



A multi-criterion feature integration framework for accurate diagnosis of Sjögren's disease using routine laboratory tests



Shu Liu^{1,5}, Guanhao Wu^{1,5}, Mengge Pan^{1,5}, Qi Sun², Chenjing Gao¹, Xianming Long³, Chaogui Tang⁴, Xinran Yuan¹✉ & Lingyun Sun¹

Sjögren's disease (SjD) is a common systemic autoimmune disease that remains difficult to diagnose early due to non-specific symptoms and lack of definitive biomarkers. This multicentre retrospective cohort study analysed data from 34,958 patients across three hospitals in China to assess the diagnostic potential of routine laboratory tests and to develop a robust, low-cost artificial intelligence model, the Sjögren Multi-criterion Feature Integration Framework (SMFIF). This study is registered with ClinicalTrials.gov (NCT06982482). The model was built using 16 optimal features selected through ensemble learning and SHAP analysis, and validated internally ($n = 9329$) and externally ($n = 545$). SMFIF achieved high diagnostic performance, with AUCs of 0.929, 0.934, and 0.964 in testing, internal, and external validation sets, respectively, outperforming conventional biomarkers such as anti-SSA/Ro and ANA. Calibration curves and confusion matrices confirmed its reliability. SMFIF is publicly available and provides probabilities of SjD based on laboratory data, offering a practical diagnostic tool for clinical use.

Sjögren disease (SjD) is one of the most common systemic autoimmune diseases, yet effective targeted therapies remain elusive^{1,2}. The prevalence of SjD ranges from 0.01% to 0.05%, mostly affecting adult females^{3,4}. The clinical course of SjD is characterised by significant heterogeneity, ranging from mild, gradually progressive exocrine dysfunction to severe systemic manifestations. Extra-glandular involvement is frequently observed and often precedes diagnosis in nearly half of all patients⁵. People with or without extra-glandular involvement have a lower quality of life compared to the general population^{3,6–8}. A comprehensive epidemiological meta-analysis has highlighted an increased incidence of malignancies (excluding lymphoma), infections, and cardiovascular diseases among individuals with SjD^{9,10}. However, diagnostic delay is common among patients with SjD¹¹. In a substantial international cohort of SjD patients, 7–8% lacked ocular or oral symptoms at the time of diagnosis, resulting in frequent misdiagnoses^{12,13}. A study of 524 individuals presenting with dry eyes and dry mouth revealed that 75.2% had initially been misdiagnosed with rheumatoid arthritis (RA),

systemic lupus erythematosus (SLE), or systemic sclerosis (SSc). Of these, 46.5% were subsequently correctly identified as having SjD¹⁴. These findings emphasise the critical importance of early and accurate recognition of SjD in order to improve patient outcomes.

The clinical diagnosis of SjD relies on expert clinical assessment. While established classification criteria, such as those from the American College of Rheumatology (ACR)-European League Against Rheumatism (EULAR), are valuable research tools for defining standardized cohorts, they are not used for individual patient diagnosis^{2,15,16}. These classification criteria prominently feature the presence of anti-SSA/Ro autoantibody as a key component. However, only 75% of patients with clinically diagnosed SjD test positive for these antibodies¹⁷. Clinicians often utilize several elements incorporated within these research criteria to support the diagnostic evaluation. These include positive findings on minor salivary gland biopsy demonstrating focal lymphocytic sialadenitis, reduced tear and salivary flow rates as assessed by Schirmer's test, and abnormal findings on salivary gland

¹Department of Rheumatology and Immunology, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, Jiangsu, China.

²Department of Endocrinology and Metabolism, Affiliated Hospital of Nantong University, Nantong, Jiangsu, China. ³The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, Jiangsu, China. ⁴The Affiliated Huai'an No. 1 People's Hospital of Nanjing Medical University, Huai'an First People's Hospital, Nanjing Medical University, Huai'an, Jiangsu, China. ⁵These authors contributed equally: Shu Liu, Guanhao Wu, Mengge Pan. ✉e-mail: xinran_yuan@outlook.com

imaging like delayed contrast elimination on sialography. Other commonly observed laboratory abnormalities in SjD include antinuclear antibodies (ANA), low complement 4 (C4) levels, rheumatoid factor (RF), and elevated immunoglobulin G (IgG)^{18–23}. The lack of a single, highly specific and sensitive biomarker, combined with the high cost of certain tests and the invasiveness of some procedures (like biopsy), contributes significantly to the challenges in diagnosis, leading to underdiagnosis and misdiagnosis.

Advancements in artificial intelligence (AI) and the growing availability of big data have opened new avenues for disease screening. The present retrospective multicentre study was conducted with the objective of developing a robust, cost-effective and generalisable Sjögren Multi-criterion Feature Integration Framework (SMFIF) that leverages routine laboratory tests to assist in the identification of SjD.

Results

Baseline characteristics of the study population and cohort composition

Between January 1, 2013, and January 1, 2023, a total of 34,958 individuals were finally included (Fig. 1). The DTH group (training set) included a total of 17,558 participants, of whom 9685 were in the Sjögren's group and 7873 were in the control group. The DTH cohort (testing set) comprised 7526 participants, with 3429 in the Sjögren's group and 4097 in the control group. The internal validation set included 9329 participants, of which 4203 were in the Sjögren's group and 5126 in the control group. The external validation set, the SMH and HFPN cohort, consisted of 545 participants, with 139 in the Sjögren's group and 406 in the control group. The baseline characteristics of the study population are summarized (Table 1). The baseline data results from the three datasets showed significant differences between the control and the SjD group in many haematological characteristics (Supplementary Table 2). The median age of the testing cohort was 52 years. Of these, 89.9% were female and 8.7% were smokers. In the internal validation cohort, the median age was 53 years, 94.7% were female, and 12.8% were smokers. In the external validation cohort, the median age was similar to the

testing cohort at 56 years. 91.4% were female and 9.3% were smokers. The median symptom duration in SjD group was longer across all sets (testing: 3.5 years; internal validation: 4.0 years; external validation: 2.8 years) compared to controls (0.8 years, 0.6 years, 0.6 years). Organ involvement was elevated in SjD groups, with higher rates of liver (6.5–9.0% vs 3.0–5.0%), renal (14.4–17.0% vs 4.9–8.0%), lung (7.9–13.0% vs 3.0–5.0%), cardiovascular (18.0–25.5% vs 16.0–19.0%), and metabolic disorders (17.3–19.0% vs 13.5–15.0%) versus controls. Based on further tests, there were 21.6%, 19.7% and 23.0% with swollen parotid gland; 89.1%, 88.3% and 90.4% with dry eye syndrome; 60.8%, 55.9% and 29.6% with positive parotid sialography; 30.0%, 23.5% and 21.2% with positive minor salivary glands in the testing, internal validation and external validation cohorts, respectively.

Feature selection strategy and optimization outcomes

The construction of SMFIF model adopts the method of a multi-criterion feature integration framework (Fig. 2; Supplementary Table 3). Initially, the optimal features and intersecting features were selected through feature selection algorithms for model construction. A detailed evaluation of the performance of the models and the SHAP dependence plots of the relevant features is provided in the supplementary materials (Supplementary Tables 5–11; Supplementary Figs. 1 and 2). The SHAP values computed for the ML models built using all features are shown (Fig. 3A–G), while the performance of the seven models is summarized in the supplementary materials (Supplementary Table 4). Correlation analysis revealed significant relationships among most features (Fig. 3H). Through the above analysis (refer to the Methods section), 16 core features were identified and selected as the final input for SMFIF: Creatinine (Crea), γ -glutamyl transferase (GGT), Uric acid (UA), Total protein (TP), Apolipoprotein AI (apoAI), Alanine transaminase (ALT), Phosphorus (P), Adenosine deaminase (ADA), Glucose (Glu), Direct bilirubin (DBIL), Chloride ion (Cl^-), Globulin (GLO), Alkaline phosphatase (ALP), estimated Glomerular Filtration Rate (eGFR), High density lipoprotein cholesterol (HDL-C), and Complement 4 (C4). The SHAP

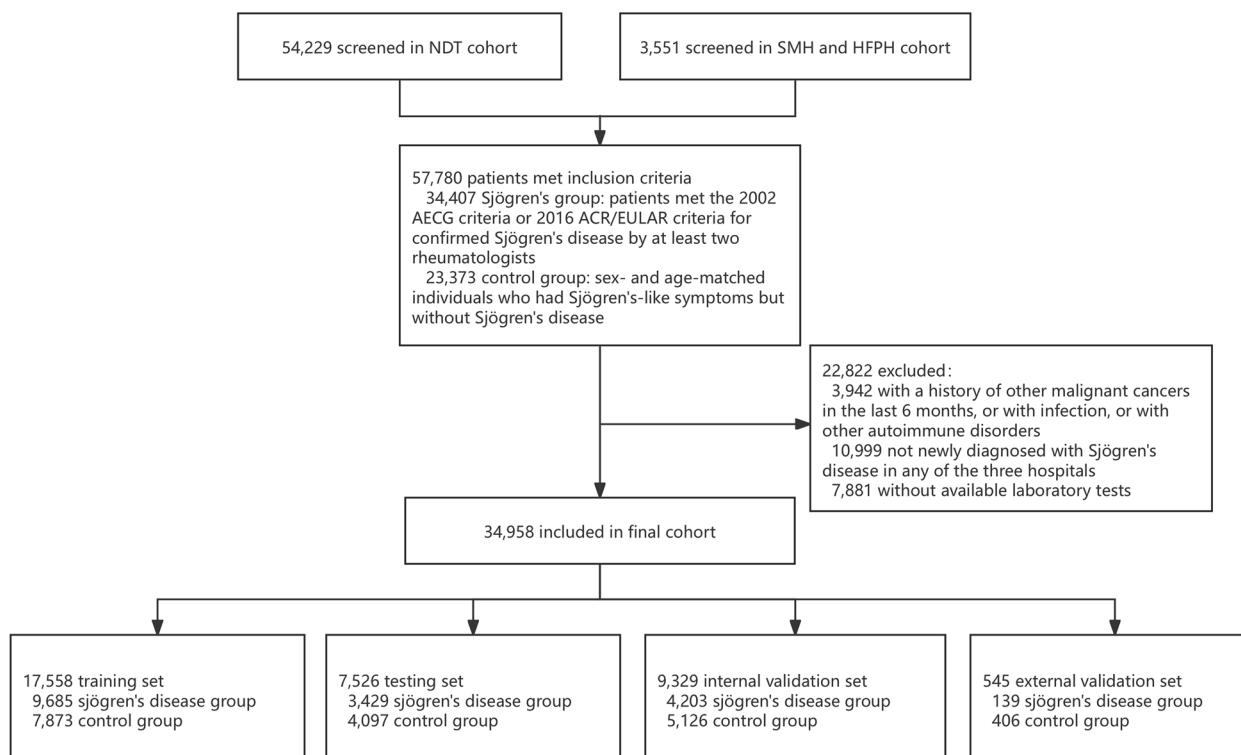


Fig. 1 | Study flowchart. NDT cohort = patients from Nanjing Drum Tower Hospital, the affiliated hospital of Nanjing University Medical School. SMH cohort = patients from Suzhou Municipal Hospital, the affiliated Suzhou hospital of Nanjing

Medical University. HFPN cohort = patients from the Huai'an First People's Hospital, the affiliated Huai'an No. 1 People's Hospital of Nanjing Medical University.

Table 1 | Baseline characteristics of included population

Characteristics	Testing set NDT cohort = 7526	Validation set NDT cohort = 9329	External validation set SMH and HFPH cohort = 545
Sjögren's disease	3429 (45.1%)	4203 (45.1%)	139 (25.5%)
Non-Sjögren's disease	4097 (54.9%)	5126 (54.9%)	406 (74.5%)
Age, years			
Sjögren's disease	52 (48–62)	53 (41–63)	56 (47–64)
Control groups	39 (24–49)	37 (29–47)	45 (26–51)
Female			
Sjögren's disease	3048/3429 (89.9%)	3981/4203 (94.7%)	127/139 (91.4%)
Control groups	3585/4097 (86.7%)	4793 /5126 (93.5%)	334/406 (82.3%)
Smokers			
Sjögren's disease	295/3429 (8.7%)	541/4203 (12.8%)	13/139 (9.3%)
Control groups	153/4097 (3.7%)	283/5126 (5.5%)	14/406 (3.4%)
Symptom duration, years			
Sjögren's disease	3.5 (0.5–8.0)	4.0 (1.0–9.5)	2.8 (0.3–6.0)
Control groups	0.8 (0.1–3.0)	0.6 (0.1–2.5)	0.6 (0.1–1.5)
Co-morbid conditions			
Sjögren's disease			
Liver diseases	254/3429 (7.5%)	378/4203 (9.0%)	9/139 (6.5%)
Renal diseases	543/3429 (16.0%)	714/4203 (17.0%)	20/139 (14.4%)
Cardiovascular diseases	712/3429 (21.0%)	1073/4203 (25.5%)	25/139 (18.0%)
Lung diseases	356/3429 (10.5%)	546/4203 (13.0%)	11/139 (7.9%)
Metabolic disorders	627/3429 (18.5%)	799/4203 (19.0%)	24/139 (17.3%)
Control groups			
Liver diseases	166/4097 (4.0%)	256/5126 (5.0%)	12/406 (3.0%)
Renal diseases	207/4097 (5.0%)	410/5126 (8.0%)	20/406 (4.9%)
Cardiovascular diseases	703/4097 (17.0%)	973/5126 (19.0%)	65/406 (16.0%)
Lung diseases	124/4097 (3.0%)	256/5126 (5.0%)	16/406 (3.9%)
Metabolic disorders	558/4097 (13.5%)	769/5126 (15.0%)	57/406 (14.0%)
Swollen parotid gland			
Sjögren's disease	732/3429 (21.6%)	830/4203 (19.7%)	32/139 (23.0%)
Control groups	82/4097 (2.0%)	301/5126 (5.8%)	8/406 (2.0%)
Dry eyes syndrome ^a			
Sjögren's disease	3021/3429 (89.1%)	3712/4203 (88.3%)	126/139 (90.4%)
Control groups	2406/4097 (58.2%)	1724/5126 (33.6%)	263/406 (64.8%)
Parotid sialography (+) ^b			
Sjögren's disease	2061/3429 (60.8%)	2353/4203 (55.9%)	41/139 (29.6%)
Control groups	459/4097 (11.1%)	589/5126 (11.5%)	58/406 (14.3%)
Chisholm stage ^c			
Sjögren's disease			
Stage I	468/3429 (13.8%)	410/4203 (9.8%)	3/139 (2.2%)
Stage II	231/3429 (6.8%)	90/4203 (2.1%)	9/139 (6.5%)
Stage III	288/3429 (8.5%)	219/4203 (5.2%)	4/139 (2.9%)
Stage IV	729/3429 (21.5%)	769/4203 (18.3%)	27/139 (19.3%)
N/A	1675/3429 (49.4%)	2715/4203 (64.6%)	96/139 (69.1%)
Control groups			
Stage I	798/4097 (19.3%)	761/5126 (14.8%)	65/406 (16.0%)
Stage II	289/4097 (7.0%)	294/5126 (5.7%)	7/406 (1.7%)
Stage III	210/4097 (5.1%)	78/5126 (1.5%)	24/406 (6.0%)
Stage IV	211/4097 (5.1%)	211/5126 (4.1%)	20/406 (4.9%)

Table 1 (continued) | Baseline characteristics of included population

Characteristics	Testing set NDT cohort = 7526	Validation set NDT cohort = 9329	External validation set SMH and HFPH cohort = 545
N/A	2627/4097 (63.5%)	3782/5126 (73.8%)	290/406 (71.4%)

Data are n (%) or median (IQR). NDT cohort = patients from Nanjing Drum Tower Hospital, the affiliated hospital of Nanjing University Medical School. SMH cohort = patients from Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University. HFPH cohort = patients from the Huai'an First People's Hospital, The Affiliated Huai'an No. 1 People's Hospital of Nanjing Medical University.

^aAn ocular surface disease index of ≥ 3 in either eye, or the Schirmer test was considered positive if the filter paper was wet to an extent of ≤ 5 mm after 5 min confirmed a diagnosis of dry eyes syndrome.

^bA positive diagnosis was confirmed when the iodine formed globular pools in the terminal glands, with a tree- or snowflake-like pattern but the dominant duct was not obstructed.

^cLabial gland biopsies with a pathological grade of at least grade III were considered as being SjD positive based on the Chisholm criteria.

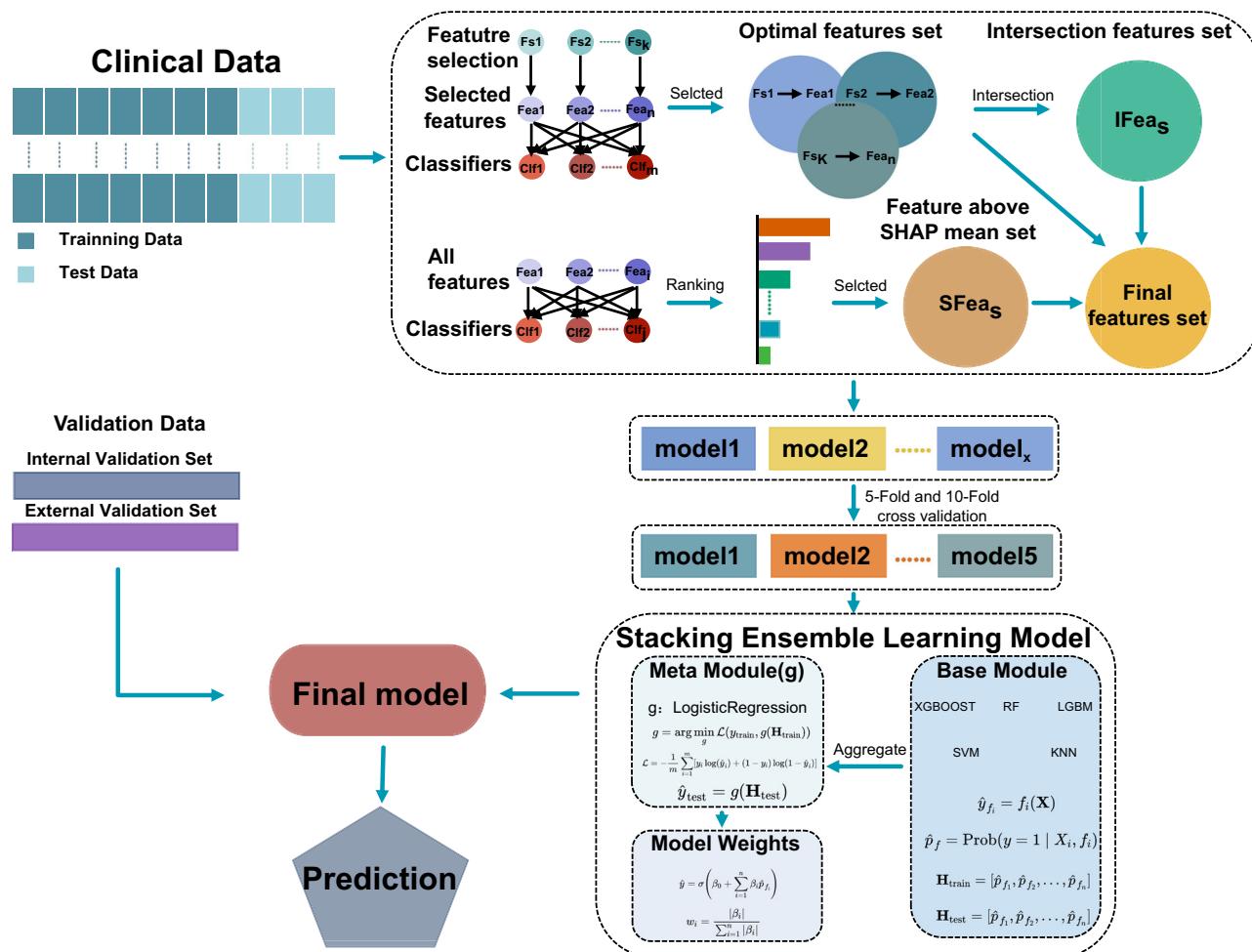


Fig. 2 | Graphical illustration of the construction and validation process

of SMFIF. The SMFIF model was developed using a systematic approach, involving meticulous preprocessing of clinical and laboratory data, feature selection with six methods, and the creation of 42 feature-model combinations. After refining feature

subsets, a final set of 16 key features was identified and used to train seven classifiers. The top five models were selected for ensemble construction using a weighted approach, and the model's performance was rigorously validated with internal and external datasets.

dependence plots for these features are included in the supplementary materials (Supplementary Fig. 3).

In our study, 16 features were finally used to calculate the comprehensive performance indicators of each classifier in all datasets. Compare and analyse the top five classifiers with the highest performance as base classification models, integrated into a stacked ensemble learning model (Supplementary Table 12). This ensemble model synthesises predictions from meta classifiers and ultimately constructs the SMFIF model. To rigorously evaluate the performance of the SMFIF model, five-fold and ten-fold cross-validation and 95% CI analysis were performed on all performance evaluation metrics of the model. SMFIF model has been made publicly available via GitHub.

Performance of the SMFIF model and multicenter validation results

The SMFIF model showed exemplary predictive ability in discriminating patients with SjD from those with SjD-like manifestations without SjD. The AUC for the test set was 0.923 (95% CI, 0.923–0.935), for the internal validation set was 0.934 (95%CI, 0.929–0.939) and for the external validation set was 0.964 (95%CI, 0.943–0.986) (Table 2). The SMFIF model showed superior and more stable performance compared to all previously established models, especially in terms of AUC, accuracy and specificity. In addition, a comparative analysis was performed between 16 feature-based SMFIF and the independent conventional marker-based SMFIF. The results showed that 16 feature-based SMFIF

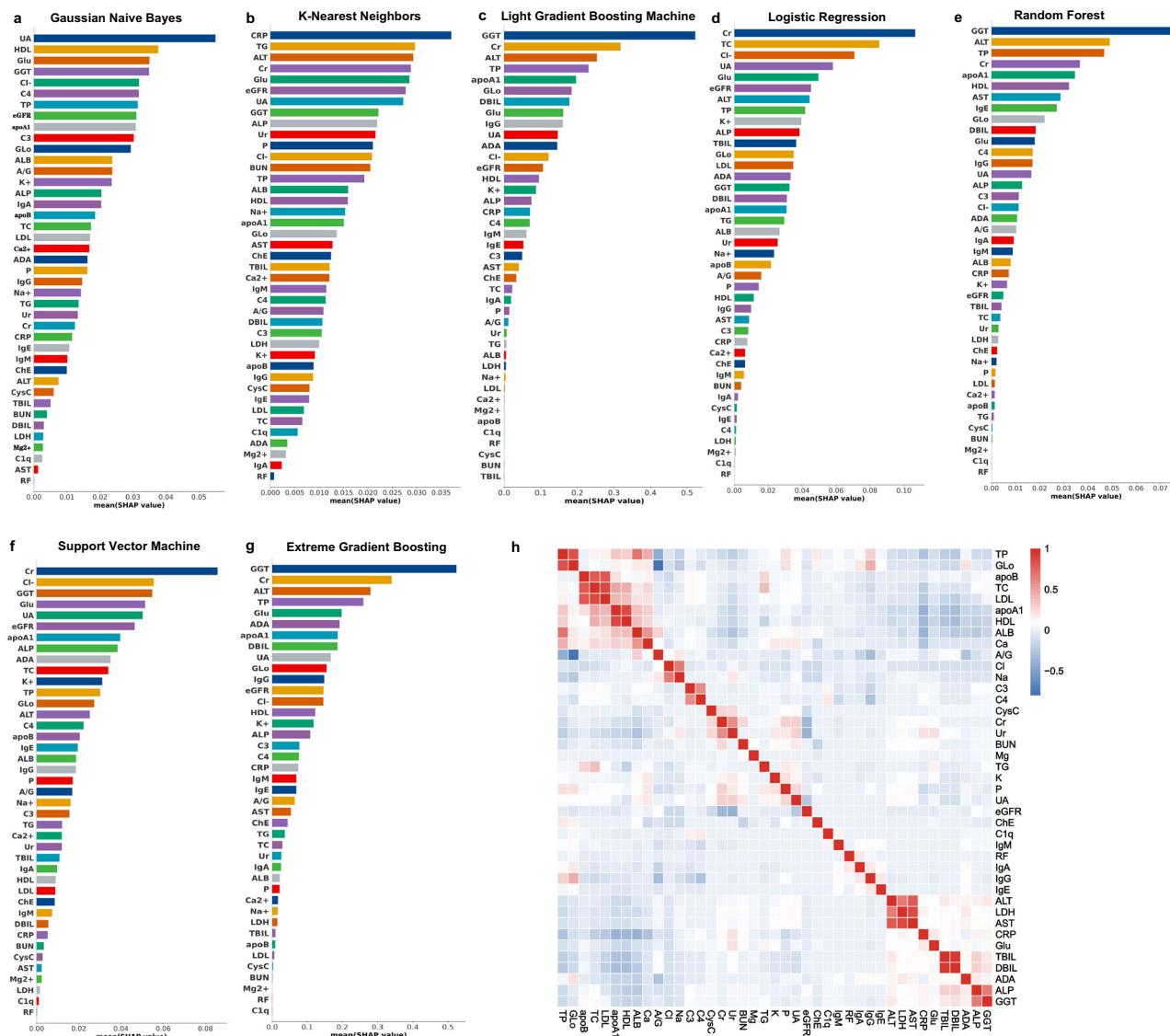


Fig. 3 | SHapley Additive exPlanations plot: the impact of clinical features for diagnosing Sjögren's disease in the Gaussian Naive Bayes, K-Nearest Neighbors, Light Gradient Boosting Machine, Logistic Regression, Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Logistic Regression.

a Gaussian Naive Bayes, b K-Nearest Neighbors, c Light Gradient Boosting Machine, d Logistic Regression, e Random Forest, f Support Vector Machine, g Extreme Gradient Boosting, h Pearson correlation coefficients between the included features. Red represents a positive correlation between the two features. The redder the colour is, the higher the positive correlation coefficient is. Blue represents a negative correlation. The bluer the colour is, the higher the negative correlation coefficient is. All the features were clustered into two groups according to the correlation. TP Total protein, GLO Globulin, apoB Apolipoprotein B, TC Total cholesterol, LDL-C Low

density lipoprotein cholesterol, apoA1 Apolipoprotein AI, HDL-C High density lipoprotein cholesterol, ALB Albumin, Ca2+ Calcium, A/G Albumin to globulin ratio, Cl- Chloride ion, Na+ Sodium ion, C3 Complement 3, C4 Complement 4, CysC Cystatin C, Cr Creatinine, Ur Urea, BUN Blood urea nitrogen, Mg2+ Magnesium ion, TG Triglycerides, K+ Potassium ion, P Phosphorus, UA Uric acid, eGFR Estimated Glomerular Filtration Rate, ChE Cholinesterase, C1q Cholinesterase, IgM Immunoglobulin M, RF Rheumatoid factor, IgA Immunoglobulin A, IgG Immunoglobulin G, IgE Immunoglobulin E, ALT Alanine transaminase, LDH Lactate dehydrogenase, AST Aspartate transaminase, CRP C-reactive protein, Glu Glucose, TBIL Total bilirubin, DBIL Direct bilirubin, ADA Adenosine deaminase, ALP Alkaline phosphatase, GGT γ -glutamyl transferase.

outperformed those based on ANA (AUC: 0.709; 0.728–0.846), SSA (AUC: 0.705; 0.703–0.699), SSB (AUC: 0.575; 0.569–0.583), RF (AUC: 0.506; 0.506–0.529), GLO (AUC: 0.688; 0.690–0.592), IgG (AUC: 0.677; 0.685–0.740), C3 (AUC: 0.672; 0.682–0.736) and C4 (AUC: 0.671; 0.683–0.621) in testing, internal validation and external validation, respectively (Fig. 4).

The calibration curves for the testing set, internal validation set, and external validation set demonstrate a generally good agreement between predicted probabilities and actual outcomes, indicating that the SMFIF model is well-calibrated across different datasets. Notably, the calibration curves for the testing and internal validation sets show a high degree of consistency, suggesting strong generalizability of the

model to unseen data. However, the calibration curve for the external validation set exhibits greater variability, indicating potential challenges in adapting to external data distributions. In terms of classification performance, the confusion matrices across the testing, internal validation and external validation data sets consistently show high sensitivity and specificity with minimal false classifications, highlighting the robust and reliable diagnostic performance of the model across all evaluation phases (Fig. 5).

Overall, these results indicate that the SMFIF model provides robust and reliable probabilistic predictions with strong diagnostic performance. Nevertheless, while the model generalizes well to internal validation and testing data, its performance on external validation data suggests that further

Table 2 | Performance of SMFIF for classifying Sjögren's disease versus controls

	Testing set (3429 vs 4097)	Internal validation set (4203 vs 5126)	External validation set (139 vs 406)
ACC	0.858(0.850–0.866)	0.859 (0.852–0.866)	0.960(0.942–0.973)
Recall	0.832(0.821–0.845)	0.825 (0.813–0.833)	0.871(0.822–0.917)
F1-score	0.842(0.832–0.851)	0.841 (0.833–0.849)	0.917(0.883–0.947)
PPV	0.852(0.839–0.862)	0.857 (0.847–0.867)	0.968(0.929–0.992)
AUC	0.929(0.923–0.935)	0.934 (0.929–0.939)	0.964(0.943–0.986)
Specificity	0.879(0.869–0.888)	0.887 (0.880–0.897)	0.990(0.979–0.998)

Data are metrics (95% CI).

ACC accuracy, AUC area under the receiver-operating characteristic curve, PPV positive predicted value.

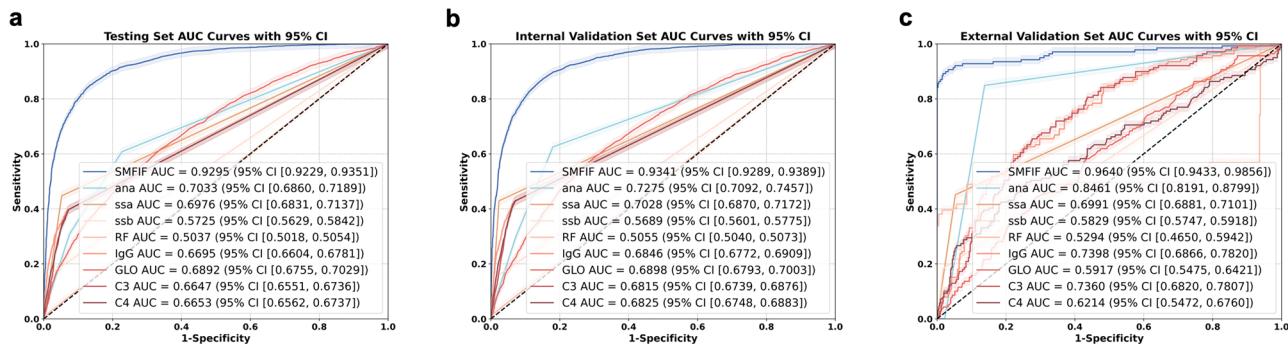


Fig. 4 | Diagnostic performance of the final machine learning models for Sjögren's disease in testing, internal validation and external validation cohorts vs. SSA/SSB/ANA/RF/IgG/GLO model. a Testing cohort, **b** Internal Validation cohort, **c** External Validation cohort. AUC area under the receiver-operating characteristic curve.

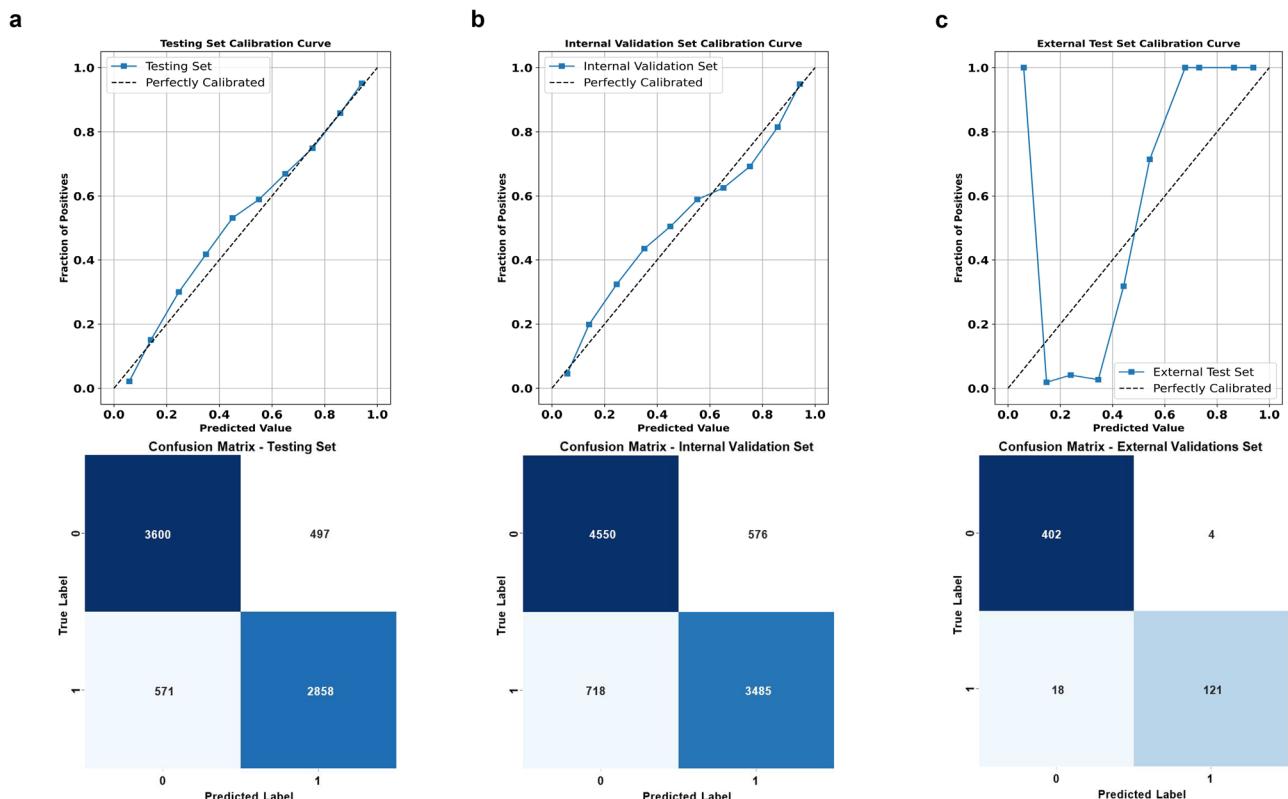


Fig. 5 | Calibration curves and confusion matrix to evaluate performance of SMFIF. a Testing cohort, **b** Internal validation cohort, **c** External validation cohort.

refinement may be needed to enhance sensitivity when applied to different external conditions.

Discussion

In this multicenter retrospective study, we screened 70 features from 123 routine blood test results through data preprocessing. By multi-algorithm feature selection, SHAP value analysis, model performance evaluation, and ensemble learning strategies, we developed the SMFIF model based on 16 optimal features for accurate SjD diagnosis and prediction, the SMFIF achieved consistent and good performance both in the internal set and the external validation sets, and outperformed anti-SSA/Ro, anti-SSB/La, ANA in identifying SjD-like individuals with atypical symptoms. The model was developed into a user-friendly prediction tool, which is publicly available and provides an estimated probability of SjD based on routine laboratory test results. To our knowledge, this study presents the first AI machine learning model developed for diagnosing SjD using big data in clinical practice.

The prevalence of SjD varies widely in the population. Due to the lack of typical early symptoms and its slow progression, especially when combined with smoking, diabetes, menopausal syndrome, or infection, clinicians in primary care settings fail to identify the primary disease. As a result, patients may already have interstitial lung disease, arthritis, polyneuropathy or even malignancies when a SjD diagnosis is made, leading to delays in treatment of the primary disease in SjD. One advantage of the SMFIF model is that it provides clinicians with a tool that can help them identify patients who potentially have SjD, especially in primary care settings where routine physical examinations or clinical experience with rheumatological diseases is limited. Compared with SMFIF_{ANA}, SMFIF_{anti-SSA/Ro}, SMFIF_{anti-SSB/La} and SMFIF_{RF/Glo/IgG/C3/C4}, the combination of laboratory tests and auto-antibodies contained in SMFIF has improved predictive performance, as indicated by the increased AUC and sensitivity, with a much lower cost than conventional autoantibodies, labial gland biopsy, parotid sialography or ultrasonography, tear flow rate and corneal fluorescein staining.

One important factor in the generalisability and stability of our prediction model is the representativeness of the data. We first trained, tested, and internally validated the model using data from a large regional hospital, and selected two independent local mergers as primary care providers for external validation. In the three cohorts, the median age at diagnosis of SjD was 52–56 years, which is representative of the Chinese and other Asian populations^{24,25}. Another important factor in improving the predictive performance of the model is our proposed multi-criterion feature integration strategy. The approach not only ensures the identification of the most relevant and robust features but also significantly enhances the model's interpretability and generalizability. By systematically integrating the results from multiple feature selection algorithms and SHAP-based evaluations, we derived a final feature set comprising 16 key features. Such optimization effectively minimizes feature redundancy while maximizing the predictive power of the selected features. It is worth noting that our final 16-feature model—which primarily includes non-classical immune biomarkers of liver, kidney, and metabolic function—reflects both the data-driven feature selection process and the systemic pathophysiological characteristics of SjD. Although baseline analysis revealed significant differences between the control and the SjD group in terms of classical immune biomarkers, multivariate feature selection prioritised conventionally available and cost-effective indicators after rigorous multi-algorithm filtering and SHAP-based ranking. Most cases in the control group were clinically suspected of SjD but were ultimately diagnosed with various chronic diseases, ensuring the model could distinguish SjD from complex non-autoimmune diseases rather than healthy states. Additionally, the SjD group exhibited higher actual organ involvement (co-morbid conditions) compared to the control group, indicating that autoimmune disease-specific biochemical abnormalities play a critical role in disease progression, while the value of routine blood tests may be significantly underestimated.

Further, to enhance interpretability in ambiguous cases, we incorporated patient-level SHAP decision plots to illustrate how the SMFIF model makes predictions. We analysed all misclassified samples and visualized representative cases from the false negative (FN) and false positive (FP) groups (Supplementary Fig. 5). Notably, many of the FN cases exhibit atypical values for canonical SjD indicators, but these were overridden by strong negative contributions from renal or hepatic function markers, thereby leading to a non-SjD prediction. In contrast, some FP cases present borderline elevations in immune-related indicators, which could reasonably lead to model suspicion of SjD.

Given the development of big data in genomics, transcriptomics, proteomics, imaging, therapeutics, and electronic health records (EHRs), machine learning models are extremely attractive in the field of rheumatic diseases such as SLE and RA^{26,27}. A few studies have been already conducted in SjD. Dros et al. developed a classification model using primary care electronic health records (EHRs) data and hospital claims data (HCD) from 1411 SjD patients. The AUC of logistic regression (LR) and random forest (RF) was 0.82 and 0.84, respectively²⁸. Troncoso et al. utilize an auto-machine learning (autoML) platform for the automated segmentation and quantification of Focus Score (FS) on histopathological slides, aiming to augment diagnostic precision and speed in SS²⁹. Marlin et al. developed an AI-enabled algorithm that automatically identifies key histological features of the minor salivary gland of SS patients³⁰. Wu et al. developed a graph-learning-based model named Cell-tissue-graph-based pathological image analysis model for automatic diagnosis of SS³¹. Marugán et al. applied ML methodologies to classify SLE and SjD patients and designed an analysis pipeline to train an eXtreme Gradient Boosting (XGBoost) multiclass predictor for each type of data³². However, all published studies above were limited by the inclusion of only general basic information or a single pathological type, small cohort sizes, and lack of external validation. In contrast, our study was conducted on a large-scale multicentre dataset that included multiple routine laboratory tests, autoantibodies, minor salivary gland function assessments and pathological types of labial gland biopsies. The clinical utility of SMFIF depends on its ability to distinguish SjD not only from healthy individuals but, more importantly, from patients with other immune-mediated and chronic conditions. We additionally applied the 16-feature model to four comparator conditions using database from DTH. The results demonstrate that SMFIF effectively distinguishes SjD from RA (AUC = 0.8336), SLE (AUC = 0.8896), SSc (AUC = 0.8416), and osteoarthritis (AUC = 0.9149) (Supplementary Table 13 and Supplementary Fig. 4). We interpret the strong performance against RA, SLE, SSc, and osteoarthritis (OA) as evidence that SMFIF captures a signature relatively specific to SjD pathology compared to these distinct inflammatory and non-inflammatory (OA) rheumatic disorders, significantly enhancing its potential clinical utility for differential diagnosis. The SMFIF modelling process used multi-algorithm feature selection, SHAP value analysis, model performance evaluation, and ensemble learning strategies, which has advantages over traditional logistic regression models, and was validated by an independent external large cohort, making the results more reliable.

Our study has several limitations. First, SMFIF was not evaluated in a prospective cohort. Second, loss of data was inevitable in this real-world retrospective study. Third, due to data-sharing restrictions and annotation costs, the cohort was exclusively derived from three Chinese hospitals, representing a single-nation dataset. This necessitated reliance on a second independent external validation cohort within the same healthcare system. These constraints collectively limit the extrapolation of our findings to other ethnic populations or distinct healthcare settings. Furthermore, outpatient recruitment was confined to two local hospitals, resulting in imbalanced sample sizes between internal and external validation sets. The model's generalizability across other countries and regions requires further investigation, including validation with more patients from primary care settings to

confirm its broader utility. Future multinational studies are essential to establish the model's global applicability. Fourth, multimodal features such as clinical symptoms, medical history characteristics, image characteristics, and genetic characteristics were not included. Finally, it must be acknowledged that the clinical application of the developed model still faces challenges due to practical considerations such as ethnic diversity, the attribution of clinical use, and full transparency of the decision-making process. Prospective cohorts, improved annotation of multimodal data, and more reasonable regulations for artificial intelligence are required to ensure better application of SMFIF.

In summary, SMFIF based on 16 key features of routine laboratory items achieved satisfactory and stable performance and was significantly better than the traditional markers SSA/Ro, SSB/La and ANA. This model provides a low-cost, accessible and accurate diagnostic tool for SjD. Prospective studies and data sets from other regions are needed to further confirm the feasibility and generalizability of SMFIF.

Methods

Study design and participants

This retrospective, multicentre study screened 198,922 laboratory examinations from three hospitals in China between Jan 1, 2013, and Jan 1, 2023. The 123 laboratory items were selected based on the tests' universality in routine clinical use by rheumatologists (Supplementary Table 1). The training (testing) set was selected from a large-scale regional centre hospital, Nanjing Drum Tower Hospital (NDT cohort). The external validation set was selected from two local hospitals, Suzhou Municipal Hospital (SMH cohort) and Huai'an First People's Hospital (HFPH cohort). All participants recruited were aged between 18 and 65 years and underwent pathological, haematological and imaging tests to confirm the presence of SjD. Laboratory tests were measured from one month before diagnosis to any treatment. Patients diagnosed with Sjögren's disease were included in the SjD group, while the control group comprised sex- and age-matched individuals exhibiting similar or suspected Sjögren-like symptoms, such as dry mouth and eyes, fatigue, joint pain, and weight loss, but who were excluded on screening for SjD and other autoimmune diseases. The study excluded cases of combined pregnancy, breastfeeding, a clear diagnosis of other autoimmune diseases, infection and malignant tumours. The diagnosis of SjD is based on an assessment by at least two rheumatologists using the 2002 American-European Consensus Group (AECG) or 2016 ACR/EULAR classification criteria^{15,16}. Meeting any one of the two classification criteria is considered a definite diagnosis.

Ethical statement

The study protocol was approved by the Medical Ethics Committee of Nanjing Drum Tower Hospital (B2022-529-04), the Medical Ethics Committee of the First People's Hospital of Huai'an City (YX-Z-2024-032-01), and the Medical Ethics Committee of the Suzhou Municipal Hospital (K-2022-074-K03). The study complied with the ethical principles of Declaration of Helsinki of 1975. All case data were anonymised, and the Institutional Review Board waived the requirement for written informed consent. This study is registered with ClinicalTrials.gov (Registration number: NCT06982482, Date of registration: May 13, 2025, Registry name: Artificial Intelligence Based Models for Primary Sjögren's Syndrome Diagnosis). A copy of the study protocol can be found in supplemental file.

Data preprocessing

Laboratory items with different units were unified. Characteristics of study cohorts in which more than 50% of missing data in each row were removed from analysis and more than 80% of missing data in each column after cleaning were removed. After data cleaning, missing values were imputed using the random forest imputation method, which leverages predictive modelling to estimate missing entries based on available features. To minimize institutional bias introduced by the multicenter nature of the dataset, we implemented a harmonization procedure prior to modeling.

Laboratory items from the three centres were first aligned using a shared data dictionary, ensuring consistent variable naming, definitions, and measurement units³³. Additionally, we visually inspected the distributions of key variables across cohorts and applied min-max normalization to reduce centre-specific variability³⁴. Data imbalance was resolved via the adaptive synthetic sampling method with a balancing ratio of 0.5³⁵.

Predictive modelling

After data preprocessing, the 70 laboratory items were used as the candidate features for the model building. The predictive model for SjD diagnosis was developed using a structured, methodologically rigorous framework (Fig. 2). Six feature selection algorithms were first applied to identify optimal feature subsets, which were then input into seven machine learning classifiers to calculate SHAP values, quantifying feature importance. After the first step of screening, 41 candidate features were obtained. To strengthen model robustness, the intersection of features from six algorithms formed a core feature set. Meanwhile, the performance of all features was assessed by constructing models using the seven classifiers, followed by ranking features based on their SHAP values. By integrating all feature sets, optimal feature subsets and the core feature set with SHAP-based rankings, and a final set of 16 significant features was established. To enhance the reliability of the features incorporated into the final model, we computed the Pearson correlation matrix for all candidate features. No pair of features exceeded the commonly accepted threshold of $r > 0.90$. Additionally, we calculated the Variance Inflation Factors (VIFs) for the features. Eleven features had a VIF of less than 5, and five features (apoA1, HDL, DBIL, TP and GLO) had a VIF between 5 and 10; the latter were retained due to their distinct clinical relevance. This achieves control over multicollinearity. The predictive model was then built using this curated set and trained with multiple classifiers. The top five performing models were selected as base classifiers and integrated into a stacking ensemble learning model, SMFIF, which employs a meta-classifier to further improve accuracy³⁶. The model's generalizability and accuracy were validated using both internal and external datasets, demonstrating its potential for clinical application. The model's generalizability and accuracy were validated using both internal and external datasets, demonstrating its potential for clinical application.

Performance evaluation

The performance of the SMFIF model was assessed by using seven significant metrics: area under the receiver operating characteristic curve (AUC), accuracy (ACC), recall, F1-score, positive predicted value (PPV), specificity, and negative predictive value (NPV). Independent classification models were constructed using feature inputs derived from the anti-SSA/Ro antibodies, anti-SSB/La antibodies, ANA, RF, GLO, IgG, C3, and C4 datasets, and their performance was compared to that of the SMFIF model. The evaluation of the SMFIF model was further augmented with calibration curves and confusion matrices to assess its clinical utility and reliability.

Feature evaluation

The primary outcome was the prediction accuracy of the model in identifying SjD, based on an optimally selected feature set to ensure the best predictive performance. The SHAP method was used to analyse the contribution of each feature in order to quantify its relative importance. The diagnostic performance of the SMFIF was found to be heavily influenced by the selection of input features. To optimize the model, 16 significant features were identified and prioritized based on their frequency of selection and importance rankings across all classification models.

Statistical analysis

Patient data were analysed separately for continuous and categorical variables, with results reported as median and interquartile range (IQR) for continuous variables and frequency (%) for categorical variables. Comparisons of continuous variables between groups were conducted using the Mann-Whitney U test, while categorical variables were compared using the

Chi-square test or Fisher's exact test, as appropriate. A P -value < 0.05 was considered indicative of statistical significance.

Data availability

The data that support the findings of this study are available on reasonable request from the corresponding author, X.Y. Please note that patient personal information and correspondence will not be shared, as they are protected by the corresponding authors' institutions to ensure patient privacy.

Code availability

Data analysis was conducted using R software (Version 4.3.2) and Python (Version 3.7). The core code for the research can be found on the corresponding author's GitHub page (<https://github.com/guanhaowu123/SMFIF>).

Received: 16 April 2025; Accepted: 22 October 2025;

Published online: 27 November 2025

References

1. Mariette, X. & Criswell, L. A. Primary Sjögren's syndrome. *N. Engl. J. Med.* **378**, 931–939 (2018).
2. Ramos-Casals, M. et al. 2023 International Rome consensus for the nomenclature of Sjögren disease. *Nat. Rev. Rheumatol.* **21**, 426–437 (2025).
3. Belenguer, R. et al. Influence of clinical and immunological parameters on the health-related quality of life of patients with primary Sjögren's syndrome. *Clin. Exp. Rheumatol.* **23**, 351–356 (2005).
4. Beydon, M. et al. Epidemiology of Sjögren syndrome. *Nat. Rev. Rheumatol.* **20**, 158–169 (2024).
5. Asmussen, K., Andersen, V., Bendixen, G., Schiødt, M. & Oxholm, P. A new model for classification of disease manifestations in primary Sjögren's syndrome: evaluation in a retrospective long-term study. *J. Intern. Med.* **239**, 475–482 (1996).
6. Huang, H., Xie, W., Geng, Y., Fan, Y. & Zhang, Z. Mortality in patients with primary Sjögren's syndrome: a systematic review and meta-analysis. *Rheumatology* **60**, 4029–4038 (2021).
7. Gairy, K., Knight, C., Anthony, P. & Hoskin, B. Burden of illness among subgroups of patients with primary Sjögren's syndrome and systemic involvement. *Rheumatology* **60**, 1871–1881 (2021).
8. Seror, R. et al. Outcome measures for primary Sjögren's syndrome: a comprehensive review. *J. Autoimmun.* **51**, 51–56 (2014).
9. Pego-Reigosa, J. M., Restrepo Vélez, J., Baldini, C. & Rúa-Figueroa Fernández de Larrinoa, I. Comorbidities (excluding lymphoma) in Sjögren's syndrome. *Rheumatology* **60**, 2075–2084 (2021).
10. Chatzis, L. G. et al. Clinical picture, outcome and predictive factors of lymphoma in primary Sjögren's syndrome: results from a harmonized dataset (1981–2021). *Rheumatology* **61**, 3576–3585 (2022).
11. Wang, B. et al. Early diagnosis and treatment for Sjögren's syndrome: current challenges, redefined disease stages and future prospects. *J. Autoimmun.* **117**, 102590 (2021).
12. Abrol, E., González-Pulido, C., Praena-Fernández, J. M. & Isenberg, D. A. A retrospective study of long-term outcomes in 152 patients with primary Sjögren's syndrome: 25-year experience. *Clin. Med.* **14**, 157–164 (2014).
13. Soret, P. et al. A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nat. Commun.* **12**, 3523 (2021).
14. Rasmussen, A. et al. Previous diagnosis of Sjögren's Syndrome as rheumatoid arthritis or systemic lupus erythematosus. *Rheumatology* **55**, 1195–1201 (2016).
15. Shibuski, C. H. 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.* **69**, 35–45 (2017).
16. Vitali, C. et al. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann. Rheum. Dis.* **61**, 554–558 (2002).
17. Retamozo, S. et al. Influence of the age at diagnosis in the disease expression of primary Sjögren syndrome. Analysis of 12,753 patients from the Sjögren Big Data Consortium. *Clin. Exp. Rheumatol.* **39**, 166–174 (2021).
18. Ramos-Casals, M. et al. Systemic involvement in primary Sjögren's syndrome evaluated by the EULAR-SS disease activity index: analysis of 921 Spanish patients (GEAS-SS Registry). *Rheumatology* **53**, 321–331 (2014).
19. Baldini, C. et al. Primary Sjögren's syndrome as a multi-organ disease: impact of the serological profile on the clinical presentation of the disease in a large cohort of Italian patients. *Rheumatology* **53**, 839–844 (2014).
20. Maślińska, M., Mańczak, M. & Kwiatkowska, B. Usefulness of rheumatoid factor as an immunological and prognostic marker in PSS patients. *Clin. Rheumatol.* **38**, 1301–1307 (2019).
21. Scofield, R. H., Fayyaz, A., Kurien, B. T. & Koelsch, K. A. Prognostic value of Sjögren's syndrome autoantibodies. *J. Lab Precis. Med.* **3**, 10–21037 (2018).
22. Nocturne, G. et al. Rheumatoid factor and disease activity are independent predictors of lymphoma in primary Sjögren's syndrome. *Arthritis Rheumatol.* **68**, 977–985 (2016).
23. Tzioufas, A. G., Boumba, D. S., Skopouli, F. N. & Moutsopoulos, H. M. Mixed monoclonal cryoglobulinemia and monoclonal rheumatoid factor cross-reactive idiotypes as predictive factors for the development of lymphoma in primary Sjögren's syndrome. *Arthritis Rheum.* **39**, 767–772 (1996).
24. Qin, B. et al. Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Ann. Rheum. Dis.* **74**, 1983–1989 (2015).
25. Ji, J., Sundquist, J. & Sundquist, K. Gender-specific incidence of autoimmune diseases from national registers. *J. Autoimmun.* **69**, 102–106 (2016).
26. Munguía-Realpozo, P. et al. Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: a systematic review. *Autoimmun. Rev.* **22**, 103294 (2023).
27. Danieli, M. G. et al. Machine learning application in autoimmune diseases: state of art and future prospectives. *Autoimmun. Rev.* **23**, 103496 (2024).
28. Dros, J. T. 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* **23**, 199 (2022).
29. Álvarez Troncoso, J. et al. Classification of salivary gland biopsies in Sjögren's syndrome by a convolutional neural network using an auto-machine learning platform. *BMC Rheumatol.* **8**, 60 (2024).
30. Marlin, M. C. et al. AB0802 AI-Enabled tissue classifier for Sjögren's disease salivary gland identifies key histological features important for understanding disease manifestation. *Ann. Rheum. Dis.* **83**, 1695 (2024).
31. Wu, R. et al. A graph-learning based model for automatic diagnosis of Sjögren's syndrome on digital pathological images: a multicentre cohort study. *J. Transl. Med.* **22**, 748 (2024).
32. Martorell-Marugán, J., Chierici, M., Jurman, G., Alarcón-Riquelme, M. E. & Carmona-Sáez, P. Differential diagnosis of systemic lupus erythematosus and Sjögren's syndrome using machine learning and multi-omics data. *Comput. Biol. Med.* **152**, 106373 (2023).
33. Pezoulas, V. C. et al. Addressing the clinical unmet needs in primary Sjögren's syndrome through the sharing, harmonization and federated analysis of 21 European cohorts. *Comput. Struct. Biotechnol. J.* **20**, 471–484 (2022).
34. Box, G. E. P. & Cox, D. R. An Analysis of Transformations. *J. R. Stat. Soc.: Ser. B (Methodol.)* **26**, 211–243 (1964).

35. Haibo, H., Yang, B., Garcia, E. A. & Shutao, L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; 1–8 June 2008, 1322–1328 (2008).
36. Pavlyshenko, B. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*; 21–25 Aug 2018, 255–258 (2018).

Acknowledgements

We acknowledge the support of Yidu Cloud Technology Company and the Hospital Information Management System platform for routine clinical data maintenance and export. This work was supported by the National Natural Science Foundation of China (82572066, 81901649), the Natural Science Foundation of Jiangsu Province (BK20251723, BK20250237), and Jiangsu Health International Exchange Program.

Author contributions

S.L. and G.W. designed the study and wrote the main manuscript. M.P. collected the clinical data and loaded the data into the database. Q.S. and C.G. analysed and interpreted the data. X.L. and C.T. provided external validation data sets. X.Y. and L.S. checked all results and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-02110-2>.

Correspondence and requests for materials should be addressed to Xinran Yuan.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025