Clinical Medicine Research
2025, Vol. 14, No. 3, pp. 90-97
https://doi.org/10.11648/j.cmr.20251403.17

Research Article

Development of a Machine Learning-based Clinical Prediction Model for Pulmonary Arterial Hypertension in Systemic Sclerosis Patients

Zhengbo Yan, Xu Cai, Xinmin Huang, Xinpeng Chen, Yiwei Hong, Wenru Chen*, Jianwei Xiao*

Department of Rheumatology, Shenzhen Futian Hospital for Rheumatic Diseases, Shenzhen, China

Abstract

Objective: To develop a machine learning-based clinical prediction model for assessing the risk of pulmonary arterial hypertension (PAH) in patients with systemic sclerosis (SSc), with the aim of enabling early intervention and improving patient prognosis and quality of life. Methods: A retrospective study was conducted, including 65 SSc patients and 20 SSc patients with PAH diagnosed between January 2018 and June 2023. A total of 333 clinical and laboratory parameters were collected as potential predictors. Feature selection was performed using the Boruta algorithm and LASSO regression, with overlapping variables used to construct a multivariable logistic regression model, which was visualized using a nomogram. Model performance was evaluated using the C-index, ROC curve, calibration curve, and decision curve analysis. Internal validation was conducted using the K-nearest neighbors (KNN) algorithm, with ROC AUC, precision-recall curve (PR curve), and confusion matrix as evaluation metrics. Results: Five key predictive factors—age, TNF-α, interstitial lung disease (ILD), impaired pulmonary function, and chest tightness—were identified. The model demonstrated excellent performance with a C-index of 0.94 and an AUC of 0.943. Calibration curves showed good consistency, and decision curve analysis indicated maximal net benefit within a threshold probability range of 0.1 to 0.75. As part of internal validation, KNN validation yielded an AUC of 0.9362 and an F1-score of 0.79, with stable performance observed across 5-fold cross-validation. Conclusion: A highly effective clinical prediction model was successfully developed, capable of identifying SSc patients at risk of developing PAH, particularly suitable for early screening in subclinical cases. The model offers significant clinical utility for targeted interventions, though further validation with larger cohorts and integration of novel biomarkers is warranted.

Keywords

Systemic Sclerosis, Pulmonary Arterial Hypertension, Machine Learning, Clinical Prediction Model, Boruta, Lasso

1. Introduction

Systemic sclerosis (SSc) is a rare systemic autoimmune disease characterized by localized or diffuse skin thickening or fibrosis and multiorgan involvement. The disease is associated with significant disability [1], and its pathogenesis involves microvascular damage, immune dysregulation with autoantibody production, and widespread fibrosis. Epidemi-

*Corresponding author: 108812365@qq.com (Jianwei Xiao)

Received: 20 May 2025; **Accepted:** 4 June 2025; **Published:** 23 June 2025

ological studies report a prevalence of 0.6–1.22 per million, categorizing it as a rare disease [2]. SSc affects individuals worldwide, with onset typically between ages 30 and 50, being more common in women (female-to-male ratio 3–14:1) [3]. The disease exhibits clinical heterogeneity in manifestations, autoantibody profiles, organ involvement, treatment response, and outcomes, with overall survival rates of 94.2–99% at 1 year, 94.8% at 3 years, 80.0–87.6% at 5 years, and 65.7–74.2% at 10 years [4].

Among complications, PAH is particularly severe, contributing to increased pulmonary vascular resistance (PVR), right ventricular afterload, and ultimately right heart failure and multi-organ dysfunction [5]. Early identification of PAH in SSc is crucial to prevent poor outcomes.

This study aimed to develop a simple and effective clinical prediction tool to estimate the risk of PAH in SSc patients, thereby facilitating early intervention and improving prognosis.

2. Materials and Methods

2.1. Patient Cohort

This retrospective study was approved by the institutional ethics committee. Sixty-five patients with SSc and twenty with SSc-PAH hospitalized from January 2018 to June 2023 were included. Inclusion criteria: SSc diagnosis based on 2013 ACR criteria [6]; PAH diagnosis based on 2022 ESC guidelines [7]. Exclusion criteria: corticosteroid use within the past 6 months, or coexisting malignancies.

Demographic, clinical, and laboratory data—including sex, age, disease duration, Raynaud's phenomenon, skin thickening, hyperpigmentation, xerostomia, dysphagia, TNF-α, gastrointestinal symptoms, ILD, pulmonary dysfunction, chest tightness, muscle enzymes, hemoglobin, digital sclerosis, joint symptoms, ESR, proteinuria, renal and liver function, autoantibodies (ACA, Scl-70, ANA, SSA, anti-JO1, anti-RNP), IgG, complement C3—were collected, totaling 333 variables.

2.2. Feature Selection

Feature selection was conducted using both the Boruta algorithm (with doTrace=2 and maxRuns=60) and LASSO regression via 10-fold cross-validation in R (glmnet package). The intersection of both methods was used to identify the optimal predictors.

2.3. Model Construction

A multivariable logistic regression model was developed using selected features, and a nomogram was constructed. Model discrimination was evaluated via the C-index and ROC analysis. Calibration was assessed using calibration plots. Clinical utility was evaluated through decision curve analysis.

2.4. Machine Learning

Machine Learning-Based Model Validation KNN was used to validate the model. A random grid search with 20 iterations was applied to tune hyperparameters. ROC and PR curves, confusion matrix, F1-score, and 5-fold cross-validation were used for performance evaluation.

3. Results

3.1. Feature Selection and Model Construction

Boruta identified five variables: age, TNF-α, ILD, pulmonary dysfunction, and chest tightness. LASSO identified nine variables, and the intersection yielded the same five predictors (Figure 1). These were incorporated into the logistic regression model, and a nomogram was generated (Figure 2).

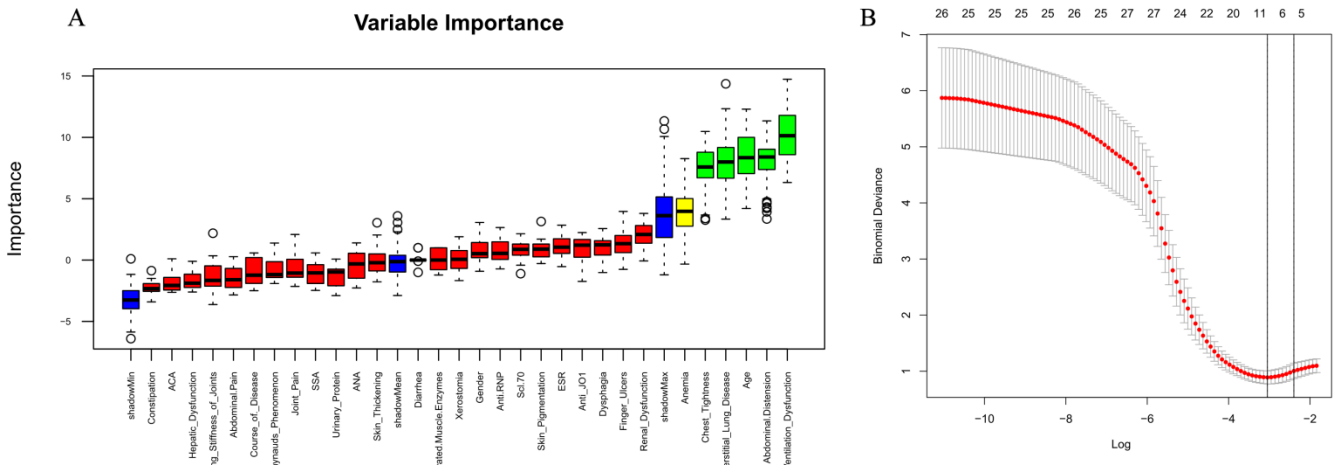

Figure 1. *Feature selection (A. Boruta B. Lasso).*

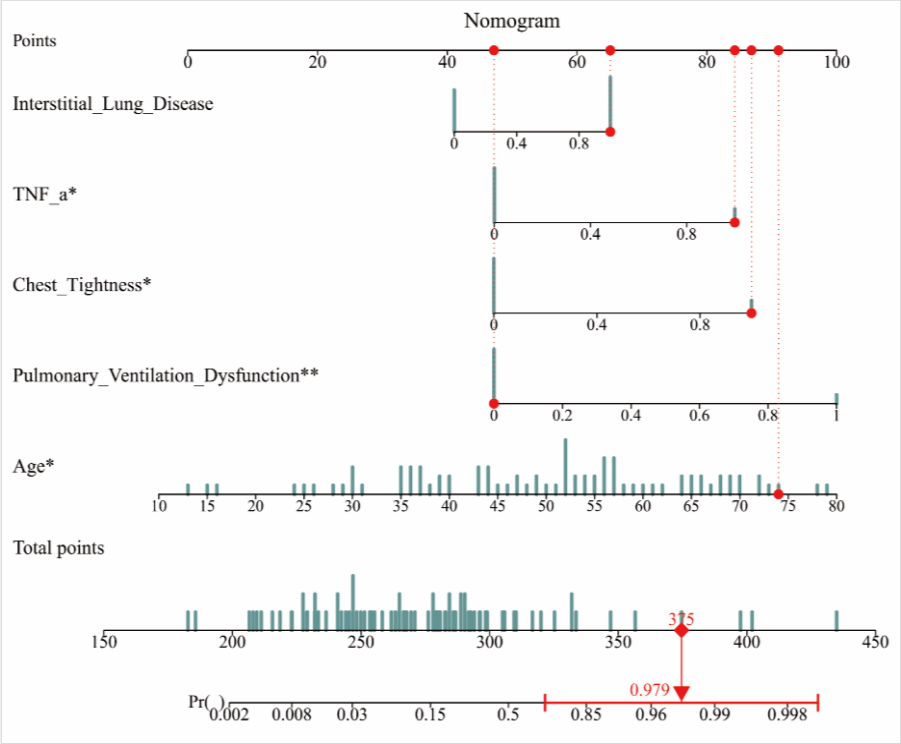

Figure 2. *Clinical prediction model nomogram.*

3.2. Clinical Model Performance

The model achieved a C-index of 0.94 and an AUC of 0.943 (Figure 3A). The calibration plot showed high agreement between predicted and observed outcomes (Figure 3B). Decision curve analysis demonstrated that the model yielded maximal net benefit within a threshold probability range of 0.1–0.75 (Figure 4).

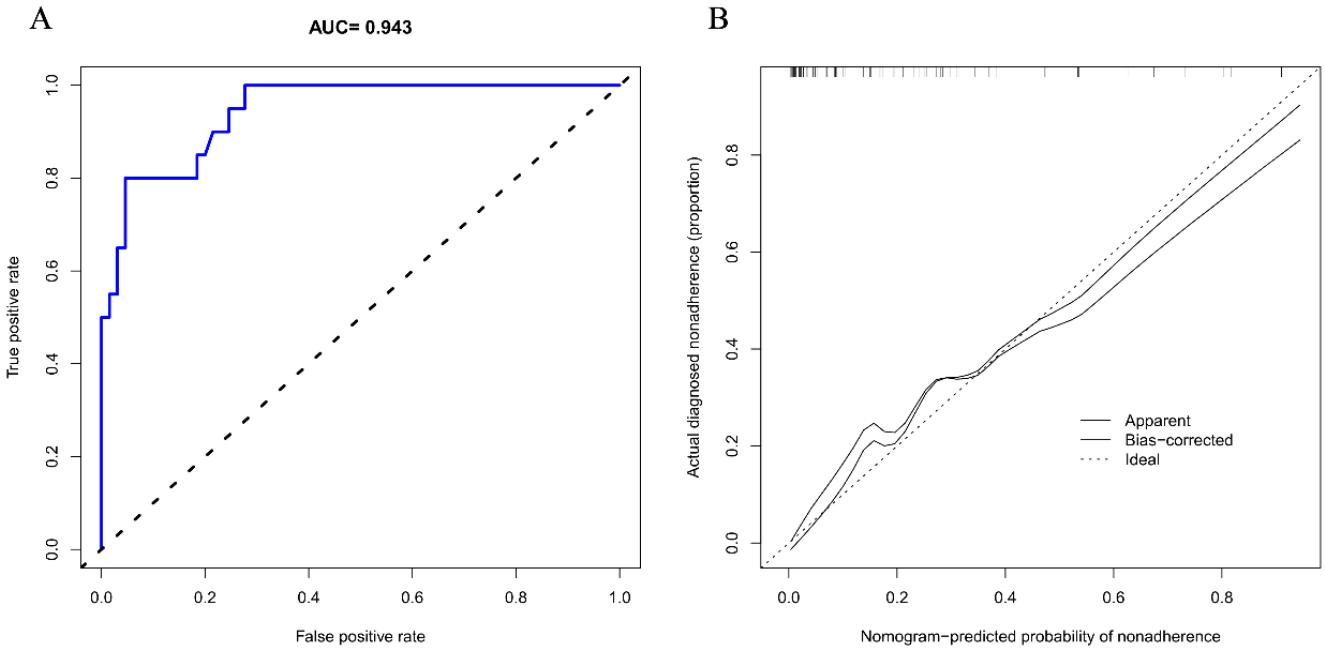

Figure 3. *ROC curve and clinical decision curve of the prediction model (A: The ROC curve shows an area under the curve (AUC) of 0.943 B: The diagonal dashed line indicates the ideal model, the solid line shows the nomogram's actual performance, and the closer the solid line is to the diagonal dashed line, the stronger the model's predictive ability.).*

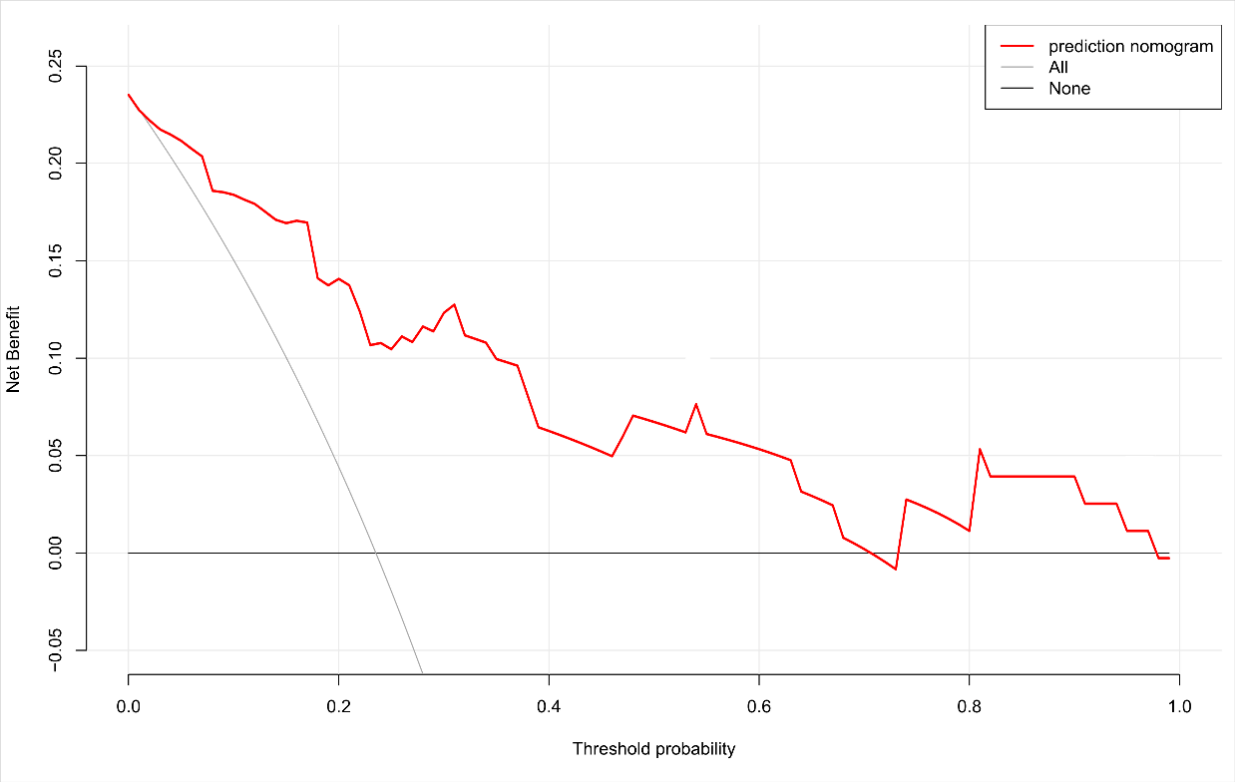

Figure 4. *Clinical decision curve.*

3.3. Machine Learning Validation Results

KNN validation showed an AUC of 0.9362 (Figure 5), PR curve area of 0.8272 (Figure 6), TPR=0.95, TNR=0.86, accuracy=0.88, and F1-score=0.79 (Figure 7). Five-fold cross-validation confirmed the model's robustness (ROC > 0.7; Figure 8).

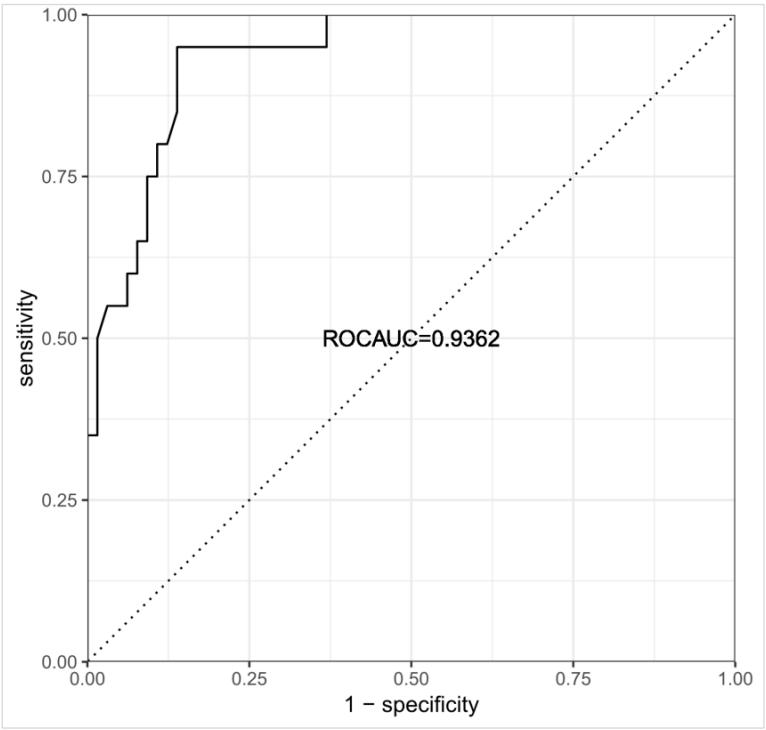

Figure 5. *ROC curve.*

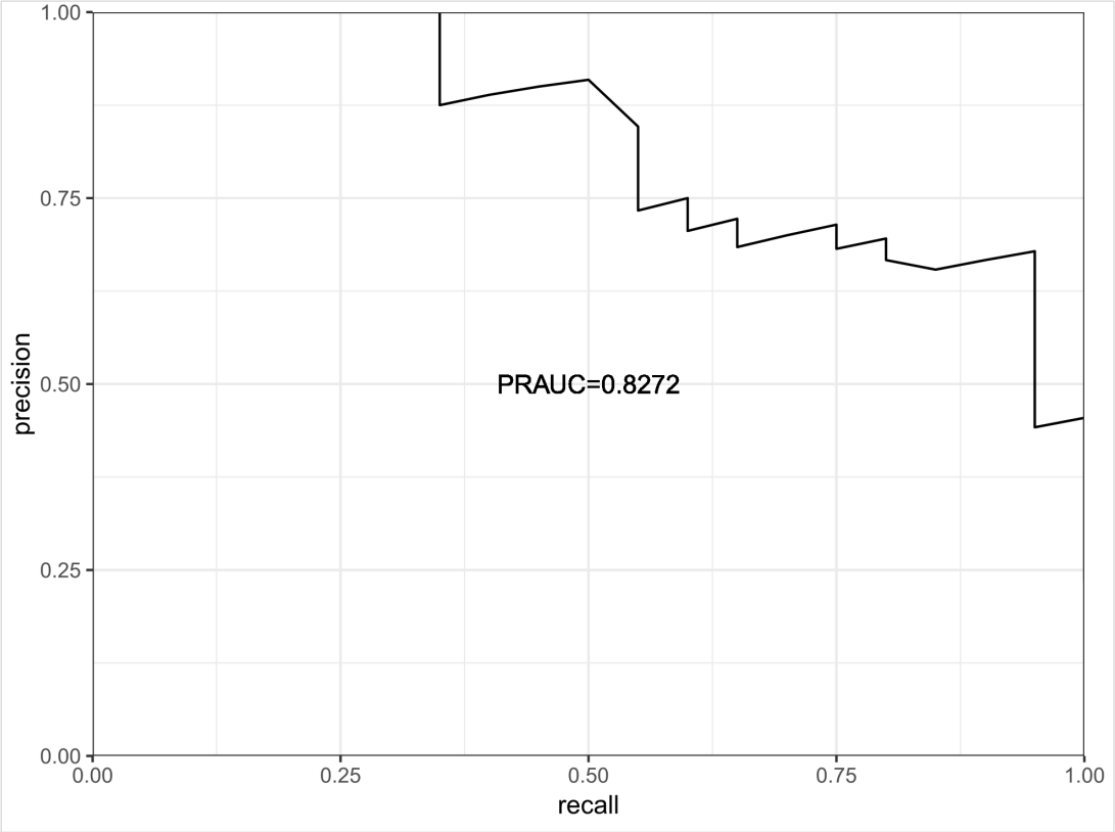

Figure 6. *PR curve.*

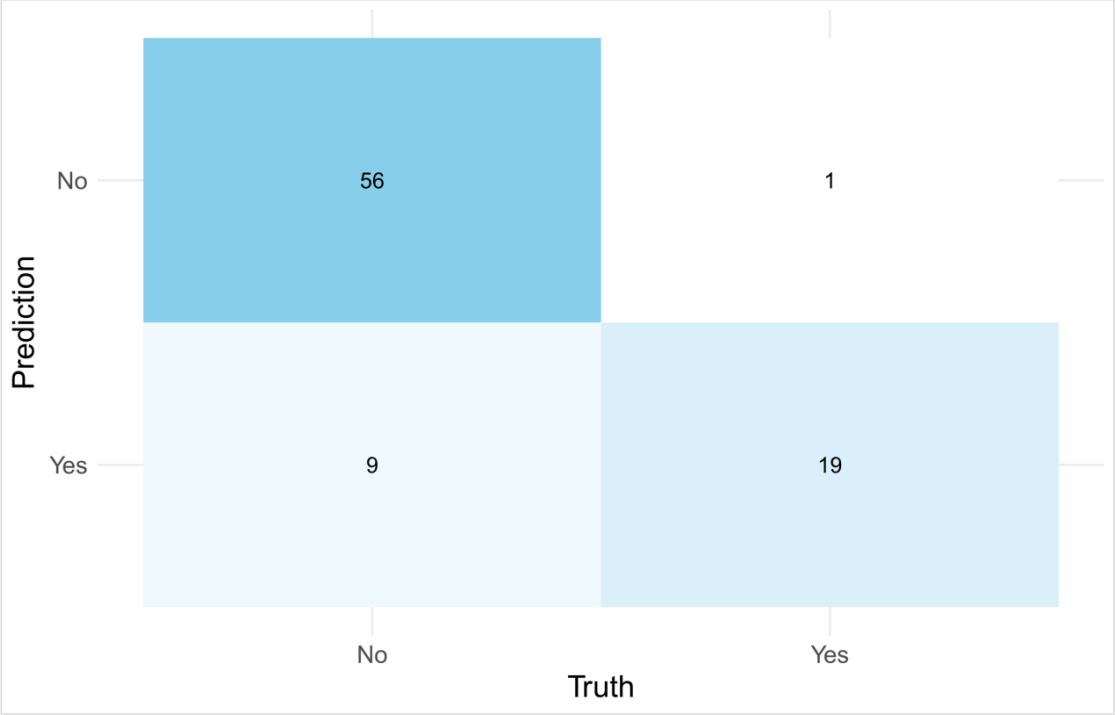

Figure 7. *Confusion matrix.*

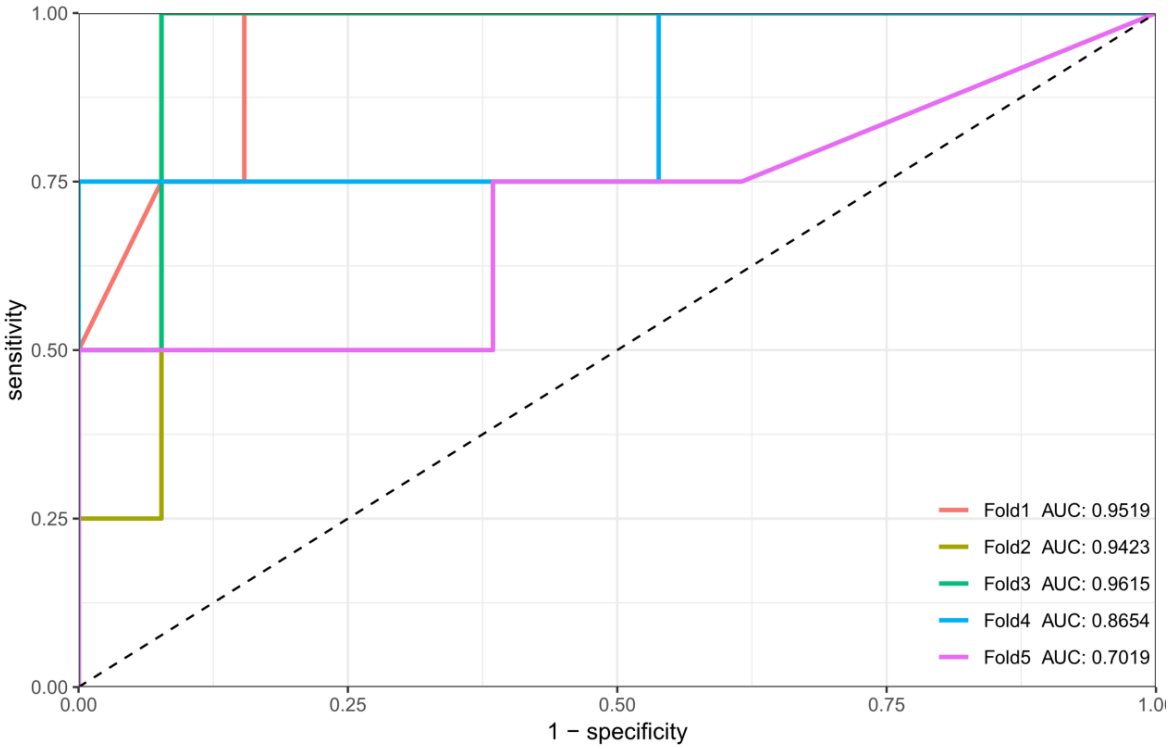

Figure 8. *Five - fold cross – validation.*

4. Discussion

SSc is characterized by multiorgan fibrosis and autoimmune dysfunction, with PAH representing a life-threatening complication. ILD and impaired pulmonary function were prominent in the risk model. Studies suggest ILD progression increases PVR due to alveolar-capillary barrier disruption and fibrotic vascular remodeling [8]. Impaired ventilation (e.g., abnormal FEV1/FVC) reflects small airway disease, hypoxia-induced vasoconstriction, and vascular remodeling [9, 10].

Advanced age (≥60) emerged as a key risk factor, consistent with declining endothelial function and immune dysregulation in older adults [11, 12]. Elevated antifibrotic antibodies and proinflammatory cytokines in elderly patients enhance vascular smooth muscle proliferation and contribute to PAH pathogenesis [13, 14].

Chest tightness, often underestimated by patients, showed strong correlation with echocardiographic markers of right ventricular overload. Its presence may reflect early cardiopulmonary dysfunction, supporting its use as an early clinical indicator of PAH [15].

TNF-α positivity, a marker of inflammatory activation, has been implicated in endothelial apoptosis and vasoconstrictor overexpression via NF-κB signaling. Clinical data indicate that TNF-α-positive patients have a 4.1-fold higher risk of PAH, possibly mediated by upregulated endothelin-1 (ET-1), a potent vasoconstrictor promoting vascular remodeling [16–18]. This highlights the potential for anti-inflammatory therapies targeting TNF-α in PAH prevention [19].

This model, comprising five predictive variables (age, ILD, TNF-α, pulmonary function, chest tightness), demonstrated excellent classification performance. Machine learning-based validation (KNN and LightGBM with SHAP analysis) confirmed these variables' contributions, aligning with nomogram results. Importantly, this model is capable of identifying subclinical PAH patients, offering a crucial window for early intervention [20].

In comparison to existing clinical models and screening protocols, our model offers several advantages [21, 22]. Firstly, it incorporates a comprehensive set of predictors, including clinical features, laboratory parameters, and inflammatory markers, providing a more holistic assessment of PAH risk. Secondly, the use of machine learning algorithms allows for the identification of complex interactions between predictors, potentially enhancing predictive accuracy. Finally, our model has been validated using multiple methods, including both traditional statistical approaches and machine learning-based techniques, ensuring its robustness and reliability.

However, it is important to acknowledge the limitations of our study. The small sample size and the exclusion of emerging biomarkers such as miRNA profiles and genetic susceptibility indicators may limit the generalizability of our findings. Additionally, external validation with a larger dataset has not been conducted, which is crucial for confirming the model's applicability in diverse clinical settings.

To address these limitations, we plan to utilize public databases to conduct preliminary analyses on the impact of

missing biomarkers, such as miRNAs, on the model's performance. This will help us understand how the integration of these biomarkers could improve the predictive accuracy and clinical applicability of the model.

In clinical practice, this model holds significant translational value. Physicians can use it to more accurately assess patients' risk of developing the disease and tailor personalized screening and treatment plans based on individual patient conditions. For high-risk patients, further examination and monitoring can be conducted in a timely manner, along with proactive interventions, such as early administration of endothelin receptor antagonists (e.g., ambrisentan), to slow disease progression. For low-risk patients, unnecessary tests and treatments can be minimized, thereby avoiding wastage of medical resources while also alleviating the financial and psychological burdens on patients [23]. Therefore, the model developed in this study provides clinicians with a powerful tool to achieve precision medicine, improve patients' quality of life and survival rates, and advance the clinical management of SSc with PAH to a new level. However, this study has some limitations, primarily the small sample size and the exclusion of emerging biomarkers such as miRNA profiles and genetic susceptibility indicators. Additionally, external validation with a larger dataset has not been conducted. Future research will continue to collect data and integrate multidimensional data using machine learning algorithms, including these emerging biomarkers and genetic susceptibility indicators, to further optimize the model and enhance its predictive performance.

Future research should focus on expanding the sample size, incorporating emerging biomarkers and genetic susceptibility indicators, and conducting external validation with diverse patient populations. Additionally, the development of user-friendly tools or applications based on this model could facilitate its integration into routine clinical practice.

5. Conclusion

In conclusion, we developed a machine learning-based clinical prediction model for PAH in SSc patients, which showed excellent performance and significant clinical utility for early intervention. However, limitations like the small sample size and the need for external validation highlight the necessity for future research to optimize the model.

Abbreviations

LASSO	Least Absolute Shrinkage and Selection Operator
TNF-α	Tumor Necrosis Factor – Alpha
ACA	Anticentromere Antibody
Scl-70	Antiscleroderma 70 Antibody
ANA	Antinuclear Antibody
SSA	Anti-Ssa Antibody
anti-JO1	Antibodies to Histidyl-tRNA Synthetase
anti-RNP	Antibodies to Ribonucleoprotein

Author Contributions

Zhengbo Yan: Data curation, Writing – original draft
Xu Cai: Writing – original draft
Xinmin Huang: Writing – original draft
Xinpeng Chen: Data curation, Formal Analysis
Yiwei Hong: Data curation, Formal Analysis
Wenru Chen: Conceptualization
Jianwei Xiao: Conceptualization

Funding

The present study was funded by the Shenzhen Futian District Health and Public Welfare Research Project (grant no. FTWS2022021, FTWS2022043, FTWS2023027, FTWS2023046) and the Shenzhen Health Economics Association Research Project (Grant Nos. 202597, 2025115).

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Denton CP, Khanna D. Systemic sclerosis [J]. Lancet, 2017, 390(10103): 1685-1699. https://doi.org/10.1016/S0140-6736(17)30933-9

[2] Barnes J, Mayes MD. Epidemiology of systemic sclerosis: incidence, prevalence, survival, risk factors, malignancy, and environmental triggers [J]. Curr Opin Rheumatol, 2012, 24(2): 165-170. https://doi.org/10.1097/BOR.0b013e3283502506

[3] Elhai M, Meune C, Avouac J, et al. Trends in mortality in systemic sclerosis: A meta-analysis of 7,000 patients over 40 years [J]. Arthritis Rheumatol, 2018, 70(10): 1653-1661. https://doi.org/10.1002/art.40535

[4] Rubio-Rivas M, et al. Survival and prognosis factors in systemic sclerosis: A systematic review and meta-analysis [J]. Semin Arthritis Rheum, 2021, 51(6): 1253-1264. https://doi.org/10.1016/j.semarthrit.2021.01.007

[5] Benza RL, Kanwar MK, Raina A, et al. Risk stratification and prognostic factors in pulmonary arterial hypertension: Insights from the REVEAL registry 2.0 [J]. Chest, 2023, 164(2): 351-362. https://doi.org/10.1016/j.chest.2023.04.007

[6] van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative [J]. Ann Rheum Dis, 2013, 72(11): 1747-1755. https://doi.org/10.1136/annrheumdis-2013-204424

[7] Volkmann ER, Andréasson K, Smith V. Systemic sclerosis [J]. Lancet, 2023, 401(10373): 304-318. https://doi.org/10.1016/S0140-6736(22)01692-0

[8] Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis: Pathogenesis and mechanisms of vascular remodeling [J]. Lancet Respir Med, 2014, 2(10): 815-826. https://doi.org/10.1016/j.lrm.2014.06.006

[9] Smith JJ, Donovan GM, Kistemaker LE, et al. Small airway dysfunction in early chronic obstructive pulmonary disease: Mechanisms and clinical implications [J]. Eur Respir J, 2023, 61(2): 2102123. https://doi.org/10.1183/13993003.02102-2022

[10] Yan T, Ma Q, Li X, et al. Establishment of a prediction model of pulmonary artery hypertension in patients with hyperthyroidism [J]. Ann Noninvasive Electrocardiogr, 2024, 29(5): e 13133. https://doi.org/10.1111/anec.13133

[11] Franceschi C, Garagnani P, Vitale G, et al. Inflamm-aging and anti-inflammaging: The dual role of the immune system in aging [J]. Nat Rev Immunol, 2023, 23(10): 633-650. https://doi.org/10.1038/s41577-023-00934-5

[12] Jiang Y, Wang X, Xia L, et al. Aging-induced T cell senescence promotes endothelial dysfunction via the CXCL16/CXCR6 axis [J]. Aging Cell, 2024, 23(3): e 14087. https://doi.org/10.1111/acel.14087

[13] Zhang Y, Wen J, Zeng M, et al. A Nomogram Prediction Model for Persistent Pulmonary Hypertension of the Newborn in Neonates Hospitalized for the First Time After Birth [J]. Pediatr Emerg Care, 2024, https://doi.org/10.1097/PEC.0000000000004027

[14] Dan W, Shuwei H, Jingke C, et al. A comprehensive study on machine learning models combining with oversampling for bronchopulmonary dysplasia-associated pulmonary hypertension in very preterm infants [J]. Respir Res, 2024, 25(1): 199. https://doi.org/10.1186/s12931-024-1285-9

[15] Hoeper MM, Pausch C, Grünig E, et al. Chest tightness as a warning sign of right ventricular decompensation in pulmonary arterial hypertension [J]. Eur Respir J, 2022, 59(1): 2101349. https://doi.org/10.1183/13993003.01349-2021

[16] Humbert M, Montani D, Savale L, et al. Proinflammatory cytokine profiling in systemic sclerosis-associated pulmonary arterial hypertension: TNF-α as a key biomarker [J]. Eur Respir J, 2023, 61(4): 2201345. https://doi.org/10.1183/13993003.01345-2022

[17] Megan G, Cedric M, K B C, et al. Abstract 15889: An Artificial Intelligence-Derived Pediatric Pulmonary Hypertension Risk Prediction Model From the Pediatric Pulmonary Hypertension Network (PPHNet) Registry [J]. Circulation, 2023, 148(Suppl_1): A 15889-A 15889. https://doi.org/10.1161/CIRCULATIONAHA.123.015889

[18] Mathai SC, Saketkoo LA, Hsu VM, et al. TNF-α-driven vascular remodeling in systemic sclerosis-associated pulmonary hypertension [J]. Eur Respir J, 2022, 59(2): 2101497. PMID: 3494782.

[19] Mengmeng D, Chunyi Z, Chaoying L, et al. Clinical characteristics and prognosis in systemic lupus erythematosus-associated pulmonary arterial hypertension based on consensus clustering and risk prediction model [J]. Arthritis Res Ther, 2023, 25(1): 155. https://doi.org/10.1186/s13075-023-03119-0

[20] Mukul S, Shashank V, Amod A. 486: PEDIATRIC PULMONARY HYPERTENSION READMISSIONS PREDICTION MODEL USING ARTIFICIAL INTELLIGENCE [J]. Crit Care Med, 2022, 50(1 Suppl 1): 234-234. https://doi.org/10.1097/01.CCM.0000892017.82321.X7

[21] Bauer, Y., de Bernard, S., Hickey, P., et al. Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort. The European respiratory journal, 57(6), 2002591. https://doi.org/10.1183/13993003.02591-2020

[22] Coirier V, Lescoat A, Fournet M, et al. Dépistage de l'hypertension artérielle pulmonaire au cours de la sclérodermie systémique: comparaison de l'algorithme DETECT à une discussion pluridisciplinaire en centre de compétence [Screening for pulmonary arterial hypertension in patients with systemic sclerosis: Comparison of DETECT algorithm to decisions of a multidisciplinary team, in a competence centre]. *Rev Med Interne*. 2017; 38(8): 502-507. https://doi.org/10.1016/j.revmed.2017.04.005

[23] Robin C, Kris B, Jennifer S, et al. Reply: External validation of the OPALS prediction model for in-hospital mortality in patients with acute decompensated pulmonary hypertension [J]. ERJ Open Res, 2022, 8(1): 03292. https://doi.org/10.1183/23120541.00329-2022