Novel multiclass classification machine learning approach for the early-stage classification of systemic autoimmune rheumatic diseases

Yun Wang, Wei Wei, Renren Ouyang, Rujia Chen, Ting Wang, Xu Yuan, Feng Wang, Hongyan Hou, Shiji Wu [ID]

Department of Laboratory Medicine, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology, Wuhan, Hubei, China

Correspondence to
Dr Shiji Wu; wilson547@tjh.tjmu.edu.cn, Dr Hongyan Hou; tjhouhongyan@tjh.tjmu.edu.cn and Dr Feng Wang; fengwang@tjh.tjmu.edu.cn

ABSTRACT

Objective Systemic autoimmune rheumatic diseases (SARDs) encompass a diverse group of complex conditions with overlapping clinical features, making accurate diagnosis challenging. This study aims to develop a multiclass machine learning (ML) model for early-stage SARDs classification using accessible laboratory indicators.

Methods A total of 925 SARDs patients were included, categorised into SLE, Sjögren's syndrome (SS) and inflammatory myositis (IM). Clinical characteristics and laboratory markers were collected and nine key indicators, including anti-dsDNA, anti-SS-A60, anti-Sm/nRNP, antichromatin, anti-dsDNA (indirect immunofluorescence assay), haemoglobin (Hb), platelet, neutrophil percentage and cytoplasmic patterns (AC-19, AC-20), were selected for model building. Various ML algorithms were used to construct a tripartite classification ML model.

Results Patients were divided into two cohorts, cohort 1 was used to construct a tripartite classification model. Among models assessed, the random forest (RF) model demonstrated superior performance in distinguishing SLE, IM and SS (with area under curve=0.953, 0.903 and 0.836; accuracy= 0.892, 0.869 and 0.857; sensitivity= 0.890, 0.868 and 0.795; specificity= 0.910, 0.836 and 0.748; positive predictive value=0.922, 0.727 and 0.663; and negative predictive value= 0.854, 0.915 and 0.879). The RF model excelled in classifying SLE (precision=0.930, recall=0.985, F1 score=0.957). For IM and SS, RF model outcomes were (precision=0.793, 0.950; recall=0.920, 0.679; F1 score=0.852, 0.792). Cohort 2 served as an external validation set, achieving an overall accuracy of 87.3%. Individual classification performances for SLE, SS and IM were excellent, with precision, recall and F1 scores specified. SHAP analysis highlighted significant contributions from antibody profiles.

Conclusion This pioneering multiclass ML model, using basic laboratory indicators, enhances clinical feasibility and demonstrates promising potential for SARDs classification. The collaboration of clinical expertise and ML offers a nuanced approach to SARDs classification, with potential for enhanced patient care.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Machine learning (ML) models have been successfully applied to various aspects of rheumatology, but the utilisation of ML for the differential classification of systemic autoimmune rheumatic diseases (SARDs), particularly during the early stages, remains a relatively underexplored area of research.

WHAT THIS STUDY ADDS

⇒ This study introduced the application of a multiclass ML model in the realm of SARDs demonstrates the feasibility and effectiveness of employing an ML model based on basic laboratory indicators for the accurate multiclass classification of these diseases.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The multiclass ML model promises to provide a more nuanced and complex classification of SARDs, ultimately paving the way for enhanced patient care and improved disease management.

INTRODUCTION

Systemic autoimmune rheumatic diseases (SARDs) encompass a diverse group of chronic and complex conditions characterised by aberrant immune responses leading to self-directed tissue and organ damage.[1] These diseases, including SLE, rheumatoid arthritis (RA), systemic sclerosis, Sjögren's syndrome (SS), mixed connective tissue disease and inflammatory myositis (IM), among others, share a common autoimmune pathogenesis while exhibiting similar or distinct clinical presentations and affecting various organs.[2] The classification of SARDs is of paramount importance in clinical practice, research and drug development, as it enables the inclusion of homogeneous patient populations and facilitates the evaluation of disease outcomes and therapeutic interventions.[3] However, prompt and accurate diagnosis of SARDs remains a significant challenge in clinical

practice due to the overlapping clinical features shared among various autoimmune disorders.[4]

The absence of definitive diagnostic criteria necessitates the reliance on classification criteria, developed primarily for research purposes, as diagnostic aids.[5] Among the widely recognised classification criteria, two of the most prominent are the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) criteria. These criteria have been refined over the years through iterative processes involving expert consensus and systematic data analysis, making them primary tools for classifying diseases such as RA,[6] SLE,[7 8] SS[9] and IM.[10] Recently, efforts have been made to develop more sensitive and specific classification criteria to capture early and milder forms of SARDs. This led to the development of the Systemic Lupus International Collaborating Clinics (SLICC) criteria for SLE[11] and the EULAR/ACR 2019 criteria,[12] which combines elements from both organisations for SLE classification. Despite these advancements, challenges remain in the classification of SARDs, particularly in distinguishing between diseases with overlapping clinical features and in identifying rare and atypical cases.[13] Consequently, delayed diagnosis and treatment initiation can result in increased disease activity, organ damage and poor patient outcomes.

Artificial intelligence (AI) and machine learning (ML) are increasingly recognised as powerful tools capable of handling complex medical tasks.[14] ML offers distinct advantages over conventionally programmed strategies, particularly in handling complex multidimensional data by identifying latent relationships that may have escaped prior recognition.[15] ML models, trained on diverse medical and biological data, have been successfully applied to various aspects of rheumatology,[16] including molecular classification of SS,[17] IM[18] and RA,[19] assessment of disease activity,[20] response to treatment[21] and prediction of disease outcomes[22] or mortality.[23] The capacity of ML to comprehensively analyse extensive and intricate clinical datasets holds immense potential. Nevertheless, its seamless translation into clinical practice within the realm of rheumatology is still at a formative stage.[24] Moreover, the utilisation of ML for the differential diagnosis of SARDs, particularly during the early stages,[25–28] remains a relatively underexplored area of research. Specifically, there is a dearth of investigations focusing on the development of ML models that rely on straightforward and readily available objective laboratory indicators, such as ANA profiles, to accurately identify common SARDs in their early stages (within 2 year of disease onset).

MATERIALS AND METHODS
Study design and study population cohorts
We conducted a single-centre observational retrospective study that included all study subjects from the Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology.

A total of 925 patient initially diagnosed with SLE (n=519), SS (n=163) and IM (n=243) were included in this study from October 2017 to June 2023. Cohort 1 (model building and internal validation): patients with SLE, IM and SS who had their initial consultations from October 2017 to May 2022 were included to establish a classification model for distinguishing between IM, SLE and SS. Cohort 2 (external validation): patients with IM, SLE and SS who were first seen between June 2022 and June 2023 were included for external validation. All patients were diagnosed for the first time within 2 years of symptom onset. The major clinical symptoms included dizziness and fatigue, cough, joint pain (synovitis involving two or more joints, characterised by swelling or effusion or tenderness in two or more joints and 30 min or more of morning stiffness), oedema (extremity oedema, eyelid oedema, abdominal oedema and pulmonary oedema etc), proteinuria (urine protein>0.5 g/24 hours), fever (>38.2°C) and rash (such as butterfly rash, subacute cutaneous lupus erythematosus, discoid lupus erythematosus, purpura etc). None of these patients had received specific treatment or previous immunosuppressive drugs. The diagnosis of SS was made according to the criteria of the 2016 ACR/EULAR Classification Criteria,[9] while the diagnosis of IM was based on the criteria of the 2017 EULAR/ACR classification criteria for adult and juvenile idiopathic inflammatory myopathies.[10] The diagnosis of SLE before 2019 followed the 2012 SLICC Classification Criteria for SLE,[11] and from 2019 to 2023, the criteria of the 2019 EULAR/ACR SLE classification criteria[12] were employed. Exclusion criteria encompassed patients who were HIV or hepatitis C virus positive and malignancies. The patients who may satisfy multiple classification criteria were also excluded.

Data cleaning
The basic information of the study subjects (age, gender, underlying disease and medical history) and the results of all first laboratory tests (antinuclear antibodies, anti-dsDNA (indirect immunofluorescence assay (IIFA)), routine blood, liver function, kidney function, ion, lipid, glucose, cardiac protein and cardiac enzyme, etc) were collected in the medical record system within 72 hours after admission. The patient ID number is used as the unique identifier. Serum ANA profiles levels were measured using BioPlex 2200 System. The IIFA for the detection of ANA and anti-dsDNA was performed using the EUROPattern Computer-aided immunofluorescence microscopy (EPA). Routine blood levels were detected by the Sysmex instrument. Liver function and kidney function levels were measured using the Roche Cobas e701 Automatic Electrochemiluminescence Immunoassay System. Test data with missing rates >30% were excluded.

Model fitting and evaluation
The flow chart for building a model using ML algorithms is shown in figure 1. ML models were constructed using the variables that exhibited statistically significant
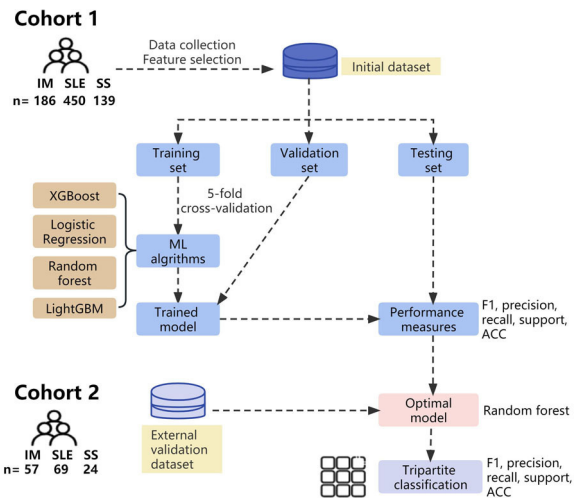
Figure 1 Flowchart of machine learning to build the diagnosis model. ACC, accuracy; IM, inflammatory myositis; SS, Sjögren's syndrome.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$$

Statistical analysis

Continuous variables were presented as either mean±SD or median (IQ), depending on the data distribution. The Student's t-test, Mann-Whitney U test or one-way analysis of variance test was used to compare the variables between groups, as appropriate. Categorical variables were compared using the χ^2 test or Fisher's exact test based on the sample size and expected cell frequencies. Receiver operating characteristic (ROC) curve analysis was performed to identify the optimal cut-off values for parameters, aiming to achieve the highest sensitivity and specificity. Lasso regression was employed for feature variable selection using the R package 'Lasso2'. To explore differences between groups based on laboratory indicators, t-distributed stochastic neighbour embedding analysis was conducted using the R package 'Rtsne', while heatmaps were generated using the R package 'pheatmap'. Correlations between indicators were analysed and correlation matrix visualisation was performed using the R package 'corrplot'. Principal component analysis (PCA) was executed using the R package 'PCA' to analyse the associations between categorical variables. These tests were employed to analyse the associations between categorical variables in the study. Statistical analyses were performed using GraphPad Prism V.9.0 (San Diego, California, USA) and SPSS V.22.0. Statistical significance was determined as p<0.05.

RESULTS
Basic characteristics of included cohorts
Comparing the IM, SLE and SS groups within cohort 1, no significant differences were identified with regard to age, sex, joint pain and fever. The predominant clinical features in IM, SLE and SS groups were dizziness and fatigue (35.48%), oedema (35.78%) and fever (23.74%), respectively. The clinical symptoms observed in cohort 2 closely paralleled those identified in the three groups within cohort 1 (table 2). Certain symptoms such as proteinuria and oedema were notably higher in SLE compared with SS and IM. Joint pain, a hallmark symptom of many rheumatic diseases, displayed varying prevalence rates, further underscoring the intricate nature of symptomatology in SARDs. Additionally, symptoms that might not immediately seem directly related to rheumatic diseases, such as cough and fever, exhibited noteworthy prevalence rates, highlighting the multisystemic nature of these conditions.

Significantly differences of ANA patterns and routine laboratory indicators were observed among SLE, SS and IM groups, including blood routine, liver function, complements and immunoglobulin (online supplemental table 2). Further insights into the ANA profiles

differences among three groups, as well as variables that have been highlighted in existing literature. To mitigate multicollinearity, correlation analysis was performed to identify indicators with high correlations. Subsequently, ROC analysis, single-factor analysis or multifactor regression were employed to identify the distinctive features (online supplemental table 1).

Cohort 1 was randomly divided into two subsets: 85% for training and 15% for testing. The training set uses fivefold cross-validation of the entire dataset into five-folds, using four of the folds as the training set to train the model, and the remaining one fold as the internal validation set to score the model, and repeating the above process five times. We compare the early classification models built by Extreme Gradient Boosting (XgBoost), Logistic Regression, Light Gradient Boosting Machine (LightGBM) and RF, and select the best performing ML algorithms and early classification models (table 1) in the validation set. Cohort 2 performs external validation of the model. The prediction models were constructed using the Beckman Coulter DxAI platform (https://www.xsmartanalysis.com/beckman/login/). Outcome was categorised into three categories (SLE, IM and SS). RF model was selected to construct classification models. The RF model is an ensemble learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting by using random feature selection and voting/averaging mechanisms. It is widely used for classification and regression tasks. Shapley additive explanation (SHAP) analysis was used to elucidate the significance and contributions of features within the RF model. To evaluate the performance of the models, accuracy was calculated to determine the model's performance. The macro and weighted precision, recall, F1 were calculated for the three classes.

$$Accuracy = \frac{TP+TN}{P+N}$$

Table 1 Results of the different machine learning models

Groups	Model type	AUC (95% CI)	Accuracy (95% CI)	Sensibility (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score (95% CI)
SLE	XGBoost	0.951 (0.931 to 0.972)	0.878 (0.876 to 0.881)	0.885 (0.858 to 0.912)	0.894 (0.888 to 0.901)	0.938 (0.890 to 0.986)	0.814 (0.780 to 0.848)	0.911 (0.874 to 0.948)
	Logistic	0.943 (0.920 to 0.967)	0.893 (0.888 to 0.898)	0.874 (0.833 to 0.914)	0.933 (0.876 to 0.989)	0.957 (0.933 to 0.980)	0.824 (0.815 to 0.833)	0.913 (0.902 to 0.925)
	RandomForest	0.953 (0.932 to 0.974)	0.892 (0.889 to 0.894)	0.890 (0.854 to 0.926)	0.910 (0.885 to 0.935)	0.922 (0.887 to 0.958)	0.854 (0.821 to 0.887)	0.905 (0.904 to 0.907)
	LightGBM	0.950 (0.929 to 0.971)	0.884 (0.881 to 0.886)	0.860 (0.838 to 0.883)	0.936 (0.898 to 0.974)	0.941 (0.897 to 0.984)	0.821 (0.790 to 0.853)	0.898 (0.891 to 0.906)
IM	XGBoost	0.911 (0.879 to 0.942)	0.853 (0.848 to 0.858)	0.878 (0.855 to 0.901)	0.810 (0.755 to 0.864)	0.800 (0.722 to 0.878)	0.865 (0.839 to 0.890)	0.837 (0.804 to 0.869)
	Logistic	0.911 (0.879 to 0.943)	0.853 (0.832 to 0.874)	0.884 (0.810 to 0.959)	0.855 (0.849 to 0.862)	0.652 (0.579 to 0.724)	0.952 (0.915 to 0.989)	0.749 (0.727 to 0.770)
	RandomForest	0.903 (0.866 to 0.940)	0.869 (0.869 to 0.869)	0.868 (0.782 to 0.953)	0.836 (0.736 to 0.936)	0.727 (0.705 to 0.750)	0.915 (0.903 to 0.927)	0.790 (0.768 to 0.813)
	LightGBM	0.911 (0.880 to 0.941)	0.856 (0.840 to 0.871)	0.873 (0.861 to 0.885)	0.827 (0.786 to 0.868)	0.758 (0.721 to 0.796)	0.878 (0.866 to 0.890)	0.812 (0.785 to 0.839)
SS	XGBoost	0.839 (0.790 to 0.888)	0.854 (0.841 to 0.867)	0.826 (0.662 to 0.989)	0.698 (0.494 to 0.902)	0.638 (0.582 to 0.694)	0.881 (0.873 to 0.890)	0.716 (0.689 to 0.743)
	Logistic	0.883 (0.840 to 0.926)	0.735 (0.704 to 0.767)	0.765 (0.572 to 0.958)	0.860 (0.752 to 0.969)	0.387 (0.359 to 0.415)	0.956 (0.951 to 0.962)	0.511 (0.492 to 0.530)
	RandomForest	0.836 (0.782 to 0.890)	0.857 (0.849 to 0.865)	0.795 (0.632 to 0.959)	0.748 (0.567 to 0.930)	0.663 (0.620 to 0.705)	0.879 (0.876 to 0.883)	0.720 (0.677 to 0.762)
	LightGBM	0.852 (0.804 to 0.899)	0.844 (0.815 to 0.872)	0.818 (0.610 to 1.026)	0.747 (0.479 to 1.014)	0.585 (0.454 to 0.715)	0.884 (0.883 to 0.886)	0.672 (0.656 to 0.688)

AUC, area under curve; IM, inflammatory myositis; LightGBM, Light Gradient Boosting Machine; Logistic, Logistic Regression; NPV, negative predictive value; PPV, positive predictive value; SS, Sjögren's syndrome; XGBoost, Extreme Gradient Boosting.

Table 2 Clinical baseline characteristics

	Cohort 1					Cohort 2				
	Total (n=775)	SLE (n=450)	IM (n=186)	SS (n=139)	P value	Total (n=150)	SLE (n=69)	IM (n=57)	SS (n=24)	P value
Age, median (IQR), year	51.00 (35.00–61.00)	49.00 (34.00–61.00)	52.00 (35.00–62.00)	53.00 (37.00–64.00)	0.220	46.00 (32.00–60.00)	34.00 (26.00–54.00)	53.00 (39.00–61.00)	58.00 (45.00–71.00)	<0.001
Sex										
Male, n (%)	190 (24.52)	103 (22.89)	50 (26.88)	37 (26.62)	0.463	31 (20.67)	13 (18.84)	9 (15.79)	9 (37.50)	0.077
Female, n (%)	585 (75.48)	347 (77.11)	136 (73.12)	102 (73.38)		119 (79.33)	56 (81.16)	48 (84.21)	15 (62.50)	
Clinical, n (%)										
Dizziness and fatigue, n (%)	159 (20.52)	62 (13.78)	66 (35.48)	31 (22.30)	<0.001	41 (27.33)	10 (14.49)	27 (47.37)	4 (16.67)	<0.001
Cough, n (%)	151 (19.48)	71 (15.78)	49 (26.34)	31 (22.30)	0.006	18 (12.00)	7 (10.14)	6 (10.53)	5 (20.83)	0.347
Joint pain, n (%)	138 (17.81)	89 (19.78)	27 (14.52)	22 (15.83)	0.230	24 (16.00)	15 (21.74)	4 (7.02)	5 (20.83)	0.063
Oedema, n (%)	190 (24.52)	161 (35.78)	19 (10.22)	10 (7.19)	<0.001	28 (18.67)	22 (31.88)	3 (5.26)	3 (12.50)	<0.001
Proteinuria, n (%)	135 (17.42)	118 (26.22)	6 (3.23)	11 (7.91)	<0.001	12 (8.00)	10 (14.49)	1 (1.75)	1 (4.17)	0.024
Fever, n (%)	162 (20.90)	89 (19.78)	40 (21.51)	33 (23.74)	0.588	22 (14.67)	16 (23.19)	1 (1.75)	5 (20.83)	0.002
Rash, n (%)	120 (15.48)	72 (16.00)	42 (22.58)	6 (4.32)	<0.001	22 (14.67)	15 (21.74)	6 (10.53)	1 (4.17)	0.059

Data are presented as number (%), or median (25th–75th percentile).
IM, inflammatory myositis; SS, Sjögren's syndrome.

are depicted (figure 2A–L). Markers specific to SLE, such as anti-dsDNA, antichromatin, anti-Sm and antiribosomal P antibodies, displayed significantly higher concentrations in the SLE group compared with the SS and IM groups. SLE-associated markers such as anti-Sm/nRNP and anti-RNP A/68 also exhibited higher concentrations in the SLE group. Notably, the IM-specific antibody anti-Jo-1 showed the highest levels in the IM group. The antibody anti-SS-B, more closely related to SS, displayed the highest concentration in the SS group. Moreover, the antibody anti-SS-A60, highly associated with both SLE and SS, demonstrated no distinction between the two groups but was significantly elevated in comparison to the IM group. However, the major clinical symptoms and routine laboratory indicators did not exhibit a clustered distribution across the three disease groups (online supplemental figure 1A). With the addition of specific laboratory indicators (ANA patterns and ANA profiles), the distinction improved to some extent, although there still remained significant overlap (online supplemental figure 1B).

Establishment of machine learning model

Following a rigorous feature selection strategy (as described in the Materials and methods section), we identified the most informative features that could serve as potential biomarkers for distinguishing between IM, SLE and SS. Ultimately, nine key features, anti-dsDNA, anti-SS-A60, anti-Sm/nRNP, antichromatin, anti-dsDNA (IIFA), haemoglobin (Hb) levels, platelet count (PLT), neutrophil percentage (NEUT%) and cytoplasmic patterns (AC-19, AC-20), were finally selected to build the ML model. Based on the afore-mentioned indicators, we opted for the top-performing model using the RF classifier model for its superior performance when distinguishing between SLE, IM and SS. For the predictive classification of SLE, the RF model demonstrated an area under curve (AUC) value of 0.953 (95% CI: 0.932 to 0.974), an accuracy of 0.892 (95% CI: 0.889 to 0.894), sensitivity of 0.890 (95% CI: 0.854 to 0.926), specificity of 0.910 (95% CI: 0.885 to 0.935), a positive predictive value (PPV) of 0.922 (95% CI: 0.887 to 0.958) and a negative predictive value (NPV) of 0.854 (95% CI: 0.821 to 0.887). For the predictive classification of IM and SS, the RF model yielded AUC values of 0.903 (95% CI: 0.866 to 0.940) and 0.836 (95% CI: 0.782 to 0.890), respectively, along with accuracy values of 0.869 and 0.857. The model also exhibited sensitivity values of 0.868 and 0.795, specificity values of 0.836 and 0.748, PPVs of 0.727 and 0.663 and NPVs of 0.915 and 0.879 (table 1). We trained and evaluated RF models to classify the three classes of our cohorts (IM, SLE and SS). The results encompassing weighted precision, recall, weighted F1 scores and support for our predictive models are comprehensively presented in table 3. In the training set, weighted precision, recall and weighted F1 scores for all three groups were >0.9. On evaluating the testing set, the RF model demonstrated its optimal diagnostic prowess in identifying SLE cases (precision=0.930, recall=0.985, F1 score=0.957). When predicting IM and SS cases, the

RF model yielded the outcomes (precision=0.793, 0.950, recall=0.920, 0.679, F1 score=0.852, 0.792), respectively. The comprehensive confusion matrix for both the training and validation sets is illustrated in figure 3A. Further elucidating the model's intricacies, the SHAP values are presented in figure 3B–C. Notably, antichromatin, anti-dsDNA and anti-SS-A60 emerged as pivotal contributors significantly influencing the model's output magnitude, while cytoplasmic patterns (AC-19, AC-20) exhibited the lowest SHAP value, implying relatively lesser impact on the model's predictions. In the PCA biplot graph, the contribution of these nine indicators to the classification of SLE, SS and IM can also be observed (online supplemental figure 2). Notably, anti-dsDNA, anti-Sm/nRNP, antichromatin and anti-dsDNA (IIFA) antibodies are oriented towards SLE and exhibit relatively significant contributions. Conversely, Hb and PLT predominantly align with the IM category, showcasing substantial contributions. NEUT% lies between SLE and IM, suggesting a middle ground. Similar to the SHAP graph, cytoplasmic patterns (AC-19, AC-20) demonstrate the least contribution. The presence of anti-SS-A antibodies between SLE and SS categories is not unexpected, given their intermediate positioning. The comprehensive evaluation of bias and model risks, conducted through the PROBAST tool, can be found in online supplemental table 3.

External verification

Using cohort 2 as an external verification set, our model's performance was evaluated, yielding an overall diagnostic accuracy of 87.3% (figure 4). Specifically, when classifying individual diseases, the model exhibited its highest diagnostic performance in correctly identifying cases of SLE, with a precision of 0.882, recall of 0.971 and an F1 score of 0.924. In terms of differentiating between IM and SS cases, the RF model demonstrated precision rates of 0.897 and 0.750, along with recall rates of 0.912 and 0.500, respectively. These outcomes translated to F1 scores of 0.904 for IM and 0.600 for SS (table 4).

DISCUSSION

The clinical presentation of SARDs is characterised by a diverse array of symptoms.[2] The intricate interplay between diverse symptoms in each SARD highlights the challenges faced by clinicians in distinguishing these conditions from one another, particularly considering the potential for overlapping symptoms.[29 30] Within clinical practice, skilled physicians can often arrive at a diagnosis of SARDs even with the presence of a limited number of highly informative manifestations in an individual. For instance, the presence of the classic malar rash coupled with anti-dsDNA can lead to a diagnosis of SLE. However, this diagnostic acumen is a manifestation of clinical expertise cultivated through experience.[31] The data emphasise the importance of developing advanced diagnostic tools, such as ML models, to aid in the accurate and timely differentiation of SARDs.
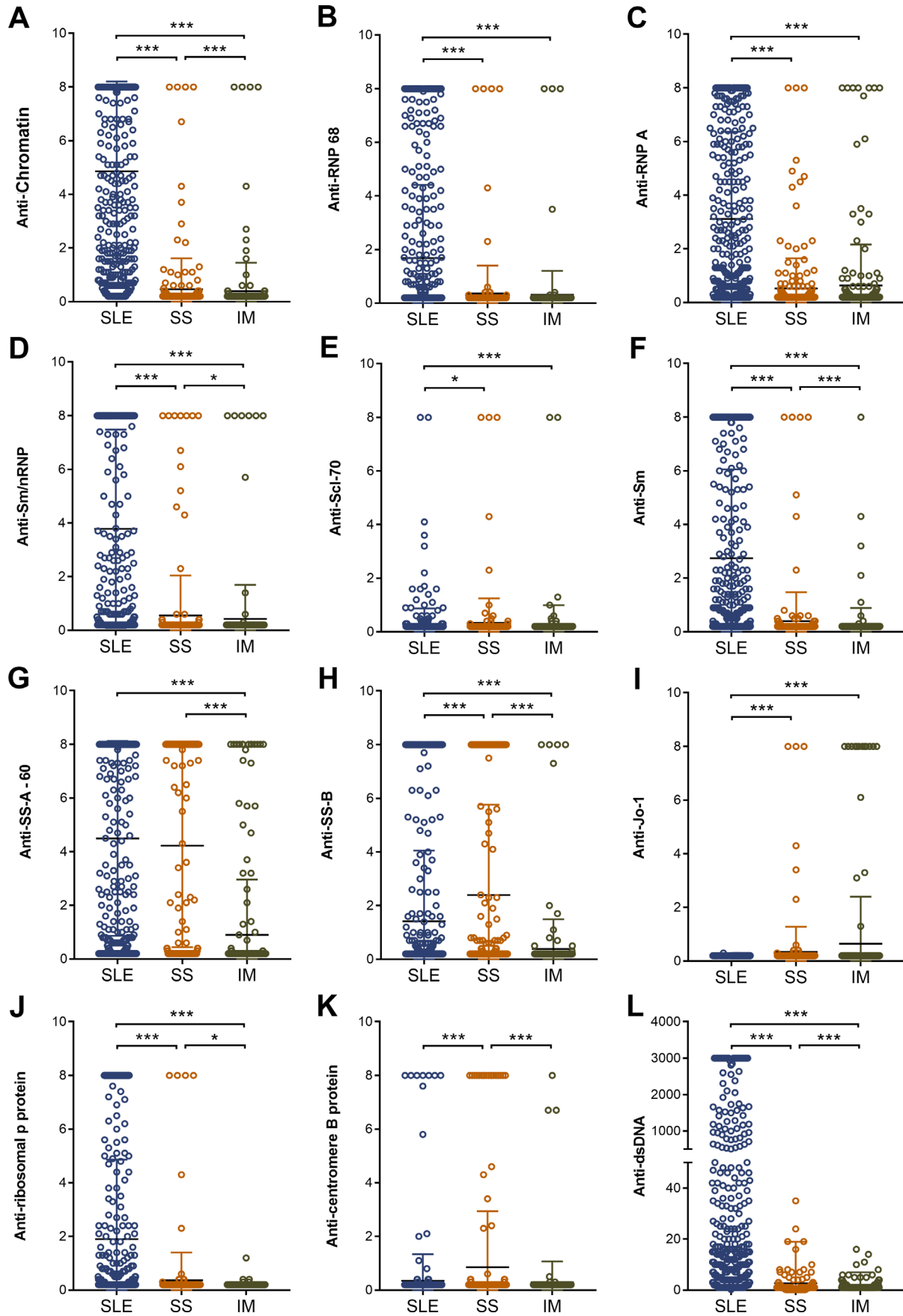
Figure 2 The characteristics of antinuclear antibody profiles in patients newly diagnosed with SLE, SS and IM (A–L). The data were presented as mean with SD. Blue circle points represent SLE, orange circle points represent SS and green circle points represent IM. *P<0.05, ***p<0.001. IM, inflammatory myositis; SS, Sjögren's syndrome.

Table 3 Performance of the built machine learning model

Groups	Training set				Testing set			
	Precision	Recall	F1 score	Support	Precision	Recall	F1 score	Support
SLE	0.975	0.975	0.975	406	0.930	0.985	0.957	67
IM	0.922	0.951	0.936	162	0.793	0.920	0.852	25
SS	0.953	0.911	0.932	112	0.950	0.679	0.792	28
Macro avg	0.950	0.946	0.948	680	0.891	0.861	0.867	120
Weighted avg	0.959	0.959	0.959	680	0.906	0.900	0.896	120

IM, inflammatory myositis; SS, Sjögren's syndrome.

ML, a computational analytical approach, is gaining rapid prominence in the realm of biomedicine.[32 33] In the context of rheumatology, the integration of ML is witnessing a gradual upsurge, with numerous studies leveraging ML techniques to classify patients with SARDs based on diverse data sources encompassing medical records,[34 35] imaging data,[36] biometric measurements[37] and gene expression profiles.[18 20 38] The modelling indicators of these ML models primarily encompass clinical symptoms, which are relatively subjective, as well as a diverse array of complex laboratory indicators, and even biologically intricate indicators that are challenging to obtain, such as genetic polymorphisms. However, comparatively simple and easily accessible laboratory indicators, such as blood routine, and crucial autoantibodies of significant relevance in SARDs diagnosis, such as ANA profiles, are seldom incorporated into consideration. This limitation to some extent impacts the applicability of these models in clinical contexts.

In this study, we present a novel ML framework that is based on readily accessible and objective laboratory indicators, notably the ANA profile and blood routine,
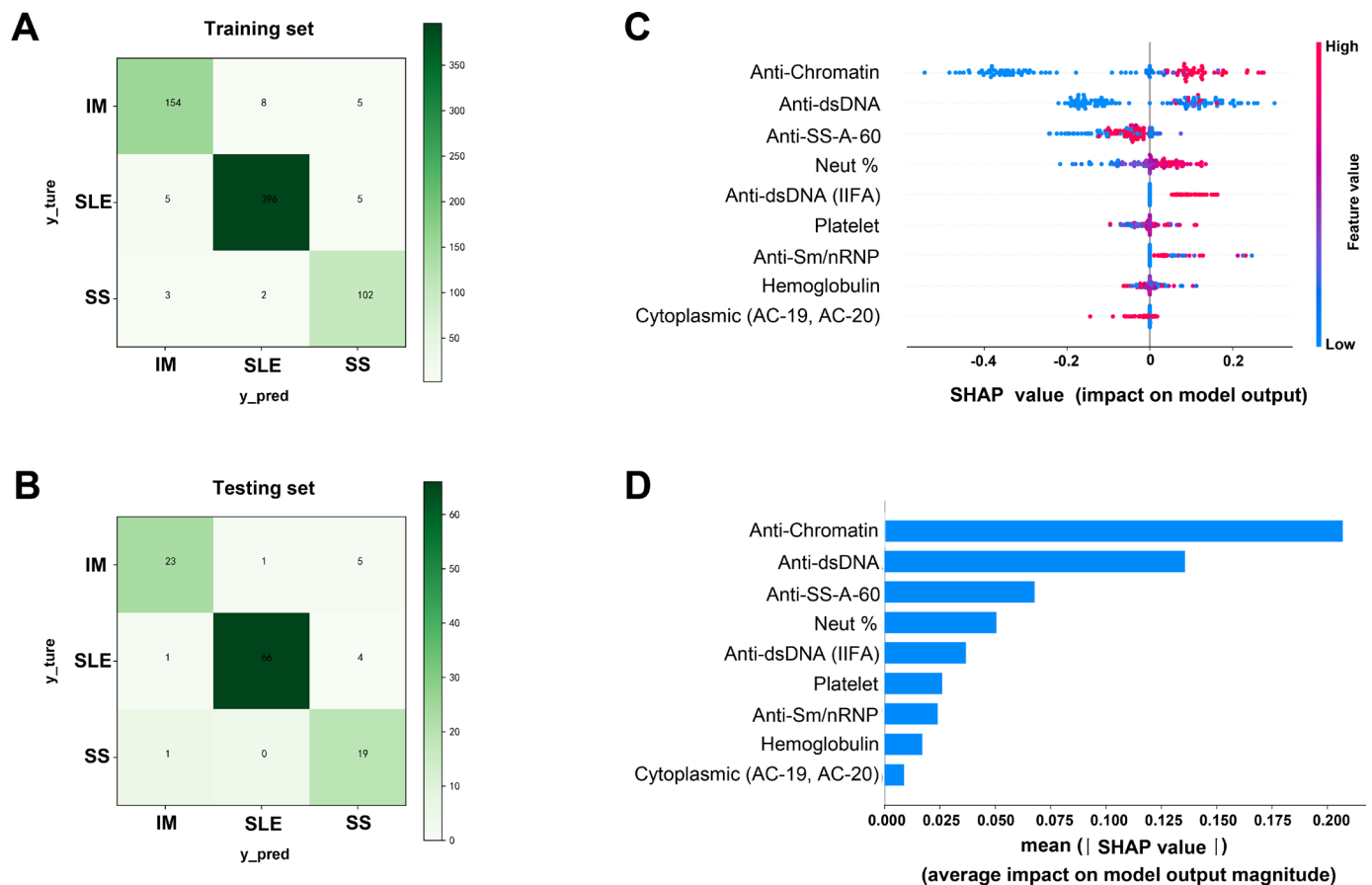
Figure 3 In the internal testing set, the random forest (RF) model built by nine key features (antichromatin, anti-dsDNA, anti-SS-A60, neutrophil percentage, anti-dsDNA IFT, platelet count, anti-Sm/nRNP, haemoglobin level and cytoplasmic (AC-19, AC-20)). (A) The comprehensive confusion matrix for the training set. (B) The comprehensive confusion matrix for the testing set. (C) The SHAP values. Each feature's SHAP value indicates its contribution to the model's prediction. (D) The weights of variables importance of top nine features. IM, inflammatory myositis; SHAP, Shapley additive explanation; SS, Sjögren's syndrome.
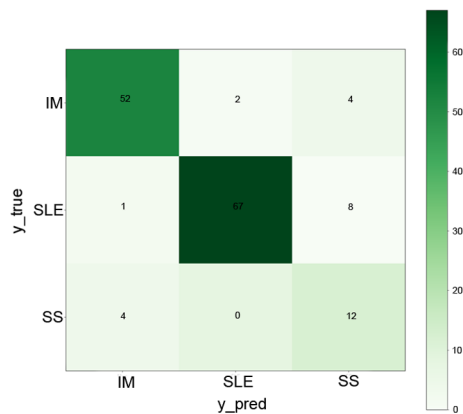
Figure 4 Confusion matrix for external validation of the diagnosis model. IM, inflammatory myositis; SS, Sjögren's syndrome.

tailored for the early-stage classification and diagnosis of three SARDs (SLE, SS and IM). This approach holds the potential to surmount the diagnostic challenges posed by the overlapping clinical presentations commonly observed in SARDs. Through an exhaustive process encompassing techniques such as differential analysis, feature selection, weight analysis, we ultimately identified laboratory indicators anti-dsDNA, anti-SS-A60, anti-Sm/nRNP, antichromatin, anti-dsDNA (IIFA), Hb, PLT, NEUT% and cytoplasmic patterns (AC-19, AC-20) for inclusion as modelling factors. The inclusion of ANA profiles as modelling indicators is expected. It is widely acknowledged that certain antibodies play pivotal roles in classifying SARDs.[39] For instance, the presence of anti-dsDNA and anti-Sm antibodies contributes to the classification criteria of SLE,[12] while the detection of anti-SS-A60 antibodies is significant for diagnosing SS.[9] Moreover, the prevalence of antichromatin antibodies is strongly associated with lupus nephritis.[40] Additionally, different methods for detecting anti-dsDNA antibodies may yield varying diagnostic performance. Currently, the most specific method acknowledged for diagnosing SLE is the IIFA.[41] Some indicators within the classification criteria, such as complement C3 and C4, were unexpectedly excluded, potentially due to their lack of specificity. Even more unexpectedly, certain highly specific antibodies such as anti-Jo-1, anti-Sm and anti-RibP antibodies were not included, possibly due to their low positive rates in

Table 4 Confusion matrix for external validation of the diagnosis model

Groups	Precision	Recall	F1 score	Support
SLE	0.882	0.971	0.924	69
IM	0.897	0.912	0.904	57
SS	0.750	0.500	0.600	24
Macro avg	0.843	0.794	0.809	150
Weighted avg	0.866	0.873	0.865	150

IM, inflammatory myositis; SS, Sjögren's syndrome.

clinic.[41] The IIFA on HEp-2 cells is widely used for detection of ANA. Fluorescence patterns may also reveal clinically relevant information. Cytoplasmic patterns (AC-19, AC-20) are associated with the distinct anti-tRNA synthetase antibodies, which are significant relevant antibodies in the context of IM.[42] Considering the comprehensive impact of SARDs on the blood system, factors such as Hb, PLT and NEUT% hold significant clinical significance. Reduced haemoglobin levels, known as anaemia, and thrombocytopenia, characterised by low PLT counts, are commonly observed in individuals with SARDs, arising from factors such as chronic inflammation, bone marrow suppression, renal involvement, immune-mediated destruction or impaired PLT production. Leucopenia is also a common symptom of SARDs and is one of the classification criteria for SLE. However, surprisingly, what was selected is the NEUT% rather than the total white blood cell counts. Neutrophils play a central role in the immune response, and alterations in NEUT%s can reflect the immune dysregulation present in SARDs. In recent years, emerging evidence highlights the potential of SLE-derived low-density granulocytes to contribute to vascular damage, heightened type I interferon synthesis, increased cell death and enhanced extracellular trap formation, all potentially significant in SLE pathogenesis and autoimmunity induction.[43]

In contrast to the prevalent binary classification methods employed by existing ML models, which primarily categorise outcomes as either 'yes' or 'no', our model represents a pioneering effort in the realm of multiclass classification, offering a novel approach where a single model can provide the risk probabilities associated with three distinct disease types for each patient which enhance clinical operability and applicability. We opted for the top-performing model using the RF Classifier model for its superior performance. The model demonstrated an accuracy of 90.0% when tested on a subset of 120 patients, randomly selected for evaluation. Notably, the accuracy rates for distinct SARDs subtypes were as follows: 98.5% (66/67) for SLE, 67.9% (19/28) for SS and 92.0% (23/25) for IM, respectively. Finally, we conducted a performance validation of the model in a separate cohort comprising 150 newly diagnosed patients, and the results consistently demonstrated a high accuracy rate (87.3%). Compared with other diagnostic models, our model demonstrates a comparable high performance. For example, Burlina *et al* achieved an 86.6% accuracy in diagnosing IM through AI learning of muscle ultrasound images.[36] Pinal-Fernandez *et al* further discovered that dermatomyositis, antisynthetase syndrome, immune-mediated necrotising myopathy and inclusion body myositis can be distinguished based on their unique gene expression patterns by applying ML algorithms to muscle biopsy transcriptomic data.[18] However, this study's outcomes generated five different ML models, a limitation that significantly constrains its practicality in clinical settings. Dros *et al* developed an ML model based on routine healthcare data, achieving an 84.0% diagnostic accuracy for primary Sjogren's

syndrome (pSS).[34] Another study, using a set of 14 signature genes that play pivotal roles in transcription regulation and disease progression in pSS, achieved an 83.2% accuracy for pSS diagnosis.[38] Our model demonstrates a lower classification accuracy for SS compared with the outcomes of these two studies. This disparity could be attributed to variations in the composition of patient groups, differences in the chosen modelling indicators, and other relevant factors. Notably, our study includes a smaller proportion of SS patients in comparison to those with SLE and IM. As a result, the selected modelling indicators, aside from anti-SS-A60, contribute less significantly to the accurate classification of SS in our model. In a study by Adamichou *et al*, an ML model relying on 14 indicators encompassing a diverse range of clinical symptoms and laboratory parameters, including complement C3 and C4 levels, achieved a diagnostic accuracy of 94.2% for SLE.[16] The research is one of the few studies that employs readily available and simple indicators for modelling, aiming to maximise the clinical applicability. However, there exists a substantial disparity in the modelling indicators used between our study and theirs. This divergence could potentially be attributed to the differences in the selected patient cohorts, as distinct patient populations often exhibit significant variations in their characteristic features, subsequently influencing the extracted feature indicators.

Our study brings forth a pioneering approach in the field of SARDs classification by introducing a novel multiclass ML model. This innovative model employs basic laboratory indicators, such as ANA profiles and blood routine results, to effectively classify patients into three distinct disease types. This approach not only offers a simplified classification process but also enhances the clinical feasibility of the model, aligning with real-world medical practices. However, it is important to acknowledge the limitation posed by our patient cohort's predominantly Han ethnicity composition, which restricts the generalisability of our findings to diverse ethnic groups. Second, in our study, autoantibodies are treated as quantitative data. It is important to note that values of antibody levels can vary significantly between different commercial assays, which may impact the generalisability of the model. The third concern is that many of our modelling indicators, such as anti-dsDNA antibodies, are autoantibodies already used in clinical classification, posing a significant risk of circular reasoning. This is also a key reason for the high-risk assessment of outcome and predictor association in the bias risk evaluation using PROBAST. However, it is crucial to clarify that the primary aim of our model is to extract information from rich laboratory parameters through ML, rather than simply reproducing clinical classification criteria. We will emphasise more clearly that, while some antibodies may be relevant to current classification criteria, the value of the model lies in the potential associations it learns from complex data. Additionally, the relatively lower accuracy in classifying SS patients highlights the need for larger sample sizes to improve SS

classification performance and ensure comprehensive applicability. Despite these limitations, our study's promising implications for accurate and expedited SARDs diagnosis underscore the necessity for further research and validation across diverse patient populations.

In conclusion, our study introduces the application of a multiclass ML model in the realm of SARDs, demonstrates the feasibility and effectiveness of employing an ML model based on basic laboratory indicators for the accurate multiclass classification of these diseases. Certainly, the wisdom and experience of clinical physicians remain essential. The amalgamation of clinical acumen with ML-driven insights promises a more nuanced and sophisticated approach to diagnosing SARDs, ultimately paving the way for enhanced patient care and improved disease management.

ORCID iD
Shiji Wu http://orcid.org/0000-0003-2337-0509

REFERENCES

1. Wahren-Herlenius M, Dörner T. Immunopathogenic mechanisms of systemic autoimmune disease. *Lancet* 2013;382:819–31.
2. Goldblatt F, O'Neill SG. Clinical aspects of autoimmune rheumatic diseases. *Lancet* 2013;382:797–808.
3. Lisnevskaia L, Murphy G, Isenberg D. Systemic lupus erythematosus. *Lancet* 2014;384:1878–88.
4. Giacomelli R, Afeltra A, Alunno A, *et al*. International consensus: what else can we do to improve diagnosis and therapeutic strategies in patients affected by autoimmune rheumatic diseases (rheumatoid arthritis, spondyloarthritides, systemic sclerosis, systemic lupus erythematosus, antiphospholipid syndrome and

Sjogren's syndrome)? The unmet needs and the clinical grey zone in autoimmune disease management. *Autoimmun Rev* 2017;16:911–24.

5 Aringer M, Johnson SR. Classifying and diagnosing systemic lupus erythematosus in the 21st century. *Rheumatology (Oxford)* 2020;59:v4–11.

6 Kay J, Upchurch KS. ACR/EULAR 2010 rheumatoid arthritis classification criteria. *Rheumatology (Oxford)* 2012;51 Suppl 6:vi5–9.

7 Hochberg MC. Updating the American college of rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;40:1725.

8 Tan EM, Cohen AS, Fries JF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1982;25:1271–7.

9 Shiboski CH, Shiboski SC, Seror R, et al. 2016 American college of rheumatology/European league against rheumatism classification criteria for primary sjögren's syndrome: a consensus and data-driven methodology involving three international patient cohorts. *Arthritis Rheumatol* 2017;69:35–45.

10 Lundberg IE, Tjärnlund A, Bottai M, et al. 2017 European league against Rheumatism/American college of rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups. *Ann Rheum Dis* 2017;76:1955–64.

11 Petri M, Orbai A-M, Alarcón GS, et al. Derivation and validation of the systemic lupus international collaborating clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum* 2012;64:2677–86.

12 Aringer M, Costenbader K, Daikh D, et al. 2019 European league against rheumatism/American college of rheumatology classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* 2019;78:1151–9.

13 Ehrenstein MR, Shipa M. SLE is not a one-size-fits-all disease. *J Exp Med* 2023;220:e20230559.

14 Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375:1216–9.

15 Vera Cruz G, Bucourt E, Réveillère C, et al. Machine learning reveals the most important psychological and social variables predicting the differential diagnosis of rheumatic and musculoskeletal diseases. *Rheumatol Int* 2022;42:1053–62.

16 Adamichou C, Genitsaridi I, Nikolopoulos D, et al. Lupus or not? SLE risk probability index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Ann Rheum Dis* 2021;80:758–66.

17 Soret P, Le Dantec C, Desvaux E, et al. A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nat Commun* 2021;12:3523.

18 Pinal-Fernandez I, Casal-Dominguez M, Derfoul A, et al. Machine learning algorithms reveal unique gene expression profiles in muscle biopsies from patients with different types of myositis. *Ann Rheum Dis* 2020;79:1234–42.

19 Kedra J, Davergne T, Braithwaite B, et al. Machine learning approaches to improve disease management of patients with rheumatoid arthritis: review and future directions. *Expert Rev Clin Immunol* 2021;17:1311–21.

20 Kegerreis B, Catalina MD, Bachali P, et al. Machine learning approaches to predict lupus disease activity from gene expression data. *Sci Rep* 2019;9:9617.

21 Guan Y, Zhang H, Quang D, et al. Machine learning to predict anti-tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. *Arthritis Rheumatol* 2019;71:1987–96.

22 Choi MY, Chen I, Clarke AE, et al. Machine learning identifies clusters of longitudinal autoantibody profiles predictive of systemic lupus erythematosus disease outcomes. *Ann Rheum Dis* 2023;82:927–36.

23 Lezcano-Valverde JM, Salazar F, León L, et al. Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach. *Sci Rep* 2017;7:10189.

24 Kingsmore KM, Puglisi CE, Grammer AC, et al. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nat Rev Rheumatol* 2021;17:710–30.

25 Mjaavatten MD, Bykerk VP. Early rheumatoid arthritis: the performance of the 2010 ACR/EULAR criteria for diagnosing RA. *Best Pract Res Clin Rheumatol* 2013;27:451–66.

26 Bellando-Randone S, Matucci-Cerinic M. Very early systemic sclerosis. *Best Pract Res Clin Rheumatol* 2019;33:101428.

27 Pons-Estel GJ, Ugarte-Gil MF, Harvey GB, et al. Applying the 2019 EULAR/ACR lupus criteria to patients from an established cohort: a Latin American perspective. *RMD Open* 2020;6:e001097.

28 Inês L, Silva C, Galindo M, et al. Classification of systemic lupus erythematosus: systemic lupus international collaborating clinics versus American college of rheumatology criteria. a comparative study of 2,055 patients from a real-life, international systemic lupus erythematosus cohort. *Arthritis Care Res (Hoboken)* 2015;67:1180–5.

29 Davies K, Dures E, Ng W-F. Fatigue in inflammatory rheumatic diseases: current knowledge and areas for future research. *Nat Rev Rheumatol* 2021;17:651–64.

30 Szekanecz Z, McInnes IB, Schett G, et al. Autoinflammation and autoimmunity across rheumatic and musculoskeletal diseases. *Nat Rev Rheumatol* 2021;17:585–95.

31 Bertsias GK, Pamfil C, Fanouriakis A, et al. Diagnostic criteria for systemic lupus erythematosus: has the time come? *Nat Rev Rheumatol* 2013;9:687–94.

32 Li J, Liu K, Ji Z, et al. Serum untargeted metabolomics reveal metabolic alteration of non-small cell lung cancer and refine disease detection. *Cancer Sci* 2023;114:680–9.

33 Papini S, Norman SB, Campbell-Sills L, et al. Development and validation of a machine learning prediction model of posttraumatic stress disorder after military deployment. *JAMA Netw Open* 2023;6:e2321273.

34 Dros JT, Bos I, Bennis FC, et al. Detection of primary Sjögren's syndrome in primary care: developing a classification model with the use of routine healthcare data and machine learning. *BMC Prim Care* 2022;23:199.

35 Forrest IS, Petrazzini BO, Duffy Á, et al. A machine learning model identifies patients in need of autoimmune disease testing using electronic health records. *Nat Commun* 2023;14:2385.

36 Burlina P, Billings S, Joshi N, et al. Automated diagnosis of myositis from muscle ultrasound: exploring the use of machine learning and deep learning methods. *PLoS One* 2017;12:e0184059.

37 Noguchi K, Saito I, Namiki T, et al. Reliability of non-contact tongue diagnosis for Sjögren's syndrome using machine learning method. *Sci Rep* 2023;13:1334.

38 Yang K, Wang Q, Wu L, et al. Development and verification of a combined diagnostic model for primary Sjögren's syndrome by integrated bioinformatics analysis and machine learning. *Sci Rep* 2023;13.

39 Bossuyt X, De Langhe E, Borghi MO, et al. Understanding and interpreting antinuclear antibody tests in systemic rheumatic diseases. *Nat Rev Rheumatol* 2020;16:715–26.

40 Gatto M, Iaccarino L, Ghirardello A, et al. Clinical and pathologic considerations of the qualitative and quantitative aspects of lupus nephritogenic autoantibodies: a comprehensive review. *J Autoimmun* 2016;69:1–11.

41 Agmon-Levin N, Damoiseaux J, Kallenberg C, et al. International recommendations for the assessment of autoantibodies to cellular antigens referred to as anti-nuclear antibodies. *Ann Rheum Dis* 2014;73:17–23.

42 Damoiseaux J, Andrade LEC, Carballo OG, et al. Clinical relevance of HEp-2 indirect immunofluorescent patterns: the international consensus on ANA patterns (ICAP) perspective. *Ann Rheum Dis* 2019;78:879–89.

43 Kaplan MJ. Neutrophils in the pathogenesis and manifestations of SLE. *Nat Rev Rheumatol* 2011;7:691–9.