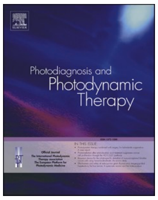
Contents lists available at ScienceDirect

Photodiagnosis and Photodynamic Therapy

journal homepage: www.elsevier.com/locate/pdpdt



Raman spectroscopy combined with machine learning algorithms for rapid detection Primary Sjögren's syndrome associated with interstitial lung disease

Xue Wu [a,b], Chen Chen [c], Xiaomei Chen [a,b], Cainan Luo [a,b], Xiaoyi Lv [d], Yamei Shi [a,b], Jie Yang [c], Xinyan Meng [a,b], Cheng Chen [d], Jinmei Su [e,*], Lijun Wu [a,b,**]

[a] *Department of Rheumatology and Immunology, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, China*
[b] *Xinjiang Clinical Research Center for Rheumatoid Arthritis, Urumqi 830001, China*
[c] *College of Information Science and Engineering, Xinjiang University, Urumqi 830001, China*
[d] *College of Software, Xinjiang University, Urumqi 830046,China*
[e] *Department of Rheumatology and Clinical Immunology, Peking Union Medical College Hospital, Beijing, China*

ARTICLE INFO

Keywords:
Primary Sjögren's syndrome
Raman spectroscopy
Interstitial lung disease
Machine learning

ABSTRACT

Background: Interstitial lung disease (ILD) is a major complication of Primary Sjögren's syndrome (pSS) patients. It is one of the main factors leading to death. The aim of this study is to evaluate the value of serum Raman spectroscopy combined with machine learning algorithms in the discriminatory diagnosis of patients with Primary Sjögren's syndrome associated with interstitial lung disease (pSS-ILD).
Methods: Raman spectroscopy was performed on the serum of 30 patients with pSS, 28 patients with pSS-ILD and 30 healthy controls (HC). First, the data were pre-processed using baseline correction, smoothing, outlier removal and normalization operations. Then principal component analysis (PCA) is used to reduce the dimension of data. Finally, support vector machine(SVM), k nearest neighbor (KNN) and random forest (RF) models are established for classification.
Results: In this study, SVM, KNN and RF were used as classification models, where SVM chooses polynomial kernel function (poly). The average accuracy, sensitivity, and precision of the three models were obtained after dimensionality reduction. The Accuracy of SVM (poly) was 5.71% higher than KNN and 6.67% higher than RF; Sensitivity was 5.79% higher than KNN and 8.56% higher than RF; Precision was 6.19% higher than KNN and 7.45% higher than RF. It can be seen that the SVM (poly) had better discriminative effect. In summary, SVM (poly) had a fine classification effect, and the average accuracy, sensitivity and precision of this model reached 89.52%, 91.27% and 89.52%, respectively, with an AUC value of 0.921.
Conclusions: This study demonstrates that serum RS combined with machine learning algorithms is a valuable tool for diagnosing patients with pSS-ILD. It has promising applications.

1. Introduction

Primary Sjögren's syndrome (pSS) is an autoimmune disease characterized by chronic lymphocytic infiltration. Its incidence is related to multiple factors such as genetics and environment [1]. It affects 0.06% of the general population. The pSS mainly affects the exocrine glands and can cause symptoms such as dry eyes and dry mouth. Among them, 30–40% of patients will have in extra-glandular manifestations, such as arthritis, interstitial nephritis and neurological involvement [2,3].

Primary Sjögren's syndrome associated with Interstitial lung disease (pSS-ILD) is a major complication of pSS patients. Also, it is one of the main factors leading to the death of patients. The prevalence of pSS-ILD is about 13%. The typical pathological manifestations are diffuse parenchymal lung, pulmonary alveolitis and interstitial fibrosis. The main manifestations are weakness, dyspnea, cough, etc. [4–6]. The diagnosis of pSS-ILD currently relies on high-resolution computed tomography (HRCT) and pulmonary function testing (PFTs) [7,8]. In addition, there have been attempts to apply MRI for the diagnosis of

* Corresponding author.
** Corresponding author at: Department of Rheumatology and Immunology, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, China.
E-mail addresses: annasu2014@163.com (J. Su), wwlj330@126.com (L. Wu).

https://doi.org/10.1016/j.pdpdt.2022.103057
Received 27 May 2022; Received in revised form 15 July 2022; Accepted 5 August 2022
Available online 6 August 2022
1572-1000/

pSS-ILD [9]. However, the above methods still have some limitations. Although HRCT is the primary tool for the diagnosis of pSS-ILD, it has the disadvantage of radiation exposure, which makes it unable to be widely used in repeated screening. The sensitivity of PFTs in the early stages of the disease is quite low, so it cannot be used as a screening method for the diagnosis of pSS-ILD. Surgical lung biopsy (SLB), known as the gold standard for diagnosis, is an invasive examination. At the same time, it may produce acute exacerbation of the early disease. So SLB can only be used for the diagnosis of atypical pSS-ILD. Therefore, it is a need to develop a rapid, accurate, noninvasive screening tool for pSS-ILD that not only optimizes current diagnostic methods but also helps to achieve early treatment of pSS-ILD and improve patient prognosis.

Raman spectroscopy (RS) is a scattering spectrum [10] that reflects the energy difference between incident photons and vibrating molecules and provides the molecular structure, tissue configuration and fingerprint information of substances. RS not only has the advantages of easy information extraction, nondestructive detection and high fingerprint resolution, but also has the advantages of rapidity, non-invasiveness, low interference and accuracy [11]. In recent decades, RS has been successfully applied in some medical fields, including the detection of diseases, such as allergic diseases, diabetes, thyroid cancer, etc. [12–14].

Machine learning can further improve the efficiency of disease diagnosis. The idea of machine learning describes the ability of a system to learn from problem-specific training data to automate the analytical model building process and solve related tasks [15]. As the industrial applications of machine learning are booming rapidly, the combination with Raman spectroscopy has achieved remarkable results in disease detection and diagnosis [16]. For example, Khan S et al. used Raman spectroscopy combined with support vector machine (SVM) to screen for hepatitis B virus [17]. Haitao Song et al. used a variety of machine learning models such as decision tree (DT) combined with Raman spectroscopy to identify papillary thyroid carcinoma and papillary microcarcinoma [18]. Cheng Chen et al. used Raman spectroscopy with back propagation neural network (BP), extreme learning machine (ELM) and other algorithms based on serum samples to diagnose patients with chronic renal failure [19]. However, there is no study on the identification of pSS and pSS-ILD using RS. Hence, it is great significance to develop a non-invasive, inexpensive, and stable pSS-ILD detection technique by RS.

Due to the high dimensionality of spectral data, too much feature data will cause information redundancy and cause the degradation of model accuracy, so dimensionality reduction is usually used to select representative features. PCA reduces training cost and improves model prediction accuracy by projecting high-dimensional data into low-dimensional space, calculating principal components and selecting appropriate principal component scores [20]. In addition, a good classification model is crucial for establishing disease identification. In this experiment, three models of SVM, KNN and RF were established based on serum RS combined with machine learning algorithms. However, the selection of the kernel function of the support vector machine affects the accuracy of the classification, so we choose the polynomial kernel function as the kernel function of the SVM. By comparing the results of the three classification models, the feasibility of serum RS combined with machine learning algorithms to diagnose pSS-ILD was verified. From the perspective of spectral analysis, it provides an interesting and effective diagnostic screening strategy for the diagnosis of pSS-ILD.

2. Materials and methods

2.1. Experimental materials

In this study, 28 pSS-ILD patients who visited the Department of Rheumatology and Immunology of the People's Hospital of Xinjiang Uygur Autonomous Region from January 2020 to December 2021 were included. Age and gender matched 30 pSS patients and 30 health control were selected during the same period.we collected fresh blood from pSS patients, pSS-ILD patients and HC. The onset of the disease is characterized by a higher incidence in women, therefore, the data we collected were not subject to gender imbalance. Table 1 contains information about the patients and healthy controls, such as their age and gender. All blood samples were drawn to obtain peripheral blood samples without any anticoagulant. Centrifugation was performed at 4 °C at a high speed of 4000r/min. After centrifugation for 10 min, the top clear night was obtained and stored in a refrigerator at −80 °C for subsequent experiments. The samples for this study were obtained from the People's Hospital of Xinjiang Uygur Autonomous Region and ethical approval was obtained (Table 2).

2.2. Raman spectroscopy acquisition

The Raman spectra of the serum samples were recorded using a confocal Raman microspectrometer (LabRAM HR Evolution Raman Spectrometer, Horiba Scientific, Ltd.). The laser wavelength was 532 nm, and the laser power was 20 mW (integration time 25 s, integration times three times). At room temperature of 22 °C, the laser beam was focused on the sample surface through a 10 × objective lens to measure the sample. The spectral measurement range was from 600 to 1800 cm^{-1}. Each serum sample was measured five times from different locations. Finally, 440 spectral data were obtained, including 150 pSS data, 140 pSS-ILD data and 150 HC data. To avoid errors in the spectral acquisition process, we used the mean of 5 measurements per sample as input for the subsequent study.

2.3. Data preprocessing

In the study, the interference of factors such as fluorescence background, sample concentration difference and laser power fluctuation existed in the collected serum Raman spectra, and the intensity of the spectral signal was relatively weak. Since the fluorescence background is the main interfering factor, the spectrum is complicated by the fluorescence background noise, which affects the analysis to a great extent [21]. In order to reduce the interference of these factors in the experiment and improve the reliability of the experimental results of spectral analysis, this experiment requires relevant preprocessing of the collected spectral data.

For the original spectra, we perform band selection, baseline correction, smoothing, followed by outlier removal and normalization preprocessing. First, the 600–1800cm^{-1} of the Raman spectrum is selected, which belongs to the fingerprint region of the Raman spectrum, which can improve the accuracy of diagnosis, so this band range is usually used for analysis [22]. Among them, adaptive iteratively reweighted penalized least squares (airPLS) has the advantages of being simple, flexible, semiautomatic, fast and effective [23–25], The Savitzky-Golay filter (SG) has the advantage that the reconstructed data can better preserve local features and ensure that the shape and width of the signal remain unchanged [26]. Therefore, the baseline correction adopts the airPLS method, and the smoothing filtering adopts the SG method. In this paper, t-SNE was used to eliminate outliers, which helped us to identify correlated patterns so that similar data can be closer together in low dimensions and dissimilar data were farther apart in low

Table 1
Brief information on pSS patients, pSS-ILD patients and healthy controls.

	pSS ($n = 30$)	pSS-ILD ($n = 28$)	HC ($n = 30$)	significant p-value
Age				
Mean	54.3(±11.9)	57.5(±10.1)	56.9(±10.2)	> 0.05
Gender				
Male	1	5	7	> 0.05
Female	29	23	23	

Table 2
Confusion matrix.

Actual Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

dimensions [27]. Then, we normalized the integral area under the curve of the preprocessed Raman spectrum and scaled the eigenvalues to reduce the data complexity and improved the model convergence speed [28].

In addition, since the acquired Raman spectra are high-dimensional data, their direct input introduces noise interference and increases the computational cost and time. Therefore, we used PCA to reduce the dimensionality of the spectral data. After reducing the dimensionality of the data, selected the appropriate number of features to input into the classification model. In this study, PCA features extracted the first 31-dimensional spectral data as the input of the subsequent classification model according to the 99% principal component contribution rate [29], and three classification algorithms, SVM, KNN, and RF, were used for modeling.

2. 4. Classification algorithm

SVM is a powerful multi-class pattern recognition technique whose core idea is to apply structural risk minimization to the field of classification. It can analyze small samples, nonlinear and high-dimensional data with good generalization, versatility and robustness [30–32]. SVM is now widely used for the classification and detection of breast cancer, liver syndrome, lumbar disk herniation and other diseases [32–34]. Based on the advantages of this classifier and its availability for multiple classifications, this study performed feature extraction by PCA on spectral data before modeling and classifying pSS, pSS-ILD and HC using SVM.

KNN is one of the most practical and best classification algorithms in data mining classification techniques. Unlike other classification algorithms, KNN does not require training; it is easy to use, fast to train, and highly accurate [35]. It can directly find the k samples closest to that sample and select the sample with the largest number of samples among the k samples for classification. Therefore, the method is applicable to multivariate classification. In addition KNN has been widely used for disease diagnosis detection [36,37], so we used KNN as the second model for multivariate classification.

RF is a classifier consisting of multiple decision trees whose output classes are determined by the plural of the classes output by the individual trees. RF has many advantages, including high accuracy, good performance for small-scale data, high generalization capability, and the ability to parallelize computation [38]. Therefore, it is widely used in biomedical and food research fields [39–41]. In addition, good classification results have been achieved by applying RF to multiple classifications [42]. In summary, we chosed RF as the third classification algorithm.

2.5. Model metrics

The performance of three classification models was evaluated by Sensitivity, Precision and Accuracy, as in Eqs. (1)–(3):

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

3. Results

3.1. Spectral analysis

Fig. 1 showed the average pre-treated Raman spectra of pSS, pSS-ILD and HC in the wave number 600 to 1800 cm^{-1} range. Some differences between the spectra of the two groups of patients and the control group can be clearly seen. In this study, we selected five main peaks corresponding to Raman shifts (662, 724, 1005, 1152, 1514 cm^{-1}) for analysis. According to the Raman spectra, it can be seen that the intensity of the Raman peaks of the patients was higher than that of the healthy people at 662 and 724 cm^{-1}. At 1005, 1152, and 1514 cm^{-1}, the Raman peak intensity was higher in healthy individuals than in patients. There was also a significant difference in the peak Raman intensity between pSS and pSS-ILD patients at 1152 and 1514 cm^{-1}. This provided an important theoretical basis for subsequent classification studies.

3.2. Sample division

To prevent the spectral anomalous data from having a large impact on the experiment, we used the t-SNE method to eliminate 5 pSS, 8 pSS-ILD and 3 HC anomalous samples, respectively. The corresponding remaining 25, 20 and 27 sample data were divided into training and test sets in the ratio of 7:3. There were 51 samples in the training set (18 pSS, 14 pSS-ILD and 19 HC) and 21 samples in the test set (7 pSS, 6 pSS-ILD and 8 HC). To avoid the data division with chance and increase the robustness of the experimental model, we randomly divided the data five times according to the above method, and the mean value of the five modeling results was taken as the final result in the experiment.

3.3. PCA feature extraction

If the normalized Raman spectral data are classified directly, it is not only computationally intensive but also time-consuming. Therefore it is necessary to perform feature extraction. In this study, we performed PCA feature extraction on the divided dataset. The three main components PC1, PC2 and PC3 with the highest contributions from the training and test sets were plotted and analyzed (Fig. 2). In the scatter plot, pSS and pSS-ILD patients have samples mixed together, which was difficult to distinguish. To improve the accuracy of subsequent classification, we need to use machine learning models (Fig. 3).

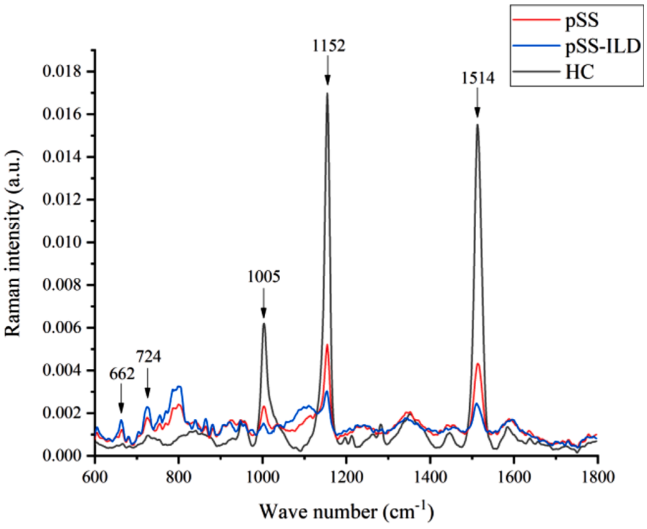


Fig. 1. Normalized mean Raman spectra of pSS, pSS-ILD and HC groups.

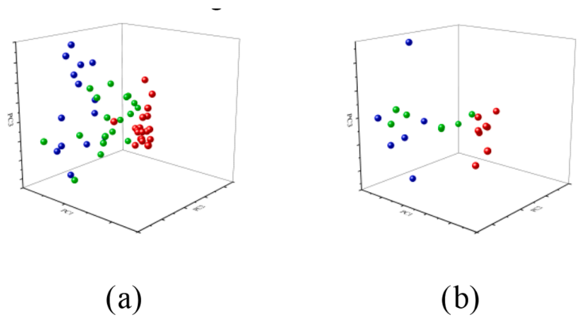


Fig. 2. (a) 3D scatterplot of the first three principal components of the training set (b) 3D scatterplot of the first three principal components of the test set Note: The green ball in the figure represents pSS; the blue ball represents pSS-ILD; the red ball represents HC.

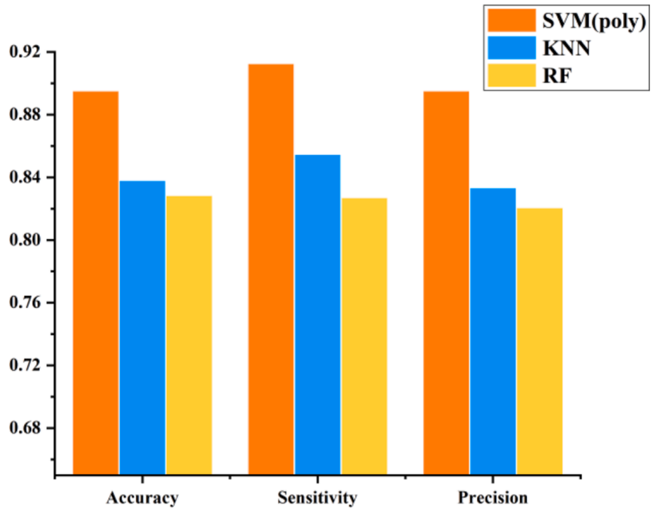


Fig. 3. Experiment results of three models.

3.4. Classification model results

In this study, SVM, KNN and RF were used as classification models, where the SVM kernel function was chosen as a polynomial kernel function (poly). The average accuracy, sensitivity, precision, and AUC results of the dimensionality-reduced data on the three models were shown in Table 3. The Accuracy of SVM(poly) was 5.71% higher than KNN and 6.67% higher than RF; Sensitivity was 5.79% higher than KNN and 8.56% higher than RF; Precision was 6.19% higher than KNN and 7.45% higher than RF. It can be seen that the SVM(poly) had better discriminative effect. In addition, we plotted the sample operating characteristic (ROC) curve, and the integral area under the ROC curve represented the AUC value. The larger the AUC value, the more reliable the model. The ROC curve of the model running average results was shown in Fig. 4, where the AUC of SVM(poly) was 0.042 higher than KNN and 0.05 higher than RF. Combining the results of each model evaluation index, we concluded that SVM(poly) was the best for patient discrimination.

Table 3
Experimental results of three machine learning models.

Model	Accuracy	Sensitivity	Precision	AUC
SVM(poly)	89.52%	91.27%	89.52%	0.921
KNN	83.81%	85.48%	83.33%	0.879
RF	82.85%	82.71%	82.07%	0.871

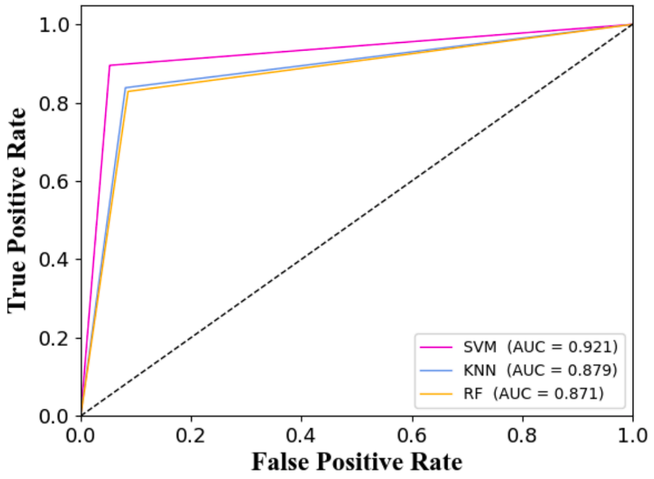


Fig. 4. ROC curves of the three models.

4. Discussion

Patients with lung involvement commonly present with ILD, which is also an important factor in poor prognosis and death in SS patients. However, there is still no "gold standard" for the diagnosis of pSS-ILD, and a multidisciplinary team (MDT) is still recommended for the diagnosis, evaluation, treatment and follow-up of patients, including rheumatologists, respiratory physicians and radiologists [43]. The monitoring of disease has a significant impact on early intervention and will strongly improve the prognosis of patients. Therefore, it is important to find simple and reproducible tools for the screening and diagnosis of CTD-ILD and the assessment of disease activity. Early identification and diagnosis of ILD is currently one of the major clinical challenges. Delayed diagnosis is a common problem in patients with ILD. A study of the United States showed that 55% of people received one misdiagnosis and 38% received two or more misdiagnoses before receiving a correct diagnosis [44]. The main reason for this is the insidious early clinical presentation of ILD and the lack of early screening tools. Although HRCT is the main tool for the diagnosis of ILD, radiation exposure and high cost prevent its use for routine screening of ILD, limiting its widespread repeat use as a screening test.PFTs are commonly used to assist in the diagnosis and detection of disease activity in ILD, however, PFTs are valuable for monitoring changes in ILD, but because of their low sensitivity in the early stages of the disease, testing Although SLB is the gold standard for diagnosis, it is an invasive test that most patients are unwilling to undergo and may also cause acute exacerbation of early disease, so SLB can only be used for the diagnosis of atypical ILDs. In addition to the lack of non-invasive, early diagnostic tools for ILD, there is also a lack of effective tools that can predict the response to treatment and prognosis of individual patients. Some progress has been made with blood biomarkers, but these are currently still in the research phase [45,46]. Therefore, if a rapid, accurate, noninvasive screening tool for ILD could be investigated not only to optimize current diagnostic methods, but also to help achieve early treatment of ILD and improve patient prognosis.

RS is a non-invasive detection technique, can reflect the structural characteristics and compositional information of biological macromolecules [47]. The peaks are mainly caused by biomolecules such as phospholipids, nucleic acids and proteins, and the intensity of the peaks can reflect the corresponding molecular concentrations [48]. Based on the spectrograms, We can see that the partial peaks in pSS and pSS-ILD patients are close to each other, indicating that the two sera have similar biomolecules. However, at different peaks, it can be seen that there are some differences between the pSS, pSS-ILD and HC groups,which may be due to differences in disease metabolism causing alterations in serum composition and thus in biomolecule concentrations. Table 4 lists the

Table 4
The major Raman bands of human serum.

Wave number (cm^{-1})	Assignment
662	C-S stretching mode of cystine (collagen type I)
724	Coenzyme A, acetylcoenzyme A
1005	Phenylalanine
1152	Proteins
1514	Carotenoid

material assignments of the Raman spectral peaks corresponding to the main bands [49–51]. Combined with Table 4 to observe Raman Peak at 662 and 724 cm^{-1}, the peaks of pSS and pSS-ILD patients at these bands are higher than that of normal people. The peak of Raman at 662 cm^{-1}corresponds to type I collagen, The peak of Raman at 724 cm^{-1}corresponds to coenzyme A and acetylcoenzyme A. This shows that type I collagen, coenzyme A and acetylcoenzyme A in patients with pSS and pSS-ILD patients are higher than normal people. The Raman peaks at 1005, 1152 and 1514cm^{-1} were higher in normal subjects than in pSS and pSS-ILD patients. The Raman peak at 1005cm^{-1}corresponds to phenylalanine, the Raman peak at 1152cm^{-1}corresponds to protein, and the Raman peak at 1514cm^{-1}corresponds to carotenoids, indicating that the metabolism of pSS and pSS-ILD patients was altered, resulting in lower levels of substances such as amino acids, proteins, and carotenoids in the patients' bodies than in normal individuals. In addition, we analyzed the correlation of the intensity of these peaks with the category, and their significance *p*-values were all less than 0.05, indicating a significant correlation between them. And the correlation coefficients are all greater than 0.8, we believe that the intensity of these peaks and the type of disease also show a strong correlation.

To increase the model robustness and determine the best classifier, we taked the mean value of each evaluation index after five runs of each model as the final evaluation criterion. In this study, the spectral data after preprocessing and PCA dimensionality reduction were taken, and the first 31 PCs were taken as input. By comparing the evaluation metrics of the three models, the accuracy of RF was relatively low, which may be due to the fact that although the RF algorithm is simple to implement and easy to model, the classification ability is reduced for datasets with high feature relevance. In contrast, KNN had high accuracy, was insensitive to outliers, and was suitable for the establishment of small sample classification models.Therefore, the model accuracy and other indicators perform relatively well in this experiment. Comparing the evaluation metrics of the three models, the comprehensive performance of SVM (poly) was better than the KNN model, because SVM can map to a high-dimensional space by using kernel functions and requires a relatively small number of samples.

To our knowledge, this is the first study to use Raman spectroscopy combined with machine learning algorithms for rapid detection pSS-ILD. This study shows that RS combined with machine learning methods has great potential research value in the early diagnosis and screening of pSS-ILD. However, due to the limited sample size, our study has some limitations. Therefore, we plan to collect more samples in the future to validate this exploratory study and further reveal the potential of Raman spectroscopy combined with powerful machine learning algorithms in the process of screening pSS-ILD. After further validation analysis, serum Raman spectroscopy combined with a powerful classifier can be prospectively extended to different stages of research.

5. Conclusion

This study used serum RS combined with machine learning algorithms to achieve a rapid and accurate diagnosis of pSS-ILD patients. In this study, the SVM(poly) performance was found to be the most stable, with its accuracy, sensitivity, precision and AUC values reaching 89.52%, 91.27%, 89.52% and 0.921, respectively. This study shows that RS combined with machine learning methods can help to understand the differences between pSS,pSS-ILD and HC in human body, and has great potential research value in the early diagnosis and screening of pSS-ILD, which provides a new idea for a non-invasive, efficient, rapid and inexpensive clinical medical diagnosis method for patients.

CRediT authorship contribution statement

Xue Wu: Writing – original draft, Data curation, Methodology, Conceptualization. **Chen Chen:** Software, Methodology, Data curation, Conceptualization. **Xiaomei Chen:** Data curation. **Cainan Luo:** Formal analysis. **Xiaoyi Lv:** Supervision. **Yamei Shi:** Writing – review & editing. **Jie Yang:** Software, Methodology, Data curation. **Xinyan Meng:** Methodology. **Cheng Chen:** Software, Methodology, Formal analysis, Data curation. **Jinmei Su:** Funding acquisition, Writing – review & editing, Supervision. **Lijun Wu:** Funding acquisition, Methodology, Supervision.

Declaration of Competing Interest

The authors declare that they have no conflicting financial interests.

Acknowledgments

This work was supported by the Youth Science Fund of Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01c144), the Distinguished Young Talents Project of Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01E11 and 2022D01E27), the Key Research and Development Project of Xinjiang Uygur Autonomous Region (2022B03002-1) and the project of scientific and technological assistance to Xinjiang (No.2020E0269).

References

[1] J.D. Paulo, C.J. Velásquez-Franco, M.C.O. Usuga, M.A. Velásquez, D.C. Montoya, J. H. Donado, Puntaje de tinción ocular en pacientes con diagnóstico de síndrome de Sjögren en una institución de salud en Medellín, Colombia, Rev. Colomb. Reumatol. 27 (2020) 15–21.

[2] B. Qin, J. Wang, Z. Yang, M. Yang, N. Ma, F. Huang, R. Zhong, Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis, Ann. Rheum. Dis. 74 (11) (2015) 1983–1989, https://doi.org/10.1136/annrheumdis-2014-205375. Nov.

[3] X. Mariette, L.A. Criswell, Primary Sjögren's syndrome, N. Engl. J. Med. 378 (10) (2018) 931–939, https://doi.org/10.1056/NEJMcp1702514. Mar 8.

[4] C. He, Z. Chen, S. Liu, H. Chen, F. Zhang, Prevalence and risk factors of interstitial lung disease in patients with primary Sjögren's syndrome: a systematic review and meta-analysis, Int. J. Rheum. Dis. 23 (8) (2020) 1009–1018, https://doi.org/10.1111/1756-185X.13881. Aug.

[5] T. Guo, Y. Long, Q. Shen, W. Guo, W. Duan, X. Ouyang, H. Peng, Clinical profiles of SS-ILD compared with SS-NILD in a Chinese population: a retrospective analysis of 735 patients, Ann. Med. 53 (1) (2021) 1340–1348, https://doi.org/10.1080/07853890.2021.1965205. Dec.

[6] T. Flament, A. Bigot, B. Chaigne, H. Henique, E. Diot, S. Marchand-Adam, Pulmonary manifestations of Sjögren's syndrome, Eur. Respir. Rev. 25 (140) (2016) 110–123, https://doi.org/10.1183/16000617.0011-2016. Jun.

[7] F.J. Martinez, A. Chisholm, H.R. Collard, K.R. Flaherty, J. Myers, G. Raghu, S. L. Walsh, E.S. White, L. Richeldi, The diagnosis of idiopathic pulmonary fibrosis: current and future approaches, Lancet Respir. Med. 5 (1) (2017) 61–71, https://doi.org/10.1016/S2213-2600(16)30325-3. Jan.

[8] W.D. Travis, U. Costabel, D.M. Hansell, T.E. King, D.A. Lynch, A.G. Nicholson, C. J. Ryerson, J.H. Ryu, M. Selman, A.U. Wells, J. Behr, D. Bouros, K.K. Brown, T. V. Colby, H.R. Collard, C.R. Cordeiro, V. Cottin, B. Crestani, M. Drent, R.F. Dudden, J. Egan, K. Flaherty, C. Hogaboam, Y. Inoue, T. Johkoh, D.S. Kim, M. Kitaichi, J. Loyd, F.J. Martinez, J. Myers, S. Protzko, G. Raghu, L. Richeldi, N. Sverzellati, J. Swigris, D. Valeyre, ATS/ERS Committee on Idiopathic Interstitial Pneumonias, An official American Thoracic Society/European Respiratory Society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias, Am. J. Respir. Crit. Care Med. 188 (6) (2013) 733–748, https://doi.org/10.1164/rccm.201308-1483ST. Sep 15.

[9] C.S. Müller, D. Warszawiak, E.D.S. Paiva, D.L. Escuissato, Pulmonary magnetic resonance imaging is similar to chest tomography in detecting inflammation in patients with systemic sclerosis, Rev. Bras. Reumatol. Engl. Ed. 57 (5) (2017) 419–424, https://doi.org/10.1016/j.rbre.2017.02.001. Sep-OctEnglish, Portuguese.

[10] E.V. Efremov, F. Ariese, C. Gooijer, Achievements in resonance Raman spectroscopy review of a technique with a distinct analytical chemistry potential,

Anal. Chim. Acta 606 (2) (2008) 119–134, https://doi.org/10.1016/j.aca.2007.11.006. Jan 14.

[11] S. Li, G. Chen, Y. Zhang, Z. Guo, Z. Liu, J. Xu, X. Li, L. Lin, Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques, Opt. Express 22 (21) (2014) 25895–25908, https://doi.org/10.1364/OE.22.025895. Oct 20.

[12] H. Zhu, S. Liu, Z. Guo, K. Yan, J. Shen, Z. Zhang, J. Chen, Y. Guo, L. Liu, X. Wu, Strong histamine torsion Raman spectrum enables direct, rapid, and ultrasensitive detection of allergic diseases, iScience 24 (11) (2021), 103384, https://doi.org/10.1016/j.isci.2021.103384. Oct 30.

[13] H. Han, X. Yan, R. Dong, G. Ban, K. Li, Analysis of serum from type II diabetes mellitus and diabetic complication using surface-enhanced Raman spectra (SERS), Appl. Phys. B 94 (4) (2009) 667–672, https://doi.org/10.1007/s00340-008-3299-5.

[14] A. Sodo, M. Verri, A. Palermo, A.M. Naciu, M. Sponziello, C. Durante, M. Di Gioacchino, A. Paolucci, A. di Masi, F. Longo, P. Crucitti, C. Taffon, M.A. Ricci, A Crescenzi, Raman spectroscopy discloses altered molecular profile in thyroid adenomas, Diagnostics (Basel) 11 (1) (2020) 43, https://doi.org/10.3390/diagnostics11010043. Dec 29.

[15] C. Cheng, C. Fang, Y. Bo, et al., A novel diagnostic method: FT-IR, Raman and derivative spectroscopy fusion technology for the rapid diagnosis of renal cell carcinoma serum, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 269 (2021), https://doi.org/10.1016/j.saa.2021.120684.

[16] G.W. Auner, S.K. Koya, C. Huang, B. Broadbent, M. Trexler, Z. Auner, A. Elias, K. C. Mehne, M.A. Brusatori, Applications of Raman spectroscopy in cancer diagnosis, Cancer Metastasis Rev. 37 (4) (2018) 691–717.

[17] S. Khan, R. Ullah, A. Khan, et al., Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning, Photodiagn. Photodyn. Ther. 23 (2018) 89–93.

[18] H. Song, C. Dong, X. Zhang, et al., Rapid identification of papillary thyroid carcinoma and papillary microcarcinoma based on serum Raman spectroscopy combined with machine learning models, Photodiagn. Photodyn. Ther. 37 (2022), 102647.

[19] C. Chen, L. Yang, H. Li, et al., Raman spectroscopy combined with multiple algorithms for analysis and rapid screening of chronic renal failure, Photodiagn. Photodyn. Ther. 30 (2020), 101792.

[20] Liang, Z.X. Liu, M.H. Yang, Y.X. Zhang, C.H Wang, Discrimination of variety and authenticity for rice based on visual/near infrared reflection spectra, J. Infrared Milim. Waves 28 (5) (2009) 353–356, https://doi.org/10.3724/SP.J.1010.2009.00353.

[21] C. Meng, H. Li, C. Chen, W. Wu, J. Gao, Y. Lai, M. Ka, M. Zhu, X. Lv, F. Chen, Serum Raman spectroscopy combined with Gaussianconvolutional neural network models to quickly detect liver cancer patients, Spectrosc. Lett. 55 (2) (2022) 79–90, https://doi.org/10.1080/00387010.2022.2027988.

[22] Z. Liao, M.G. Lizio, C. Corden, H. Khout, I Notingher, Feasibility of integrated high avenumber Raman imaging and fingerprint Raman spectroscopy for fast margin assessment in breast cancer surgery, J. Raman Spectrosc. 51 (10) (2020) 1986–1995, https://doi.org/10.1002/jrs.5937.

[23] Y. Li, T. Pan, H. Li, and S. Chen. Non-invasive quality analysis of thawed tuna using near infrared spectroscopy with baseline correction Non-invasive quality analysis of thawed tuna using near infrared spectroscopy with baseline correction. J. Food Process Eng., 2020,43(8):13445.doi:10.1111/jfpe.13445.

[24] X. Zhang, S. Chen, Z. Ling, X. Zhou, D.Y. Ding, Y.S. Kim, F. Xu, Method for removing spectral contaminants to improve analysis of Raman imaging data, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/srep39891.

[25] L. Yann, B. Yoshua, H. Geoffrey, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[26] D. Acharya, A. Rani, S. Agarwal, V. Singh, Application of adaptive Savitzky–Golay filter for EEG signal processing, Perspect. Sci. 8 (3) (2016) 677–679, https://doi.org/10.1016/j.pisc.2016.06.056.

[27] V.D.M. Laurens, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2605) (2008) 2579–2605.

[28] C. Chen, W. Jing, C. Cheng, T. Jun, L. Xiaoyi, M. Cailing, Rapid and efficient screening of human papillomavirus by Raman spectroscopy based on GA-SVM, Optik: Zeitschrift fur Licht- und Elektronenoptik, Opt. J. Light Electronopti. 210 (2020), 164514, https://doi.org/10.1016/j.ijleo.2020.164514.

[29] S. Ameek, P. Els, C. Jan, K Collin, A principal component analysis of polycyclic aromatic hydrocarbon emission in NGC2023, Mon. Not. R. Astron. Soc. 55 (1) (2020) 177–190, https://doi.org/10.1093/mnras/staa317.

[30] B. Yan, Y. Li, G. Yang, Z.N. Wen, M.L. Li, L.J. Li, Discrimination of parotid neoplasms from the normal parotid gland by use of Raman spectroscopy and support vector machine, Oral Oncol. 47 (5) (2011) 430–435, https://doi.org/10.1016/j.oraloncology.2011.02.021. May.

[31] S. Karamizadeh, S.M. Abdullah, M. Halimi, et al., Advantage and drawback of support vector machine functionality, in: Proceedings of the International Conference on Computer, IEEE, 2014.

[32] H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, Eur. J. Oper. Res. 267 (2) (2017) 687–699, https://doi.org/10.1016/j.ejor.2017.12.001.

[33] R. Naseem, B. Khan, M.A. Shah, K. Wakil, A. Khan, W. Alosaimi, M.I. Uddin, B. Alouffi, Performance assessment of classification algorithms on early detection of liver syndrome, J. Healthc. Eng. 2020 (2020), 6680002, https://doi.org/10.1155/2020/6680002. Dec 12.

[34] S. Jafari, T. Dehesh, F. Iranmanesh, Classifying patients with lumbar disc herniation and exploring the most effective risk factors for this disease, J. Pain Res. 12 (2019) 1179–1187, https://doi.org/10.2147/JPR.S189927. Apr 15.

[35] M.L. Zhang, Z.H Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048, https://doi.org/10.1016/j.patcog.2006.12.019.

[36] G. Cho, J. Yim, Y. Choi, J. Ko, S.H. Lee, Review of machine learning algorithms for diagnosing mental illness, Psychiatry Investig 16 (4) (2019) 262–269, https://doi.org/10.30773/pi.2018.12.21.2. Apr.

[37] Q. Liu, B. Pang, H. Li, B. Zhang, Y. Liu, L. Lai, W. Le, J. Li, T. Xia, X. Zhang, C. Ou, J. Ma, S. Li, X. Guo, S. Zhang, Q. Zhang, M. Jiang, Q. Zeng, Machine learning models for predicting critical illness risk in hospitalized patients with COVID-19 pneumonia, J. Thorac. Dis. 13 (2) (2021) 1215–1229, https://doi.org/10.21037/jtd-20-2580. Feb.

[38] E.O. Nsoesie, O. Oladeji, A.S.A. Abah, M.L. Ndeffo-Mbah, Forecasting influenza-like illness trends in Cameroon using Google Search Data, Sci. Rep. 11 (1) (2021) 6713, https://doi.org/10.1038/s41598-021-85987-9. Mar 24.

[39] P. Shanmugam, J. Raja, R Pitchai, An automatic recognition of glaucoma in fundus images using deep learning and random forest classifier, Appl. Soft Comput. 109 (6) (2021), 107512, https://doi.org/10.1016/j.asoc.2021.107512.

[40] S. Hwang, B. Lee, Machine learning-based prediction of critical illness in children visiting the emergency department, PLoS One 17 (2) (2022), e0264184, https://doi.org/10.1371/journal.pone.0264184. Feb 17.

[41] F.B. De Santana, W. Borges Neto, R.J Poppi, Random forest as one-class classifier and infrared spectroscopy for food adulteration detection, Food Chem. 293 (2019) 323–332, https://doi.org/10.1016/j.foodchem.2019.04.073. Sep 30.

[42] N. Chen, H.B. Wang, B.Q. Wu, J.H. Jiang, J.T. Yang, L.J. Tang, H.Q. He, D. D. Linghu, Using random forest to detect multiple inherited metabolic diseases simultaneously based on GC–MS urinary metabolomics, Talanta 235 (2021), 122720, https://doi.org/10.1016/j.talanta.2021.122720. Dec 1.

[43] Group of Pulmonary Vascular and Interstitial Diseases Associated with Rheumatic Diseases, Chinese Association of Rheumatology and Immunology Physicians; Chinese Rheumatic Disease Data Center, 2018 Chinese expert-based consensus statement regarding the diagnosis and treatment of interstitial lung disease associated with connective tissue diseases, Chin. J. Intern. Med. 57 (8) (2018) 558–565.

[44] G.P. Cosgrove, P. Bianchi, S. Danese, D.J. Lederer, Barriers to timely diagnosis of interstitial lung disease in the real world: the intensity survey, BMC Pulm. Med. 18 (2018) 9, https://doi.org/10.1186/s12890-017-0560-x.

[45] H. Ma, J. Lu, Y. Song, H. Wang, S. Yin, The value of serum Krebs von den lungen-6 as a diagnostic marker in connective tissue disease associated with interstitial lung disease, BMC Pulm. Med. 20 (1) (2020) 6, https://doi.org/10.1186/s12890-019-1043-z.

[46] H. Yamakawa, E. Hagiwara, H. Kitamura, Y. Yamanaka, S. Ikeda, A. Sekine, T. Baba, K. Okudela, T. Iwasawa, T. Takemura, K. Kuwano, T. Ogura, Serum KL-6 and surfactant protein-D as monitoring and predictive markers of interstitial lung disease in patients with systemic sclerosis and mixed connective tissue disease, J. Thorac. Dis. 9 (2) (2017) 362–371, https://doi.org/10.21037/jtd.2017.02.48. Feb.

[47] J. Krajczewski, A. Kudelski, Shell-isolated nanoparticle-enhanced Raman spectroscopy, Front. Chem. 7 (2019) 410, https://doi.org/10.3389/fchem.2019.00410. Jun 4.

[48] S.P. Mulvaney, C.D. Keating, Raman spectroscopy, Anal. Chem. 72 (12) (2000) 145R–157R, https://doi.org/10.1021/a10000155. Jun 15.

[49] A.C.S. Talari, Z. Movasaghi, S. Rehman, I. Rehman, Raman spectroscopy of biological tissues, Appl. Spectrosc. Rev. 50 (1) (2014) 46, https://doi.org/10.1080/05704928.2014.923902.

[50] R. Xiao, X. Zhang, Z. Rong, B. Xiu, X. Yang, C. Wang, W. Hao, Q. Zhang, Z. Liu, C. Duan, K. Zhao, X. Guo, Y. Fan, Y. Zhao, H. Johnson, Y. Huang, X. Feng, X. Xu, H. Zhang, S. Wang, Non-invasive detection of hepatocellular carcinoma serum metabolic profile through surface-enhanced Raman spectroscopy, Nanomedicine 12 (8) (2016) 2475–2484, https://doi.org/10.1016/j.nano.2016.07.014. Nov.

[51] W.T. Cheng, M.T. Liu, H.N. Liu, S.Y. Lin, Micro-Raman spectroscopy used to identify and grade human skin pilomatrixoma, Microsc. Res. Tech. 68 (2) (2005) 75–79, https://doi.org/10.1002/jemt.20229. Oct.