference; the identified features should therefore be interpreted as correlates of prevalent KOA rather than determinants of future disease development. Second, the marked age difference between the KOA and control groups represents an important potential source of confounding. Although the age-excluded model retained good discrimination, residual age-related confounding cannot be eliminated, and confirmation in independently recruited age-matched or age-stratified cohorts is warranted. Third, established KOA-related factors such as BMI, previous knee injury, occupational loading, physical activity, pain severity, and radiographic severity were unavailable. The model therefore captures classification information primarily from age and routine laboratory measures rather than the full clinical and etiological profile of KOA. Fourth, correlations among predictors, particularly between TG and CHOL/TG, may redistribute SHAP attribution across related features; consequently, SHAP importance should not be interpreted as an independent biological or causal effect. Fifth, the ROC-derived cut-off values based on the maximum Youden index are sample-specific statistical candidates whose reproducibility may vary with case mix, disease spectrum, and laboratory measurement conditions. They require independent validation before being considered for clinical decision-making. Finally, the temporal validation cohort was relatively small (n = 145) and originated from the same institution as the development cohort. Larger prospective multicenter studies across different geographic and clinical settings are therefore required to establish model transportability, incremental clinical value, and real-world utility.

5. Conclusions

In this study, we developed an interpretable RF-based model for KOA classification using routinely available demographic and laboratory variables. The final 11-feature model showed strong discrimination in both the internal hold-out test set (AUC = 0.978) and the independent temporal validation cohort (AUC = 0.972). SHAP analysis identified age as the most influential feature, while EO%, MONO%, HGB, UA, and other hematological and metabolic measures provided additional classification information. Candidate cut-off values were 51 years for age, 24.000 for CO_2, and 5.053 for CHOL/TG. Importantly, the model retained good discrimination after exclusion of age (AUC = 0.943), indicating that its performance was not driven solely by age-related differences. These findings support the potential value of routinely available clinical and laboratory features for KOA screening and risk stratification, although prospective multicenter studies are required to validate the model, candidate cut-off values, and incremental clinical utility.

Author Contributions

Songyao Qi: Methodology, Software, Formal analysis, Data curation, Validation, Visualization, Writing – original draft. Zhuoyan Chen: Methodology, Software, Formal analysis, Data curation, Validation, Visualization, Writing – original draft. Ziwen Xia: Investigation, Data curation, Validation. Zilin Chen: Investigation, Data curation, Resources. Zixuan Chen: Investigation, Data curation, Validation. Feng Lin: Investigation, Resources, Data curation. Jing Su: Conceptualization, Methodology, Supervision, Project administration, Resources, Writing – review & editing, Funding acquisition. Songyao Qi and Zhuoyan Chen contributed equally to this work and share first authorship. All authors read and approved the final manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and publication of this article: This work was supported by the Undergraduate Teaching Reform Research Project of Guangdong Clinical Teaching Base,Grant No:2025JD073;and the Teaching Quality and Teaching Reform Project of Guangzhou University of Chinese Medicine (GZUCM Admin. Office [2024]278, [2025]252).

Statements and Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Third Affiliated Hospital of Guangzhou University of Chinese Medicine on March 25,2026 (Approval No.:

PJ-XS-20260325-009); given that this study is a retrospective study, the requirement for patient informed consent was waived.

Declaration of conflicting interest
The authors declare that there is no conflict of interest regarding the publication of this article.

Consent for publication
Not Applicable.

Data Availability Statement
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine provided the data needed to support the study's conclusions. The supplementary files are provided in the online version of the manuscript and are available from the corresponding author upon reasonable request, if required by the journal.

Reference

1. Duong, V., et al., Evaluation and Treatment of Knee Pain: A Review. JAMA, 2023. 330(16): p. 1568 – 1580.
2. Ouyang, Y. and M. Dai, Global, regional, and national burden of knee osteoarthritis: findings from the Global Burden of Disease study 2021 and projections to 2045. J Orthop Surg Res, 2025. 20(1): p. 766.
3. Global, regional, and national burden of osteoarthritis, 1990-2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Rheumatol, 2023. 5(9): p. e508 – e522.
4. Yao, Q., et al., Osteoarthritis: pathogenic signaling pathways and therapeutic targets. Signal Transduction and Targeted Therapy, 2023. 8(1): p. 56.
5. Duong, V., et al., Risk factors for the development of knee osteoarthritis across the lifespan: A systematic review and meta-analysis. Osteoarthritis Cartilage, 2025. 33(10): p. 1162 – 1179.
6. De Roover, A., et al., Fundamentals of osteoarthritis: Inflammatory mediators in osteoarthritis. Osteoarthritis Cartilage, 2023. 31(10): p. 1303 – 1311.
7. Binvignat, M., et al., The role of obesity and adipose tissue dysfunction in osteoarthritis pain. Nature Reviews Rheumatology, 2024. 20(9): p. 565 – 584.
8. Ostojic, M., et al., Blood and urine biomarkers for the diagnosis of early stages of knee osteoarthritis: A systematic review. J Exp Orthop, 2024. 11(3): p. e12105.
9. Ma, J., et al., Poor handling of continuous predictors in clinical prediction models using logistic regression: a systematic review. Journal of Clinical Epidemiology, 2023. 161: p. 140 – 151.
10. Efthimiou, O., et al., Developing clinical prediction models: a step-by-step guide. BMJ, 2024. 386: p. e078276.
11. Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5 – 32.
12. Imrie, F., R. Davis, and M. van der Schaar, Multiple stakeholders drive diverse

interpretability requirements for machine learning in healthcare. Nature Machine Intelligence, 2023. 5(8): p. 824 – 829.
13. Lekadir, K., et al., FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. Bmj, 2025. 388: p. e081554.
14. Lundberg, S.M. and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, I. Guyon, et al., Editors. 2017.
15. Lundberg, S.M., et al., From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020. 2(1): p. 56 – 67.
16. Padmanabhan, K., et al., Explainable machine learning in healthcare: methods, interpretation, and applications for clinical research. Journal of the American Medical Informatics Association, 2026. 33(8): p. 1510 – 1522.
17. Fan, Z., et al., XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: a population-based retrospective cohort study. Arthritis Research & Therapy, 2024. 26(1): p. 213.
18. Nielsen, R.L., et al., Data-driven identification of predictive risk biomarkers for subgroups of osteoarthritis using interpretable machine learning. Nature Communications, 2024. 15(1): p. 2817.
19. Loeser, R.F., J.A. Collins, and B.O. Diekman, Ageing and the pathogenesis of osteoarthritis. Nature Reviews Rheumatology, 2016. 12(7): p. 412 – 420.
20. Acharya, M., U. Mathur, and A. Dave, Evaluation of Corneas From Donors With Septicemia for Use in Corneal Transplant. Cornea, 2016. 35(12): p. e42 – e43.
21. Sanchez-Lopez, E., et al., Synovial inflammation in osteoarthritis progression. Nature Reviews Rheumatology, 2022. 18(5): p. 258 – 275.
22. Loukov, D., et al., Monocyte activation is elevated in women with knee-osteoarthritis and associated with inflammation, BMI and pain. Osteoarthritis Cartilage, 2018. 26(2): p. 255 – 263.
23. Zhao, X., et al., CCL3/CCR1 mediates CD14(+)CD16(-) circulating monocyte recruitment in knee osteoarthritis progression. Osteoarthritis Cartilage, 2020. 28(5): p. 613 – 625.
24. Raghu, H., et al., CCL2/CCR2, but not CCL5/CCR5, mediates monocyte recruitment, inflammation and cartilage destruction in osteoarthritis. Ann Rheum Dis, 2017. 76(5): p. 914 – 922.
25. Szymczak, A., et al., Multimodal Peripheral Blood Immunophenotyping and Single-Cell Epigenetic Analysis of Knee Osteoarthritis Reveals Expansion and Epigenetic Dysregulation of Treg Cell and Nonclassical Monocyte Populations. Arthritis & Rheumatology. n/a(n/a).
26. Costantini, S., et al., Inflaming and Immune-Resolving: The Ambivalent Role of Eosinophils in Osteoarthritis. Int J Mol Sci, 2025. 26(22).
27. Di, J., et al., Eosinophil-to-Lymphocyte Ratio and Eosinophil Count as New Predictive Markers for Osteoarthritis. J Pain Res, 2024. 17: p. 3803 – 3815.
28. Yao, P., et al., Distribution of blood eosinophil counts in the Chinese general population. Eur Respir J, 2026. 68(3).
29. Ortega, H., et al., Effects of systemic corticosteroids on blood eosinophil counts

in asthma: real-world data. J Asthma, 2019. 56(8): p. 808 – 815.
30. Canney, M., et al., Regional Variation in Hemoglobin Distribution Among Individuals With CKD: the ISN International Network of CKD Cohorts. Kidney Int Rep, 2023. 8(10): p. 2056 – 2067.
31. Williams, A.M., et al., Improving Anemia Assessment in Clinical and Public Health Settings. J Nutr, 2023. 153 Suppl 1(Suppl 1): p. S29 – s41.
32. Zhang, Q., et al., A Simple Nomogram for Predicting Osteoarthritis Severity in Patients with Knee Osteoarthritis. Comput Math Methods Med, 2022. 2022: p. 3605369.
33. Zhou, W., et al., The linkage between hemoglobin-to-red blood cell distribution width ratio and osteoarthritis: a study based on the national health and nutrition examination survey. Expert Rev Hematol, 2026. 19(5): p. 571 – 577.
34. Zhong, J. and G. Yang, Metabolic syndrome-osteoarthritis axis: Beyond mechanical loading toward a cardiometabolic OA phenotype. Biomed Pharmacother, 2026. 200: p. 119583.
35. Baudart, P., et al., Association between osteoarthritis and dyslipidaemia: a systematic literature review and meta-analysis. RMD Open, 2017. 3(2): p. e000442.
36. Xie, Y., et al., Metabolic syndrome, hypertension, and hyperglycemia were positively associated with knee osteoarthritis, while dyslipidemia showed no association with knee osteoarthritis. Clin Rheumatol, 2021. 40(2): p. 711 – 724.
37. Denoble, A.E., et al., Uric acid is a danger signal of increasing risk for osteoarthritis through inflammasome activation. Proc Natl Acad Sci U S A, 2011. 108(5): p. 2088 – 93.
38. Krasnokutsky, S., et al., Serum Urate Levels Predict Joint Space Narrowing in Non-Gout Patients With Medial Knee Osteoarthritis. Arthritis Rheumatol, 2017. 69(6): p. 1213 – 1220.
39. Go, D.J., et al., Serum uric acid and knee osteoarthritis in community residents without gout: a longitudinal study. Rheumatology (Oxford), 2021. 60(10): p. 4581 – 4590.
40. Chen, H., et al., Algorithms to estimate Shapley value feature attributions. Nature Machine Intelligence, 2023. 5(6): p. 590 – 601.
41. Tang, J., et al., Development and Validation of an Interpretable Machine Learning Model for Staging Helicobacter pylori-Initiated Intestinal-Type Gastric Cancer in the Correa Cascade: Cross-Sectional Study. J Med Internet Res, 2026. 28: p. e94837.
42. Carriero, A., et al., Explainable AI in healthcare: to explain, to predict, or to describe? Diagn Progn Res, 2025. 9(1): p. 29.
43. Hassanzad, M. and K. Hajian-Tilaki, Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Med Res Methodol, 2024. 24(1): p. 84.
44. Collins, G.S., et al., Evaluation of clinical prediction models (part 1): from development to external validation. Bmj, 2024. 384: p. e074819.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialKOA10.12.docx