Journal of Inflammation Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Developing an Interpretable Machine Learning Model for Early Prediction of Cardiovascular Involvement in Systemic Lupus Erythematosus

Zixian Deng[1], Huadong Liu[1], Feng Chen[2], Qiyun Liu[1], Xiaoyu Wang[3], Caiping Wang[1], Chuangye Lyu[1], Jianghua Li[1], Tangzhiming Li[1,4]

[1]Department of Cardiology, Shenzhen People's Hospital (The First Affiliated Hospital, Southern University of Science and Technology; The Second Clinical Medical College, Jinan University), Shenzhen, 518020, People's Republic of China; [2]Cardiology Department, Peking University First Hospital, Beijing, People's Republic of China; [3]Department of Cardiology, Long Gang Central Hospital of Shenzhen, Shenzhen, Guangdong, 518116, People's Republic of China; [4]Department of Cardiology, Heping country People's Hospital, Heyuan, Guangdong, People's Republic of China

Correspondence: Jianghua Li; Tangzhiming Li, Department of Cardiology, Shenzhen People's Hospital (The First Affiliated Hospital, Southern University of Science and Technology; The Second Clinical Medical College, Jinan University), Shenzhen, 518020, People's Republic of China, Email Lijianghua06@126.com; litangzhiming@126.com

Background: Cardiovascular disease is a leading cause of death in systemic lupus erythematosus (SLE). Early prediction of cardiac involvement is critical for improving patient outcomes. This study aimed to identify key factors associated with cardiac involvement in SLE and to develop an interpretable machine learning (ML) model for risk prediction.

Methods: We conducted a retrospective analysis of 1,023 SLE patients hospitalized in Shenzhen People's Hospital between January 2000 and December 2021, with a median age of 31 years at hospitalization (IQR: 25–39 years), 92.1% being female, and 18.77% developing cardiovascular involvement during a median follow-up of 3,737 days (IQR: 1,920–5,246). The most predictive features were selected through the intersection of three feature selection techniques: Random Forest, LASSO, and XGBoost. Models were trained on 70% of the dataset and tested on the remaining 30%. Among seven evaluated algorithms, the Gradient Boosting Machine (GBM) demonstrated the best performance on the test set. Model interpretability was assessed using the DALEX package, which generated feature importance plots and instance-level breakdown profiles to visualize decision-making logic.

Results: Over a median follow-up of 3737 days, 192 (18.77%) patients developed cardiac involvement. Seven key predictors—arthritis, hypertension, HDL-C, LDL-C, total cholesterol, CRP, and ESR— were identified from 51 clinical and biological variables at admission. The Gradient Boosting Machine (GBM) model (AUC: 0.748, Accuracy: 0.779, Precision: 0.605, F1 score: 0.433, recall 0.338) performed the best of the seven models.

Conclusion: This study is the first to develop an interpretable ML model to predict the risk of cardiac involvement in SLE. Notably, the GBM model showed optimal performance, and its interpretability allowed clinicians to visualize decision-making processes, facilitating early identification of high-risk patients.

Keywords: systemic lupus erythematosus, cardiovascular involvement, machine learning, prediction model, interpretability

Introduction

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease that affects multiorgans, with the cardiovascular system being one of the most commonly impacted.[1] While the use of immunosuppressive therapies has improved the prognosis and extended the lifespan of SLE patients, cardiovascular diseases (CVD) has emerged as a leading cause of death following infection and renal complications. This shift in mortality patterns has contributed to a significant global health burden.[2] Data from previous studies have shown that women aged 44–50 with SLE have a 50-fold increased risk of myocardial infarction compared to their healthy counterparts.[3] Additionally, a cohort study in the United States that included 252,676 SLE patients also showed that SLE was associated with an increased risk of CVD (OR42, 95% CI [1.40–1.44]).[4] Mounting evidence has suggested that the rate of cardiovascular morbidity and mortality is significantly

Received: 12 April 2025
Accepted: 14 June 2025
Published: 1 July 2025

higher in patients with SLE than the general population.[5,6] Epidemiological studies report a 2- to 3-fold increase in atherosclerosis, myocardial infarction, and heart failure among SLE patients compared to healthy controls.[7] In a 5-year follow-up study, carotid plaque was found in 32% of SLE patients, versus only 4% in healthy individuals,[8] indicating accelerated subclinical atherosclerosis. Furthermore, data from over 15,000 SLE patients show that greater disease severity correlates with higher CVD risk.[9] Therefore, it is crucial to identify risk factors for cardiovascular system involvement in SLE patients at an early stage to improve the prognosis.

Conventional tools like the Framingham Risk Score,[10] QRISK,[11] and SCORE[12] are widely used in the general population, but they overlook disease-specific factors in SLE such as inflammation, disease activity, and immunosuppressive treatment. This highlights the need for tailored and interpretable models in this population.

Despite these alarming trends, most research has focused on the increased risk of CVD in SLE,[13,14] with fewer studies investigating the specific risk factors involved.[15] Moreover, no predictive model has been developed to assess cardiovascular involvement in SLE patients. Early precision identification and treatment of those at high-risk for CVD remains a major challenge in managing this patient population. Machine learning (ML) algorithm is a discipline of artificial intelligence, offers a promising solution. ML algorithms can analyze vast amounts of clinical data, continuously refining their performance through iterative processes.[16] This approach can optimize the allocation of medical resources and enhance diagnostic accuracy and treatment efficiency. In recent years, more and more ML models have been widely applied in clinical settings.[17] However, traditional ML models remain "black boxes" without intuitive visualization tools to show the internal structure of the model and the decision-making process, limiting their application in the medical field. In clinical settings, interpretability is essential to ensure transparency, build clinician trust, and support decision-making. Models that can visually demonstrate why a prediction is made are more likely to be adopted in practice, particularly in complex conditions such as SLE. Enhancing the interpretability of ML models is therefore essential for their successful integration into medical practice. In light of these considerations, our study aims to develop and validate an interpretable ML model to predict the risk of cardiovascular system involvement in SLE patients.

Methods

Patients And study Design

This retrospective study was approved by the institutional ethics committee of Shenzhen People's Hospital (LL-KY -2023213-02) and strictly adhered to with the principles of the Declaration of Helsinki. As a retrospective medical record analysis, all data were anonymized and de-identified to eliminate privacy risks. The Ethics Committee granted a waiver of informed consent since the research posed no more than minimal risk to participants, involved no interventions, and utilized pre-existing data without compromising patient rights or confidentiality. However, 34 patients rejected participation in the project after initial enrollment. These patients' data were excluded from the analysis, ensuring that the study complied with ethical guidelines and maintained the integrity of the data.

This retrospective study included SLE patients who were hospitalized from January 2000 to December 2021 in Shenzhen People's Hospital. The large cohort (n=1023) and the 21-year follow-up period ensure a robust sample size and comprehensive representation of SLE patients. Additionally, while our data come from hospitalized patients, this subset represents a high-risk group with more severe disease, which makes the model especially relevant for identifying early cardiovascular involvement in high-risk SLE patients. All of the participants enrolled satisfied the following criteria: (1) Age ≥18 years; (2) fulfilled the revised 1997 American College of Rheumatology (ACR) for classification of SLE. Exclusion criteria were as follows: (1) development of CVD events before the diagnosis of SLE (eg, coronary artery disease, cardiomyopathy, valvular heart disease, congenital heart disease, arrhythmias, heart failure, etc).; (2) development of hypertension, diabetes mellitus, cerebral infarction, and severe hepatic or renal insufficiency prior to the diagnosis of SLE; (3) patients with multiple other autoimmune diseases (including Sjögren's syndrome, rheumatoid arthritis, overlapping syndrome; Overlap syndrome refers to at least two connective tissue diseases occurring at the same time or at different times in the same patient).; (4) combined endocrine and metabolic diseases, such as hyperthyroidism or hypothyroidism; (5) patients with malignant tumors. Patients who fulfilled the inclusion and exclusion criteria were randomly assigned to the training and validation sets in a 7:3 ratio. The flow chart of the study design is shown in Figure 1.

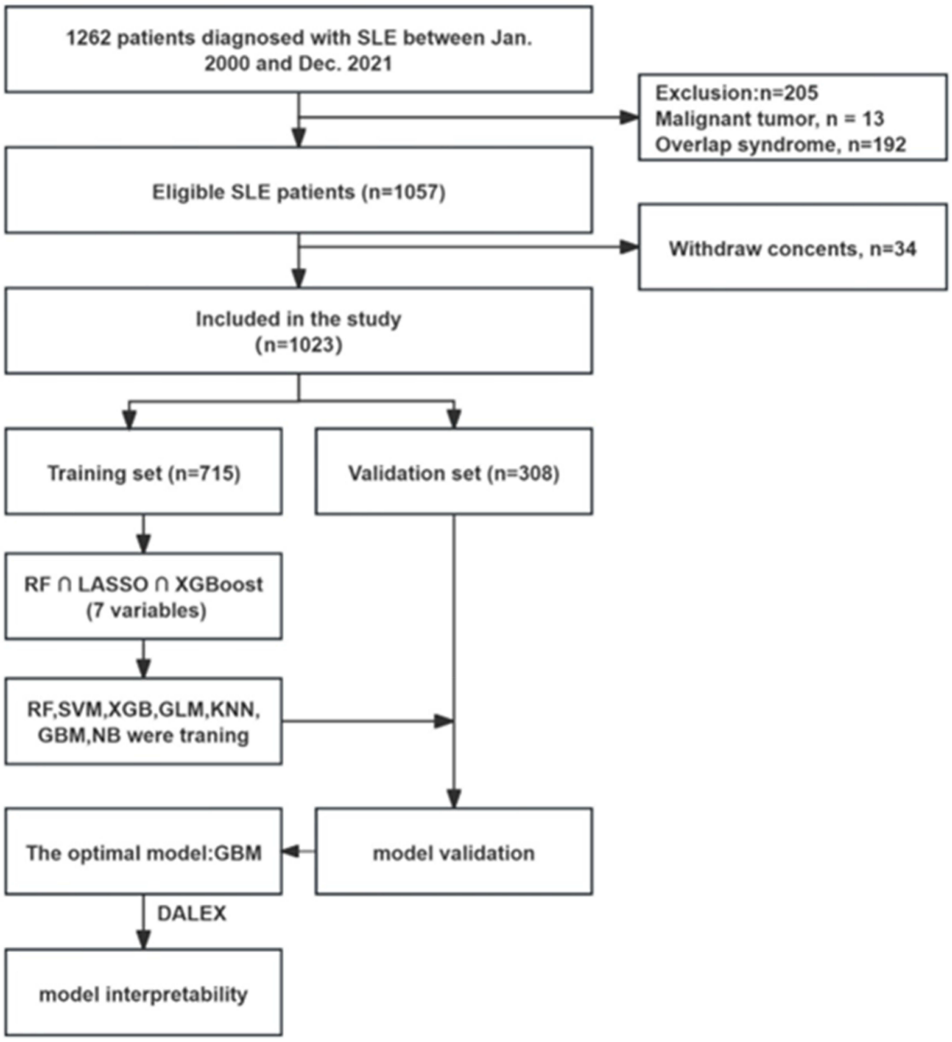

Figure 1 Flow diagram of study design.

Potential Predictive Variables

Potential predictive variables included patient characteristics at initial admission, including demographic characteristics, clinical features of SLE, cardiovascular risk factors, medications, and laboratory findings. Demographic characteristics encompassed age at diagnosis, age at hospitalization, fertility, and gender. The clinical features of SLE included systolic blood pressure (SBP), diastolic blood pressure (DBP) and heart rate (HR) on admission, hyperlipidemia, diabetes mellitus, hyperlipidemia, history of myocardial infarction (MI), family history of SLE, smoking history and chronic kidney disease (CKD). Renal involvement was defined as ACR criterion (persistent proteinuria >0.5 g/day or >3+ by urinalysis) and/or biopsy-proven lupus nephritis occurring after patients met ACR criteria, patients with renal involvement, as defined, before criteria diagnosis were excluded from these analyses. Blood system involvement included conditions such as anemia, thrombocytopenia, leukopenia, or lymphadenopathy associated with SLE, as determined by laboratory tests and clinical evaluations. In addition to these, Raynaud phenomenon, alopecia, malar rash, discoid rash, photosensitivity, oral ulcers, arthritis, serositis and organ involvements were recorded. Medications referred to hydroxychloroquine, immunosuppressants and glucocorticoids. Laboratory findings included white blood cell count (WBC), platelets (PLT), C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), creatinine (CR), triglycerides (TG), total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), complement 3 (C3), complement 4 (C4) and autoimmune antibodies such as antinuclear antibodies (ANA), anti-double-stranded

DNA (Anti-dsDNA), anti-Smith (Anti-Sm), anti-SSA/Ro60kd, anti-SSA/Ro52kd, anti-SSB/La, anti-P0, anti-U1snRNP, anti-nucleosome, anti-histone, and anti-RO52.

Outcome of Interest

The primary outcome of this study was cardiovascular involvement, which was assessed using electrocardiography (ECG), echocardiography, coronary computed tomography angiography (CTA), or coronary angiography and based on comprehensive clinical criteria, including pericardial effusion, valvular heart disease, ventricular abnormalities, arrhythmias, pulmonary artery hypertension, and other major complications. These outcomes were analyzed as categorical variables, dichotomized by presence or absence, to simplify interpretation and ensure consistency across analyses. Cardiovascular involvement in SLE was defined in this study as meeting either one or more of the following criteria:

1. **Pericardial effusion**: Identified through echocardiography.
2. **Valvular heart disease**: Structural or functional valve abnormalities detected by echocardiography, excluding physiological regurgitation.
3. **Ventricular abnormalities**: Echocardiographic evidence of abnormal ventricular wall movement or hypertrophy.
4. **Arrhythmia events**: Includes conduction disorders, atrial fibrillation, atrial flutter, ventricular tachycardia, supraventricular tachycardia, and prolonged QT syndrome.
5. **ST-T changes**: ST-segment or T-wave abnormalities observed on a resting 12-lead ECG in two or more contiguous leads.
6. **Pulmonary artery hypertension**: Pulmonary artery systolic pressure of $\geq$ 40 mm Hg as measured by 2D echocardiography.
7. **Other forms of heart damage**: Includes heart failure, atherosclerotic heart disease, and myocardial infarction.

These complications are consistent with those reported in previous studies of SLE patients and represent the most common forms of cardiovascular involvement in this population.[18] Most importantly, this study of cardiovascular system involvement in SLE has excluded congenital heart disease and cardiac manifestations that could be explained by other diseases. All patients were followed from the date of diagnosis until either the study endpoint or the last follow-up date (March 9, 2022) if no cardiovascular event occurred.

Variable Selection

Selecting variables is important for interpreting and predicting, especially in high-dimensional datasets. Three ML algorithms (RF, LASSO, XGBoost) were employed to identify optimal predictors, because they offer complementary strengths: RF provides ensemble-based importance rankings, LASSO performs sparse regularization to reduce overfitting, and XGBoost captures nonlinear interactions with high predictive accuracy. The intersection of their outputs ensures robustness and generalizability of the selected predictors. To address the "black-box" nature of ML models, we applied the DALEX package, specifically Break Down Profiles, to enhance model interpretability and allow visualization of feature contributions to individual predictions. First, we utilized the inherent ability of RF to rank features based on their importance to classification accuracy. Then, Next, the LASSO regression algorithm was performed to compress the unimportant feature coefficients to zero, so as to select the features that were strongly associated with cardiovascular system involvement. Furthermore, the XGBoost algorithm, which is based on a gradient-boosting decision tree algorithm, was adapted to evaluate the importance of each variable during model construction and identify critical variables through feature importance scores.

Receiver operating characteristic (ROC) curves were used to evaluate the predicting capability of the predictors selected by three algorithms in the test set. The intersection of features selected by RF, LASSO, and XGBoost was based on robust statistical methods to ensure the identification of key predictors. We acknowledge the importance of factors like disease severity, comorbidities, and additional biomarkers such as homocysteine, but opted not to include them in this model. Our choice was based on the available data and the need to maintain model interpretability and clinical applicability. Additionally, the selected predictors—lipid profiles and inflammation markers—have strong clinical evidence supporting their association with CVD in SLE, making them essential for predicting cardiovascular involvement.

Prediction Model Development, Validation and Explainability

The eligible patients were randomized at a 7:3 ratio into a training set (n=715) and a test set (n=308). Subsequently, the best predictors, which were selected by RF, Lasso and XGBoost, were exploited to develop the model using seven ML approaches on the training set. These ML methods included RF, Support Vector Machines (SVM), XGBoost, generalized linear model (GLM), K-nearest neighbor (KNN), Gradient Boosting Machine (GBM), and Naive Bayes (NB). For a deeper understanding of predictive process of the model and the influencing elements, we used the DALEX package to perform an interpretive analysis of the model. It provided a comprehensive assessment of model performance, including metrics such as AUC, precision, accuracy, recall, and F1 score, and plotted the reserve cumulative distribution of residuals and the box plot of residuals to get the best model. Subsequent the feature importance chart was further drawn to visually show the extent to which each feature influenced the prediction results. The contribution of each variable was calculated based on the optimal performance model, demonstrating how these variables collectively affect the model's prediction of the odds of cardiovascular involvement for each SLE patient.

Statistical Analysis

All statistical analyses were carried out using R software (version 4.2.0). Due to incomplete medical records and missing data, the missing data in this study was approximately 1%. Before the analysis, the missing values were 10-fold imputed using the "chained equation" multiple interpolation strategy. Although the missing data accounted for less than 1%, we opted to apply multiple imputation using the "chained equation" strategy to minimize potential bias and ensure robust model performance. This conservative approach helps maintain the integrity of the dataset without introducing any systematic errors that could arise from excluding cases with missing data. Categorical data were presented as numbers and percentages and were compared using the chi-squared test or Fisher's exact test. Continuous variables were expressed as means ± standard deviation (SD) or medians (interquartile ranges (IQRs)) and were analyzed using a *t*-test or Mann–Whitney *U*-test. In addition, the "randomForest" package in R software (version 4.2.0) was used for random forest, the "glmnet" package was used for LASSO analysis, "xgboost" to implement the XGBoost model, the svm function in the "e1071" package to implement the SVM model, the glm function in the "stats" package to implement the GLM model, the knn function in the "class" package to implement the KNN model, the "gbm" package to implement the GBM model, the naiveBayes function in the "e1071" package to implement the NB model. The "DALEX" package was used to interpret the model and the "pROC" package was used for ROC curve analysis. The figures were processed in Adobe Illustrator 2021. A *P*-value of less than 0.05 was considered significant (*$P < 0.05$).

Results

Baseline Characteristics

A total of 1023 patients were enrolled in the study and randomly divided into a training set (n=715) and a test set (n=308) at a ratio of 7:3. A detailed flowchart is presented in Figure 1. The demographics and clinical features of the training and test set are summarized in Supplementary Table 1. The results demonstrate no statistically significant difference in data between the two sets except for WBC and CRP. Over a median follow-up period of 3,737 days (IQR: 1,920–5,246), 192 patients (18.77%) with SLE developed cardiovascular involvement. As shown in Table 1, SLE patients with cardiovascular involvement had significantly higher rates of diabetes, chronic kidney disease (CKD), and triglyceride (TG) levels, along with more severe SLE symptoms compared to those without cardiac involvement. Interestingly, there was no significant difference in the prevalence of hyperlipidemia between the two groups, despite its known association with cardiovascular risk in other populations. Patients who experienced cardiovascular involvement had higher heart rates, CRP, ESR, creatinine (CR), and triglyceride (TG) levels at admission compared to those without such involvement. Additionally, these patients showed a higher incidence of hypertension, diabetes mellitus, myocardial infarction (MI) post-SLE diagnosis, chronic kidney disease (CKD), Raynaud phenomenon, arthritis, serositis, other systemic involvements (such as renal, neurologic, and hematologic), and Anti-nucleosome positivity. Conversely, patients with cardiovascular system involvement had a lower incidence of malar rash, C4, HDL-C, and LDL-C. Baseline clinical characteristics differences are described in Table 1.

Table 1 The Demographics and Clinical Characteristics of the SLE Patients with or Without Cardiac Involvement

Parameter	Overall	Cardiac Involved	Non-Cardiac Involved	p
n	1023	192(18.77)	831(81.23)	
Age at diagnosis, year	28.00 [23.00, 37.00]	29.00 [24.00, 39.00]	28.00 [23.00, 36.00]	0.096
Age at hospitalization, year	31.00 [25.00, 39.00]	31.00 [25.00, 40.25]	31.00 [25.00, 39.00]	0.313
Fertility (%)	621 (60.70)	126 (65.62)	495 (59.57)	0.14
Male (%)	81 (7.92)	18 (9.38)	63 (7.58)	0.458
SBP, mmHg	110.00 [100.00, 122.00]	111.50 [99.75, 128.00]	109.00 [101.00, 121.00]	0.294
DBP, mmHg	74.00 [66.00, 82.00]	76.00 [67.00, 84.00]	73.00 [66.00, 82.00]	0.082
Heart rate, bpm	88.00 [80.00, 100.00]	95.00 [80.00, 106.00]	87.00 [80.00, 98.00]	<0.001*
Smoking (%)	16 (1.56)	3 (1.56)	13 (1.56)	1
Hypertension(%)	171 (16.72)	63 (32.81)	108 (13.00)	<0.001*
Hyperlipidemia (%)	78 (7.62)	17 (8.85)	61 (7.34)	0.454
Diabetes mellitus(%)	29 (2.83)	15 (7.81)	14 (1.68)	<0.001*
The history of MI (%)	15 (1.47)	6 (3.12)	9 (1.08)	0.045*
Family history of SLE (%)	10 (0.98)	2 (1.04)	8 (0.96)	1
CKD (%)	37 (3.62)	16 (8.33)	21 (2.53)	<0.001*
Raynaud phenomenon (%)	121 (11.83)	34 (17.71)	87 (10.47)	0.009*
Alopecia (%)	316 (30.89)	55 (28.65)	261 (31.41)	0.489
Malar rash (%)	469 (45.85)	63 (32.81)	406 (48.86)	<0.001*
Discoid rash (%)	148 (14.47)	19 (9.90)	129 (15.52)	0.052
Photosensitivity (%)	119 (11.63)	25 (13.02)	94 (11.31)	0.532
Oral ulcers (%)	605 (59.14)	102 (53.12)	503 (60.53)	0.062
Arthritis (%)	262 (25.61)	128 (66.67)	134 (16.13)	<0.001*
Serositis (%)	304 (29.72)	82 (42.71)	222 (26.71)	<0.001*
Renal involvement (%)	93 (9.09)	28 (14.58)	65 (7.82)	0.005*
Neuropsychiatric lupus (%)	299 (29.23)	72 (37.50)	227 (27.32)	0.006*
Blood system involvement (%)	97 (9.48)	35 (18.23)	62 (7.46)	<0.001*
Interstitial lung disease (%)	967 (94.53)	178 (92.71)	789 (94.95)	0.22
WBC, 10^9/L	5.10 [3.58, 7.12]	5.44 [3.70, 7.78]	5.09 [3.57, 7.00]	0.147
Platelets, 10^9/L	196.00 [137.00, 252.50]	181.00 [122.00, 273.50]	197.00 [143.00, 252.00]	0.701
CRP, mg/L	3.08 [1.00, 10.62]	4.70 [2.35, 18.75]	2.44 [0.90, 8.53]	<0.001*
C3(g/L)	0.66 [0.43, 0.90]	0.60 [0.35, 0.90]	0.67 [0.44, 0.90]	0.061
C4(g/L)	0.12 [0.06, 0.19]	0.10 [0.05, 0.19]	0.12 [0.06, 0.19]	0.046*
Creatinine, mmol/L	60.00 [52.00, 71.00]	62.50 [53.00, 76.25]	59.00 [51.00, 69.40]	0.003*
TG, mmol/L	1.19 [0.84, 1.77]	1.35 [0.98, 2.13]	1.16 [0.80, 1.71]	<0.001*
TC, mmol/L	4.10 [3.35, 4.92]	4.00 [3.23, 4.81]	4.12 [3.36, 4.95]	0.128
HDL-C, mmol/L	1.01 [0.76, 1.32]	0.87 [0.63, 1.13]	1.05 [0.81, 1.33]	<0.001*
LDL-C, mmol/L	2.27 [1.76, 2.88]	2.08 [1.68, 2.77]	2.29 [1.79, 2.90]	0.05*
ESR, mm/H	38.00 [16.00, 67.00]	50.00 [24.00, 84.00]	36.00 [15.00, 62.00]	<0.001*
ANA (%)	886 (86.61)	170 (88.54)	716 (86.16)	0.413
Anti-dsDNA (%)	569 (55.62)	111 (57.81)	458 (55.11)	0.52
Anti-Sm (%)	394 (38.51)	84 (43.75)	310 (37.30)	0.101
Anti-SSA/Ro60kd (%)	614 (60.02)	114 (59.38)	500 (60.17)	0.87
Anti-SSA/Ro52kd (%)	537 (52.49)	111 (57.81)	426 (51.26)	0.109
Anti-SSB/La (%)	221 (21.60)	46 (23.96)	175 (21.06)	0.382
Anti-P0 (%)	180 (17.60)	28 (14.58)	152 (18.29)	0.248
Anti-U1snRNP (%)	462 (45.16)	99 (51.56)	363 (43.68)	0.053
Anti-nucleosome (%)	379 (37.05)	91 (47.40)	288 (34.66)	0.001*
Anti-histone (%)	285 (27.86)	65 (33.85)	220 (26.47)	0.049*
Anti-RO52 (%)	173 (16.91)	36 (18.75)	137 (16.49)	0.455

Notes: Data is depicted as a number (percentage, %), the median ± standard deviation or the median (interquartile range). *=P<0.05, statistically significant difference.SBP, systolic blood pressure; DBP, diastolic blood pressure; CKD, chronic kidney disease; MI, myocardial infarction; WBC, white blood cell count; CRP, C-reactive protein; C3, complement 3; C4, complement 4.

Abbreviations: TG, triglycerides; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; ESR, erythrocyte sedimentation rate; ANA, antinuclear antibodies.

Feature Selection

We collected 51 clinical and biological variables at the time of patient admission. To improve the diagnostic performance of the model, three popular machine learning algorithms, RF, LASSO, and XGBoost, were used to select features for predicting cardiovascular involvement. RF was used to rank the importance of all 51 feature variables, and the top 15 feature variables were selected to construct models(Figure 2A). 28 potential predictors with non-zero coefficients were selected in the LASSO logistic regression model (Figure 2B). The XGBoost model identified 15 features with high importance scores as the optimal predictors (Figure 2C). The selection of lipid and inflammation markers was driven by strong evidence linking these factors with cardiovascular risk in SLE patients. Although factors such as disease severity and comorbidities are important, they were not included in the final model to maintain focus on the most significant and actionable clinical parameters. The robust predictive performance of the model (AUCs: 0.803, 0.780, and 0.799) demonstrates the utility of the selected features for early cardiovascular risk assessment in SLE patients (Figures 2D–F). Subsequently, a Venn diagram was used to determine the intersecting feature sets from the three algorithms (Figure 2G). This process identified a common set of seven best predictors: arthritis, hypertension, HDL-C, TC, LDL-C, ESR, and CRP.

Model Performance Comparisons and Explainability

Using the selected optimal predictors, we trained seven different models: Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB), Generalized Linear Model (GLM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), and Naive Bayes (NB). We used both the confusion matrix and the DALEX package to evaluate model performance to provide complementary perspectives. While the confusion matrix offers a straightforward evaluation of classification performance, including precision, recall, and F1 score, the DALEX package was used for a deeper interpretability analysis, allowing us to understand feature contributions and model residuals. Together, these methods provide a comprehensive assessment of the model's predictive power and interpretability, ensuring both performance and clinical relevance. The results showed that the XGBoost model has the largest AUC (0.753), followed by the GBM (0.748) and GLM (0.745) models, respectively (Figure 2H). However, the accuracy of GLM (0.786) is higher than GBM (0.779) and XGBoost (0.776).

To better visualize these results, a heat map of the performance metrics was created (Figure 3F). The heat map revealed that the predictive capabilities of the models were quite similar and difficult to distinguish. To identify the optimal model, we used the DALEX package to calculate residuals for each model and further analyze their performance. The residual distributions for each model were plotted to identify the best-performing model. As shown in Figure 3A and B, the GBM model demonstrated the most consistent performance across various metrics and was selected as the optimal model for its balance of precision, recall, and interpretability. Figure 3C and D showed histograms of residuals and precision-recall curves. The importance ranking of the selected features is displayed in Figure 3E, with the features ranked in descending order of importance as arthritis, hypertension, HDL-C, LDL-C, CRP, ESR and TC. To illustrate model prediction, we presented examples of two patients from the dataset. However, validation testing was conducted using a dedicated test set, achieving an AUC of 0.748, accuracy of 0.779, and precision of 0.605. These results demonstrate robust performance. One data was from patients with cardiovascular system involvement, with a predicted probability of 0.817 (Figure 4A) and the other without, with a predicted probability of 0.132 (Figure 4B).

Discussion

Early identification and control of risk factors for cardiovascular system involvement in SLE are crucial for improving patient outcomes. Through using three popular algorithms (RF, LASSO and XGBoost), we identified seven optimal predictors out of 51 variables. These predictors were then used to develop seven ML models to predict cardiovascular involvement in SLE patients. Among these models, the GBM model demonstrated the best predictive performance. Feature importance analysis revealed that the most influential variables in the GBM model, in descending order, were arthritis, hypertension, HDL-C, LDL-C, CRP, ESR, and TC. The model's interpretability further clarified how these clinical features contributed to individual patient outcomes. Cardiovascular system involvement injury often goes unrecognized in its early stages, making it challenging to detect cardiac damage until serious complications arise.

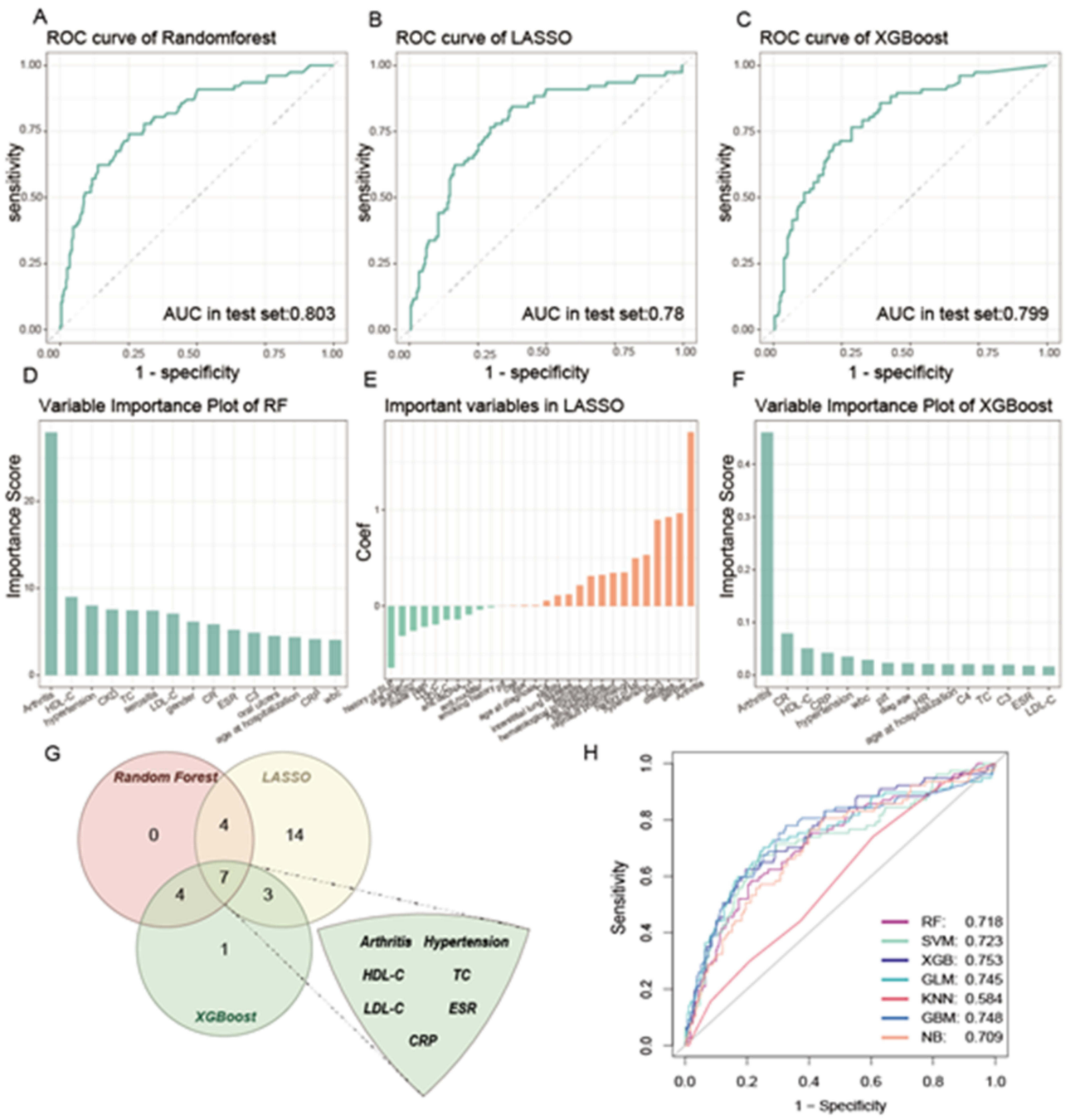

Figure 2 Feature screening and performance of models. (**A**) The variable importance plot of RF algorithms. (**B**) The important variables in LASSO algorithms. (**C**) The variable importance plot of XGBoost algorithms. (**D**) ROC of the RF model in the test set. (**E**) ROC of the LASSO model in the test set. (**F**) ROC of the XGBoost model in the test set. (**G**) Wayne diagram showing the intersections of the variables of the three algorithms. (**H**) ROC of the seven models in the test set.

Therefore, early prediction and diagnosis of patients with SLE who have not yet developed cardiac symptoms are crucial for improving long-term prognosis.

Currently, this study focuses on developing a specific risk prediction model for cardiovascular involvement in SLE patients, leveraging machine learning methods. While tools such as PREVENT ASCVD are well-established for predicting atherosclerotic cardiovascular disease in the general population, they may not adequately capture the unique risk profile of SLE patients, such as inflammation-driven mechanisms, autoimmune activity, and medication effects (eg, corticosteroids). A direct comparison with PREVENT ASCVD[19] was not performed in this study as our model

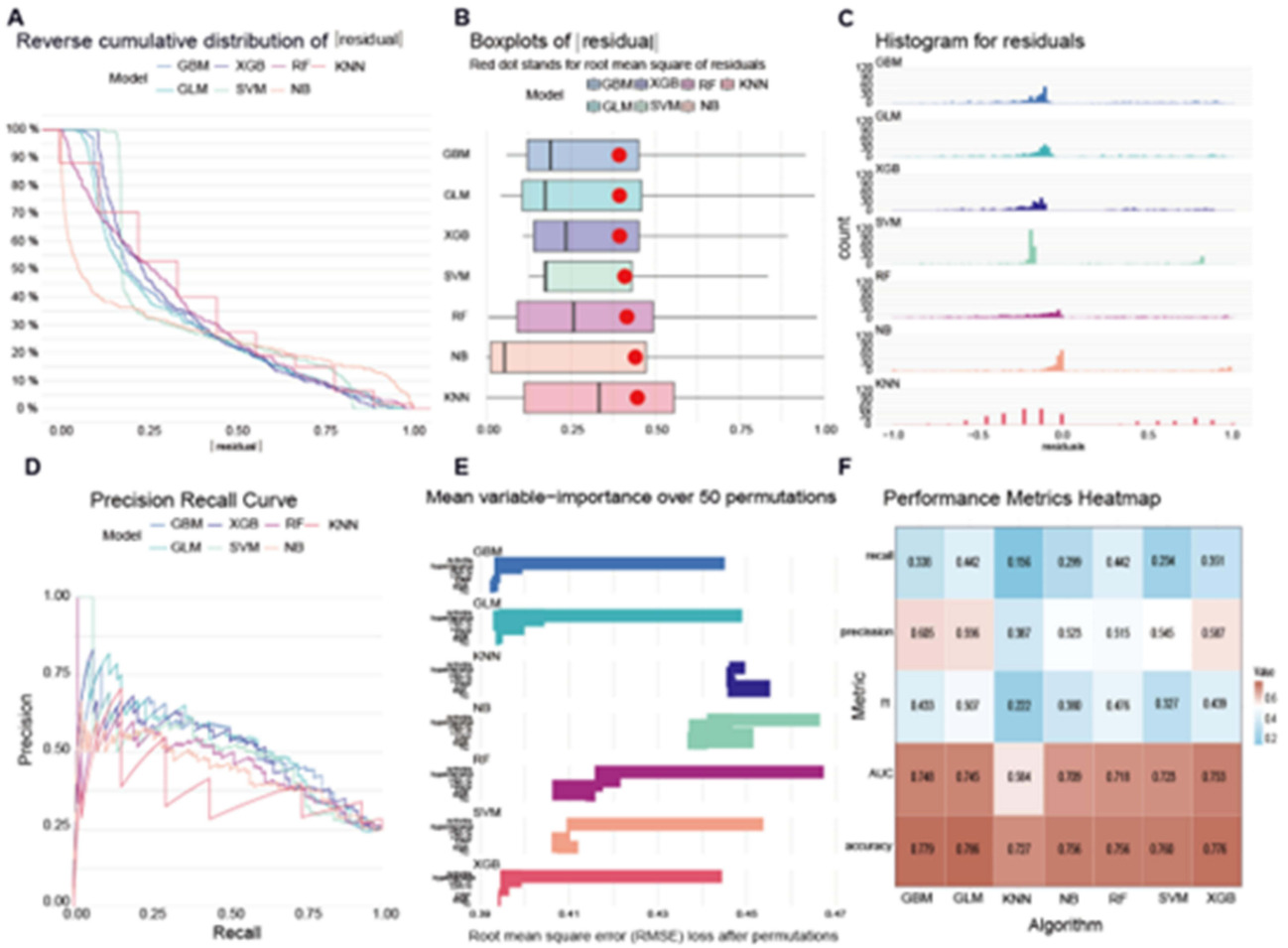

Figure 3 The model interpretation. (**A**) Reverse cumulative distribution of residuals of seven models. (**B**) Residual boxplot of seven models. (**C**) Histograms for residuals of seven models. (**D**) Precision-recall curve of seven models. (**E**) Feature importance plots of seven models. (**F**) Performance metrics heatmap of seven models.
Abbreviations: RF, random forest; SVM, support vector machines; XGBoost, Xtreme Gradient Boosting; GLM, generalized linear model; KNN, K-nearest neighbor; GBM, gradient boosting machine; NB, Naive Bayes.

specifically incorporates SLE-specific variables (eg, CRP, ESR, and arthritis), which are absent in generic risk calculators. Future studies could explore the performance of our model relative to PREVENT ASCVD or similar tools in both SLE and general populations. As far as we know, no risk prediction model for cardiovascular involvement in SLE has been developed until now. This study provides a visual ML tool that can estimate the risk of cardiovascular involvement in SLE patients based on demographic and clinical information collected at admission. To our knowledge, this is the first interpretable ML risk prediction model about cardiovascular system involvement in SLE.

ML models have made significant strides in clinical applications. Although many previous studies have demonstrated excellent performance from ML models, the internal decision-making processes of these models often remain obscure, commonly referred to as the "black box model" problem.[20] His lack of transparency means that clinicians are unable to fully understand how the model extracts features from input data, assigns weights, and ultimately generates predictions. To address this issue, we used the DALEX package to build a model predicting the cardiovascular system involvement of SLE. This approach allows users to visualize the prediction process, assess the importance of features, and evaluate the bias in the analysis model, thereby improving both the interpretability and transparency of the model. Furthermore, the prediction results and the internal mechanisms of the model can be displayed through intuitive graphs and visualizations, making the interpretation process clearer and more accessible for clinicians. Although the GBM model achieved an AUC of 0.748, which is acceptable for clinical risk stratification, the moderate performance may be attributed to several

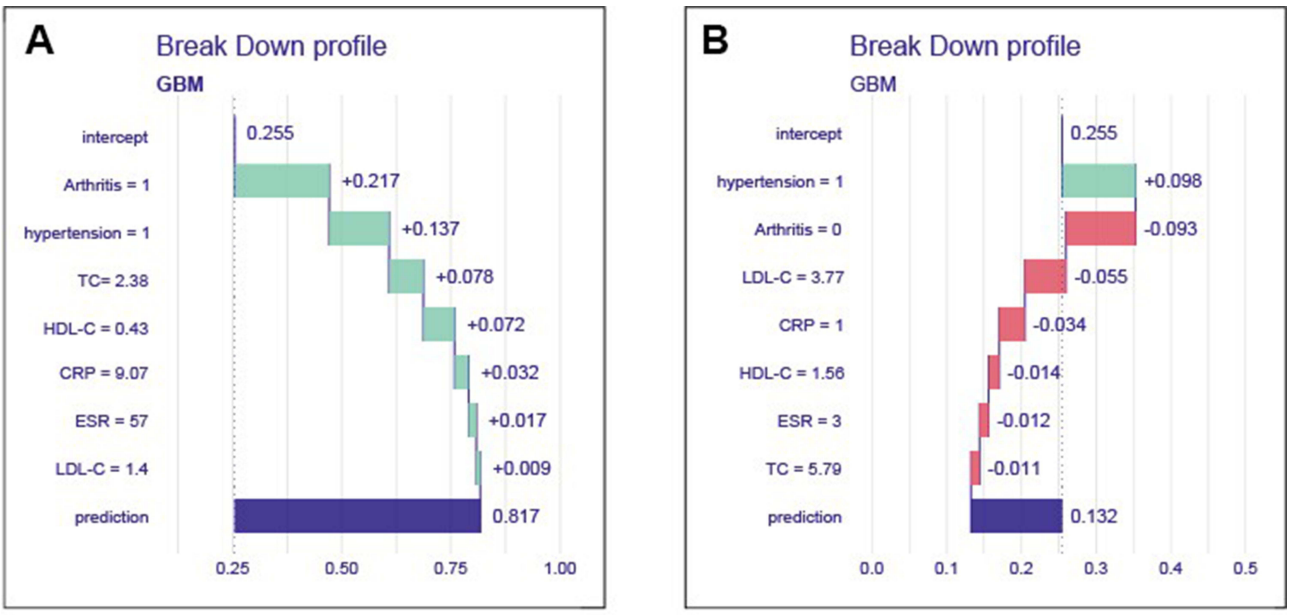

Figure 4 This figure was made with the DALEX package to explain Gradient Boosting Machine model predictions. The numbers on the X-axis indicate the predicted probability of cardiovascular involvement in SLE obtained from the model's estimation based on a patient's features. (**A**) The prediction results of one SLE patient with cardiac involvement. (**B**) The prediction results of one.

Abbreviations: SLE, patient without cardiac involvement; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol.

factors, including the absence of formal disease activity indices (eg, SLEDAI), treatment history, and some biomarkers (eg, homocysteine). Incorporating these parameters in future prospective studies may enhance prediction accuracy.

In this study, the feature importance analysis identified that arthritis as the most significant risk factor for cardiovascular system involvement. Early studies have demonstrated that inflammatory arthritis was associated with increased risk of cardiovascular mortality and morbidity, independent of traditional risk factors such as hypertension and hypercholesterolemia.[21,22] It was found that the risk of CVD in case of autoimmune disease was further increased when arthritis was present.[22] A meta-analysis showed that patients with arthritis had a 48% increased risk of developing CVD, with a 68% increased risk of myocardial infarction.[23] Therefore, clinicians should be highly alert to the risk of cardiovascular events in patients with SLE complicated with arthritis.

Mechanistically, chronic joint inflammation may promote systemic endothelial dysfunction[24] through elevated levels of pro-inflammatory cytokines such as tumor necrosis factor-alpha (TNF-α),[25] interleukin-6 (IL-6),[26] and C-reactive protein (CRP).[27] These mediators contribute to vascular injury, impaired lipid metabolism, and accelerated atherogenesis. Moreover, arthritis in SLE often reflects systemic immune activation, which may amplify cardiovascular risk through shared autoimmune and inflammatory pathways.[28]

Traditional cardiovascular risk factors such as hypertension and dyslipidemia have long been associated with risk factors for cardiovascular diseases in the general population.[29] Consistent with previous findings, in this study, hypertension was the second most important risk factor for cardiovascular system involvement in SLE patients. Intriguingly, a paradoxical association—commonly referred to as the "lipid paradox"—was observed between lipid levels and cardiovascular involvement in SLE patients, as illustrated in Figure 4. Specifically, a higher risk of cardiovascular system involvement was linked to lower LDL-C and TC levels, and conversely, higher HDL-C levels. Accumulating evidence suggests enhanced risk of cardiovascular disease in chronic systemic inflammation combined with low LDL-C and low TC environments.[30] For instance, Myasoedova et al reported a non-linear relationship between TC and CVD risk in RA, with risk increasing threefold when TC < 4 mmol/L.[31] While the "lipid paradox" is well-documented in RA, it may manifest differently in SLE due to distinct immunopathological features. In our preliminary unpublished findings, we observed that this paradox may be linked to disruptions in HDL/LDL homeostasis, involving

a shift toward pro-inflammatory lipid profiles. Chronic inflammation may impair the antioxidant capacity of HDL, transforming it into a pro-atherogenic particle.[32] These observations underscore the need for mechanistic investigations into lipid metabolism under chronic autoimmune inflammation. They also challenge the conventional assumption that "the lower the LDL-C, the better", highlighting the importance of context-specific lipid management in SLE.

Beyond traditional risk factors, numerous preclinical and clinical studies indicate that inflammation is involved throughout the development of atherosclerosis and related cardiovascular events.[33] Inflammatory and immune mechanisms have been proposed that can independently impact the relationship between SLE and atherosclerosis.[34,35] CRP and ESR, which are markers of the systemic inflammatory response, are commonly elevated in SLE as a sign of a hyperinflammatory response. It has been shown that a state of high inflammatory burden reflected by an ESR ≥30 mm/h and a CRP ≥3 mg/dl is associated with a 2-fold and 5-fold increased risk for future MACE, respectively.[36] Our result is consistent with those of previous studies demonstrating that positively linked to the risk of CVD. These findings emphasize the importance of testing for CRP and ESR which can be helpful in early detection and stratifying risk for cardiovascular system involvement in SLE.

Our cohort primarily consisted of hospitalized patients, which inherently reflects a higher disease severity. While this may limit direct comparisons to outpatient SLE populations, it allows for more accurate prediction in higher-risk patients who are more likely to develop cardiovascular involvement. The model's focus on identifying high-risk patients early is an essential feature that addresses this issue. The decision to collect predictive variables at the time of admission, rather than at first SLE diagnosis, was based on the clinical relevance of variables that are directly associated with the risk of cardiovascular involvement. While we acknowledge that early SLE disease activity and comorbidities at diagnosis are important, data collected at admission are more reflective of the patient's current disease status and cardiovascular risk. This approach enhances the practical applicability of the model in real-world clinical settings, where timely interventions are essential.

This retrospective study included SLE patients who were hospitalized from January 2000 to December 2021 in Shenzhen People's Hospital, a tertiary care center in southern China. While the single-center design provided a homogeneous cohort for analysis, it may limit the generalizability of the findings to other populations or settings. Future studies should include multi-center data and diverse populations to validate the robustness and applicability of the predictive model.

In this study, we identified predictors of cardiovascular involvement in SLE and developed a risk prediction model using multiple ML algorithms. In addition, the DALEX package was employed to explain the decision-making process of the best model, helping to visualize and explain the working mechanism of the complex model clinically. The simplicity of input variables—routine lipid and inflammation markers—makes the model potentially suitable for outpatient clinics and low-resource settings. However, practical barriers include limited access to digital infrastructure, lack of clinician training in ML interpretation, and regulatory concerns. Future work should focus on developing user-friendly interfaces and real-world validation in diverse care settings.

However, this study has several limitations. The retrospective design allowed for the analysis of a large, long-term cohort spanning over 21 years, it may introduce potential biases related to data collection and patient selection. To minimize these biases, strict inclusion and exclusion criteria were applied, and data were cross-validated for consistency. Future prospective studies are warranted to validate the findings and further mitigate potential biases. Besides, this study did not include formal disease activity indices such as SLEDAI, due to the unavailability of complete scoring data across the cohort. Instead, individual markers of disease activity (eg, CRP, ESR, renal involvement) were used as proxies. The lack of standardized scoring may limit the granularity and clinical interpretability of the model. Future prospective studies should incorporate validated disease activity indices to enhance predictive robustness and improve applicability to clinical decision-making. Furthermore, the lack of external validation in independent cohorts limits the generalizability of the model. Future studies should focus on external validation in multi-center cohorts to assess its applicability in diverse clinical settings and patient populations. Last but not the least, we acknowledge that residual confounders may still exist. Variables such as disease activity score, treatment history (eg, corticosteroids and immunosuppressants), and socioeconomic factors were not comprehensively included in this study due to data limitations. Future studies should aim to incorporate a broader range of potential confounders to ensure the robustness of the findings.

Conclusion

In conclusion, this study successfully developed a valuable risk prediction model for cardiovascular involvement in SLE, with the Gradient Boosting Machine (GBM) model identified as the optimal choice. The use of the DALEX package enhanced the interpretability of the model, providing tools for personalized risk assessment. Key predictors in the model included arthritis, hypertension, HDL-C, LDL-C, CRP, ESR, and TC. The findings of this study can be helpful in efficiently and rapidly identifying a high-risk target population and provide an early warning sign for clinical practice.

Data Sharing Statement

The data utilized and analyzed in the present study are accessible from the corresponding author upon justified request.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was supported by the National Natural Science Foundation of China (Grants number 82000058 and 82070517), the Science and Technology Planning Project of Shenzhen City (JCYJ20240813104300002), GuangDong Basic and Applied Basic Research Foundation (2025A1515012768), Natural Science Foundation of Shenzhen (Grants number JCYJ20190807145015194), and Shenzhen People's Hospital Research Cultivation Project (Grants number SYJCYJ202014 and SYLCYJ202119), Sanming Project of Medicine in Shenzhen (Grants number No. SZSM201412012), Shenzhen Key Medical Discipline Construction Fund (Grants number No. SZXK003), the Medical Scientific Research Foundation of Guangdong Province of China (Grants number no. A2018530), and the Science and Technology Planning Project of Shenzhen Municipality (Grants number no. JCYJ20190806153207263).

Disclosure

The authors report no conflicts of interest in this work. This paper has been uploaded to "SSRN" as a preprint:https://ssrn.com/abstract=4941311.

References

1. Kiriakidou M, Ching CL. Systemic lupus erythematosus. *Ann Internal Med*. 2020;172:Itc81–itc96. doi:10.7326/AITC202006020
2. Carter EE, Barr SG, Clarke AE. The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat Rev Rheumatol*. 2016;12:605–620. doi:10.1038/nrrheum.2016.137
3. Manzi S, Meilahn EN, Rairie JE, et al. Age-specific incidence rates of myocardial infarction and angina in women with systemic lupus erythematosus: comparison with the Framingham Study. *Am J Epidemiology*. 1997;145:408–415. doi:10.1093/oxfordjournals.aje.a009122
4. Katz G, Smilowitz NR, Blazer A, Clancy R, Buyon JP, Berger JS. Systemic lupus erythematosus and increased prevalence of atherosclerotic cardiovascular disease in hospitalized patients. *Mayo Clin Proc*. 2019;94:1436–1443. doi:10.1016/j.mayocp.2019.01.044
5. Broadhead RS, Facchinetti NJ. Clinical clerkships in professional education: a study in pharmacy and other ancillary professions. *Soc Sci Med*. 1985;20:231–240. doi:10.1016/0277-9536(85)90236-9
6. Björnådal L, Yin L, Granath F, Klareskog L, Ekbom A. Cardiovascular disease a hazard despite improved prognosis in patients with systemic lupus erythematosus: results from a Swedish population based study 1964-95. *J Rheumatol*. 2004;31:713–719.
7. Yazdany J, Pooley N, Langham J, et al. Systemic lupus erythematosus; stroke and myocardial infarction risk: a systematic review and meta-analysis. *RMD Open*. 2020;6.
8. Baragetti A, Ramirez GA, Magnoni M, et al. Disease trends over time and CD4(+)CCR5(+) T-cells expansion predict carotid atherosclerosis development in patients with systemic lupus erythematosus. *Nutr Metab Cardiovasc Dis*. 2018;28:53–63. doi:10.1016/j.numecd.2017.09.001
9. Alpizar-Rodriguez D, Romero-Diaz J. Are cardiovascular events and mortality in patients with systemic lupus erythematosus predictable at diagnosis? *Rheumatology*. 2020;59:467–468. doi:10.1093/rheumatology/kez539
10. Petruzzo M, Reia A, Maniscalco GT, et al. The Framingham cardiovascular risk score and 5-year progression of multiple sclerosis. *Eur J Neurol*. 2021;28:893–900. doi:10.1111/ene.14608
11. Summers DJ. Cardiovascular disease: what does QRISK actually measure? BMJ. *BMJ (Clinical Research Ed.)*. 2023;382:1502. doi:10.1136/bmj.p1502

12. Graham IM, Di AE, Huculeci R. New Way to "SCORE" risk: updates on the ESC scoring system and incorporation into ESC cardiovascular prevention guidelines. *Current Cardiology Reports*. 2022;24:1679–1684. doi:10.1007/s11886-022-01790-6
13. Barbhaiya M, Feldman CH, Guan H, et al. Race/ethnicity and cardiovascular events among patients with systemic lupus erythematosus. *Arthritis Rheumatol*. 2017;69:1823–1831. doi:10.1002/art.40174
14. Gao N, Kong M, Li X, et al. Systemic lupus erythematosus and cardiovascular disease: a Mendelian randomization study. *Front Immunol*. 2022;13:908831. doi:10.3389/fimmu.2022.908831
15. Troncoso J Á, Soto Abanades C, Á RM, et al. Cardiac involvement in a Spanish unicentric prospective cohort of patients with systemic lupus erythematosus. *Lupus*. 2023;32:111–118. doi:10.1177/09612033221141264
16. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349:255–260. doi:10.1126/science.aaa8415
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Med*. 2019;25:44–56. doi:10.1038/s41591-018-0300-7
18. Riboldi P, Gerosa M, Luzzana C, Catelli L. Cardiac involvement in systemic autoimmune diseases. *Clin Rev Allergy Immunol*. 2002;23:247–261. doi:10.1385/CRIAI:23:3:247
19. Khan SS, Matsushita K, Sang Y, et al. Development and Validation of the American Heart Association's PREVENT Equations. *Circulation*. 2024;149:430–449. doi:10.1161/CIRCULATIONAHA.123.067626
20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach Intell*. 2019;1:206–215. doi:10.1038/s42256-019-0048-x
21. Radner H, Lesperance T, Accortt NA, Solomon DH. Incidence and prevalence of cardiovascular risk factors among patients with rheumatoid arthritis, psoriasis, or psoriatic arthritis. *Arthritis Care Research*. 2017;69:1510–1518. doi:10.1002/acr.23171
22. Heslinga M, Nielen MMJ, Smulders Y, Simsek S, Nurmohamed MT. Amplified prevalence and incidence of cardiovascular disease in patients with inflammatory arthritis and coexistent autoimmune disorders. *Rheumatology*. 2020;59:2448–2454. doi:10.1093/rheumatology/kez650
23. Avina-Zubieta JA, Thomas J, Sadatsafavi M, Lehman AJ, Lacaille D. Risk of incident cardiovascular events in patients with rheumatoid arthritis: a meta-analysis of observational studies. *Ann Rheumatic Dis*. 2012;71:1524–1529. doi:10.1136/annrheumdis-2011-200726
24. Steyers CM 3rd, Miller FJ. Endothelial dysfunction in chronic inflammatory diseases. *Int J Mol Sci*. 2014;15:11324–11349. doi:10.3390/ijms150711324
25. Caraba A, Stancu O, Crişan V, Georgescu D. Anti TNF-alpha treatment improves microvascular endothelial dysfunction in rheumatoid arthritis patients. *Int J Mol Sci*. 2024;26:25. doi:10.3390/ijms26010025
26. Protogerou AD, Zampeli E, Fragiadaki K, Stamatelopoulos K, Papamichael C, Sfikakis PP. A pilot study of endothelial dysfunction and aortic stiffness after interleukin-6 receptor inhibition in rheumatoid arthritis. *Atherosclerosis*. 2011;219:734–736. doi:10.1016/j.atherosclerosis.2011.09.015
27. Dimitroulas T, Hodson J, Sandoo A, Smith J, Kitas GD. Endothelial injury in rheumatoid arthritis: a crosstalk between dimethylarginines and systemic inflammation. *Arthritis Res Ther*. 2017;19:32. doi:10.1186/s13075-017-1232-1
28. Ambler WG, Kaplan MJ. Vascular damage in systemic lupus erythematosus. *Nat Rev Nephrol*. 2024;20:251–265. doi:10.1038/s41581-023-00797-8
29. Shimokawa H, Suda A, Takahashi J, et al. Clinical characteristics and prognosis of patients with microvascular angina: an international and prospective cohort study by the Coronary Vasomotor Disorders International Study (COVADIS) Group. *Eur Heart J*. 2021;42:4592–4600. doi:10.1093/eurheartj/ehab282
30. González-Gay MA, González-Juanatey C. Inflammation and lipid profile in rheumatoid arthritis: bridging an apparent paradox. Annals of the rheumatic diseases. *Ann Rheumatic Dis*. 2014;73:1281–1283. doi:10.1136/annrheumdis-2013-204933
31. Myasoedova E, Crowson CS, Kremers HM, et al. Lipid paradox in rheumatoid arthritis: the impact of serum lipid measures and systemic inflammation on the risk of cardiovascular disease. *Ann Rheumatic Dis*. 2011;70:482–487. doi:10.1136/ard.2010.135871
32. Wu GC, Liu HR, Leng RX, et al. Subclinical atherosclerosis in patients with systemic lupus erythematosus: a systemic review and meta-analysis. *Autoimmunity Rev*. 2016;15:22–37. doi:10.1016/j.autrev.2015.10.002
33. Libby P, Hansson GK. from focal lipid storage to systemic inflammation: JACC review topic of the week. *J Am Coll Cardiol*. 2019;74:1594–1607. doi:10.1016/j.jacc.2019.07.061
34. Motoki Y, Nojima J, Yanagihara M, et al. Anti-phospholipid antibodies contribute to arteriosclerosis in patients with systemic lupus erythematosus through induction of tissue factor expression and cytokine production from peripheral blood mononuclear cells. *Thrombosis Research*. 2012;130:667–673. doi:10.1016/j.thromres.2011.11.048
35. Hollan I, Meroni PL, Ahearn JM, et al. Cardiovascular disease in autoimmune rheumatic diseases. *Autoimmunity Rev*. 2013;12:1004–1015. doi:10.1016/j.autrev.2013.03.013
36. Shi LH, Lam SHM, So H, Meng H, Tam LS. Impact of inflammation and anti-inflammatory therapies on the incidence of major cardiovascular events in patients with ankylosing spondylitis: a population-based study. *Semin Arthritis Rheum*. 2024;67:152477. doi:10.1016/j.semarthrit.2024.152477