Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 4533108, 13 pages
https://doi.org/10.1155/2023/4533108

Research Article

Application of Serum Mid-Infrared Spectroscopy Combined with Machine Learning in Rapid Screening of Breast Cancer and Lung Cancer

Kejing Zhu,[1] **Jie Shen**,[2] **Wen Xu**,[3] **Keyu Yue**,[4] **Liying Zhu**,[5] **Yulin Niu**,[1] **Qing Wu**,[6] **and Wei Pan**[3]

[1]*Organ Transplantation Department, The Affiliated Hospital of Guizhou Medical University, 28 Guiyi Rd, Guiyang, Guizhou 550004, China*
[2]*Department of Clinical Examination, Taixing People's Hospital, 1 Changzheng Rd, Taixing, Jiangsu 225400, China*
[3]*Guizhou Prenatal Diagnosis Center, The Affiliated Hospital of Guizhou Medical University, 28 Guiyi Rd, Guiyang, Guizhou 550004, China*
[4]*Institute of Rail Transit, Tongji University, 4800 Caoan Highway, Shanghai 201804, China*
[5]*Center for Clinical Laboratories, The Affiliated Hospital of Guizhou Medical University, 28 Guiyi Rd, Guiyang, Guizhou 550004, China*
[6]*Innovation Laboratory, The Third Experiment Middle School, Guizhou Key Laboratory for Information System of Mountainous Areas and Protection of Ecological Environment, Guizhou Normal University, 116 Baoshan North Rd, Guiyang, Guizhou 550001, China*

Correspondence should be addressed to Yulin Niu; 304781311@qq.com, Qing Wu; wq401@163.com, and Wei Pan; 313831139@qq.com

Received 21 November 2022; Revised 14 March 2023; Accepted 28 April 2023; Published 9 May 2023

Academic Editor: Vasudevan Rajamohan

Cancer is an increasing burden on global health. Breast and lung cancers are the two tumors with the highest incidence rates. The study shows that early detection and early diagnosis are important prognostic factors for breast and lung cancers. Due to the great advantages of artificial intelligence in feature extraction, the combination of infrared analysis technology may have great potential in clinical applications. This study explores the potential application of mid-infrared spectroscopy combined with machine learning for the differentiation of breast and lung cancers. The experiment collects blood samples from clinical sources, separates serum, trains classification models, and finally predicts unknown sample categories. We use k-fold cross-validation to determine the training set of 301 cases and the test set of 50 cases. Through differential spectrum analysis, we found that the intervals of 1318.59–1401.03 cm^{-1}, 1492.15–1583.27 cm^{-1}, and 1597.25–1721.64 cm^{-1} have significant differences, which may reflect the absorption of key chemical bonds in protein molecules. We use a total of 24 models such as decision trees, discriminant analysis, support vector machines, and K-nearest neighbor to train, identify, and distinguish spectra. The results show that under the same conditions, the prediction model trained based on fine KNN has the best performance and can perform 100% prediction on the test set samples. This also shows that our model has important potential for auxiliary diagnosis of serum breast cancer and lung cancer. This method may help to further achieve comprehensive screening of associated cancers in underserved areas, thereby reducing the cancer burden through early detection of cancer and appropriate treatment and care of cancer patients.

1. Introduction

Cancer is a general term for a group of diseases that can affect any part of the body and refers to many diseases associated with uncontrolled cell mutations caused by tumor expansion [1]. Cancer is currently the leading cause of death in the country and an important barrier to extending the average life expectancy in the country. Worldwide, there are an estimated 19.3 million new

cancer cases and nearly 10 million cancer deaths, based on GLOBOCAN 2020 cancer incidence and mortality estimates compiled by the International Agency for Research on Cancer. Overall, the burden of cancer incidence and mortality is rapidly increasing worldwide. Scholars predict that by 2040, there will be an estimated 28.4 million new cancer cases worldwide, an increase of 47% from 19.3 million in 2020 [2].

Breast cancer is the most commonly diagnosed cancer worldwide. Breast cancer is a disease in which breast epithelial cells proliferate uncontrollably under the action of various carcinogenic factors. Breast cancer accounts for 11.7% of new cases, and its burden has been rising over the past few decades. Now, breast cancer has replaced lung cancer as the most commonly diagnosed cancer worldwide [3]. Breast cancer accounts for a quarter of all cancer cases in women, is by far the most common cancer in women in 2020, and its burden is increasing in many parts of the world [4]. The number of newly diagnosed breast cancers is expected to grow by more than 40% by 2040, to about 3 million cases per year. At the same time, the number of breast cancer deaths will increase from 685,000 in 2020 to 1 million in 2040, an increase of more than 50% [5]. Lung cancer, which used to rank first today, still poses a serious burden to global health. Lung cancer is a malignant tumor originating from the bronchial mucosa or glands of the lung, accounting for 11.4% of new cases. Lung cancer is now the second most commonly diagnosed cancer in 2020. Lung cancer is the second most common cancer in men after prostate cancer and has the highest morbidity and mortality [6, 7].

According to the World Health Organization (WHO), the number of diagnosed breast cancer patients (7.8 million) is much higher than the number of breast cancer deaths (685,000) [2]. These data are smaller than that of other types of cancer, suggesting that breast cancer outcomes can be improved. The key lies in the accurate and timely detection of diseases. Lung cancer, on the other hand, is mostly found incidentally during radiology tests performed in the presence of other symptoms. The disease usually occurs at an advanced stage when it is detected. The treatment of lung cancer is mainly to find and remove the tumor through surgery, supplemented by radiotherapy and/or chemotherapy. Scholars estimate that the five-year survival rate of lung cancer is 70% in the clinical stage [8]. In general, breast cancer and lung cancer are the two tumor types with the largest global health burden. Timely detection and treatment are of great significance for alleviating the status quo of these two diseases. However, most diagnoses of such diseases rely on radiological examination, ultrasound and magnetic resonance imaging (MRI), and tissue biopsy. (These methods are considered the gold standard) [9]. There is an inevitable problem: due to the constraints of cost and expertise, these technologies are not available in rural and remote areas for the detection of these cancers. Therefore, alternative or auxiliary diagnostic techniques that are quick to detect, easy to operate, low cost, and conducive to universal screening must have important clinical needs.

There are issues related to slow reporting of results and delays in patient care due to existing medical workflows. As a result, patient treatment and prognosis are affected, which will bring potential financial burden to medical institutions. In developed countries, the number of people over the age of 65 is expected to increase by 71% by 2050, and the aging population and increasing burden of chronic diseases require higher diagnostic capabilities to diagnose and stratify patients [10]. This will increase the pressure and burden on the existing medical platform. The development of novel, low cost, and rapid diagnostic platforms is the key to breaking through the current dilemma of diagnosis and treatment. The development of novel, low cost, and rapid diagnostic platforms is the key to breaking through the current dilemma of diagnosis and treatment. Infrared spectroscopy (IR) has been used for a long time to characterize compounds, but its suitability in analyzing biological materials with highly complex chemical compositions is debatable. Nevertheless, the latest advancements in infrared spectroscopy have significantly improved its ability to analyze various types of biological specimens. As a result, more and more research is exploring the use of IR spectroscopy in the screening and diagnosis of various diseases [11–13]. In recent two years, Atiqah et al. developed a fast, noninvasive detection method for early diagnosis of osteoarthritis (OA) by analyzing blood serum samples from 15 OA patients and 10 healthy volunteers with no clinical symptoms of the disease. With support from chemical quantification methods, discriminant analysis was used to distinguish between the OA and healthy groups, yielding an overall classification accuracy of 74.47% [14]. As research progresses, the detection limit of infrared spectroscopy has also been increased, with some small molecular metabolic components such as amino acids being detectable by infrared spectroscopy. Thulya et al. improved the detection limit of low molecular weight amino acids by using glycine as a model analyte combined with mid-infrared (MIR) and near-infrared spectroscopy data, with a linear regression coefficient of determination R^2 reaching 0.997. This study successfully expanded the application limits of infrared spectroscopy technology [15]. The potential of using biological specimen analysis as a diagnostic tool for cancer has been recognized for a long time [16]. Ma et al. used infrared spectroscopy to collect serum sample data from patients with cervical cancer, CIN I, CIN II, CIN III, and uterine fibroids and compared various deep learning models. The results showed that PSO–CNN was the best and the discrimination accuracy for the five sample types could reach 87.2% [17]. Wang et al. collected SERS data from 729 patients with prostate cancer or benign prostatic hyperplasia (BPH) in their serum and established an artificial intelligence-assisted diagnostic model based on convolutional neural network (CNN). The accuracy of prostate cancer was $85.14 \pm 0.39\%$. After adjusting the model based on patient age and prostate-specific antigen (PSA), the accuracy of the multimodal CNN could reach $88.55 \pm 0.66\%$ [18]. In recent research, Alexandra et al. investigated the application of ATR-FTIR spectroscopy on serum and used machine learning algorithms to differentiate between cancer ($n = 92$) and healthy controls ($n = 88$), with a sensitivity of 100% and specificity of 100%. A receiver operating characteristic (ROC) analysis yielded an area under the curve (AUC) of 0.95 [19]. There has been a recent trend in the use

of serum infrared spectroscopy in clinical research, with artificial intelligence algorithms gradually replacing traditional, simpler chemical quantification methods to achieve better performing recognition models. The sensitivity of IR spectroscopy to chemical changes during the transition from normal to pathological states or during treatment can lead to the identification of novel biological markers associated with disease. As a result, IR spectroscopy is a powerful tool with enormous potential for clinical applications, extending beyond cancer screening, diagnosis, and prognosis to continuous monitoring of treatment responses and disease progression or regression in personalized medicine [20].

IR is the ability of molecules to selectively absorb infrared rays of certain wavelengths. The result of detecting the absorption of infrared rays is the infrared absorption spectrum of the corresponding substance. The main principle is that in organic molecules, the atoms that make up chemical bonds or functional groups are in a state of constant vibration, and its vibration frequency is equivalent to that of infrared light. When a beam of infrared light with a continuous wavelength passes through a substance, and the vibration frequency or rotation frequency of a certain group in the substance molecule is the same as the frequency of the infrared light, the molecule absorbs energy and transitions from the original ground state vibration (rotation) kinetic energy level to energy higher vibration (rotation) kinetic energy level. Molecules undergo vibration and rotational energy level transitions after absorbing infrared radiation, and the light at this wavelength is absorbed by the substance [21]. The infrared spectrum is obtained by recording the absorption of infrared light by the molecule with an instrument. Based on the above principles, various materials or biological samples can be identified through the unique energy states of specific functional groups. For the detection of vibrational energy states, infrared light can be used, such as near-infrared or mid-infrared light. Since the mid-infrared light can resonate with the fundamental frequency vibration, stronger spectral absorption intensity and more identification features can be obtained. Mid-infrared radiation is usually defined as electromagnetic waves with a wavelength of 2.5–25 μm or a frequency of 400–4000 cm^{-1} (the unit is the wave number, and the product of the constant speed of light is the frequency, the unit is Hz) [22]. Fat, amino acid, protein, sugar, heme, enzyme, hormone, etc. are all organic compounds in living organisms. The key covalent bonds of these compounds are N-H, O-H, C-H, C-C, C-O, C-N, C≡C, C≡N, C=C, C=O, C=N. The typical infrared absorption of chemical bonds is located at 600–4000 cm^{-1}, and both fall in the mid-infrared range. Based on the above analysis, the application of mid-infrared spectroscopy to the analysis of clinical biological samples may have important potential.

The essence of the infrared spectrum of a substance is that when infrared light with a continuous wavelength passes through a substance, the molecular components of the substance absorb infrared radiation under certain conditions. The spectrogram obtained by recording the absorption of infrared light by a molecule with an instrument is the infrared spectrum of the corresponding substance. The analysis of infrared spectroscopy has traditionally relied on the analysis of professional chemometric software. With the rise of infrared spectroscopy in the medical field, the clinical situation is more complicated, such as the differences between biological samples, and the specific course of the disease of individual patients is also different [23]. These problems make infrared-related research still need to solve many problems, such as infrared data processing requires stronger computing power, and also need to be "smart" to realize our data processing needs. This has caused the traditional chemometrics software to be difficult to meet our needs. The essence of each piece of infrared spectral data is a collection of thousands of wave points. Each wave point is a set of data (wavenumber + absorbance). In other words, each infrared spectrum is a data set consisting of thousands of data. Artificial intelligence (AI) is a branch of computer science aimed at creating computer systems that can think, learn, reason, and make decisions independently. The goal is to simulate various aspects of human intelligence, enabling computers to perform tasks that typically require human thought, intellect, knowledge, or skills. AI is a broad field that encompasses machine learning, natural language processing, image recognition, and intelligent robots, among others [24]. Machine learning, a subfield of AI, focuses on allowing computers to automatically generate models through the study of large amounts of data, in order to recognize and predict patterns in the data [25]. This aligns with our requirements for analyzing infrared spectroscopy data, and the application of machine learning to the analysis and processing of such complex data may yield satisfactory results [24]. We have previously used similar approaches to solve some clinical problems and obtained satisfactory results [26]. This study builds and develops new methods for assisting in the diagnosis of breast and lung cancers based on decision trees, discriminant analysis, Bayesian classification, support vector machines, and K-nearest neighbors in machine learning. It aims to quickly identify and distinguish breast and lung cancers in the population.

2. Materials and Methods

2.1. Sample Collection. Serum samples were obtained from the Affiliated Hospital of Guizhou Medical University (Guiyang, China). They come from 98 cases of breast cancer patients diagnosed by clinicians admitted to the Affiliated Hospital of Guizhou Medical University. The age of the patients ranged from 24 to 81 years old, with an average age of (50.86 ± 12.04) years old. Among these patients, there were 93 cases of invasive ductal carcinoma of the breast, 1 case of invasive ductal carcinoma of the breast with intraductal carcinoma, 1 case of invasive ductal carcinoma of the breast with ductal carcinoma in situ, 1 case of invasive ductal carcinoma of the breast with eczema-like carcinoma, and 1 case of invasive ductal carcinoma of the breast with eczema-like carcinoma. Two cases of sex cord-stromal tumors concurrently with fibroadenomas were observed. Another part of the samples came from 95 patients diagnosed with lung cancer who were diagnosed by clinicians in the Affiliated Hospital of Guizhou Medical University. The age

of the patients ranged from 31 to 83 years old, with an average age of (57.18 ± 9.81) years old including 51 cases of adenocarcinoma (ACC), 28 cases of squamous cell carcinoma (SCC), 7 cases of large cell carcinoma (LCC), and 9 cases of small cell lung cancer. In addition, 158 serum samples were collected from healthy people in the Affiliated Hospital of Guizhou Medical University in June 2021, aged 26–74 years, with an average age of (47.75 ± 10.49) years. This study was approved by the Human Research Ethics Committee of the Affiliated Hospital of Guizhou Medical University.

2.2. Instrument and Software. This study utilized the Thermo Scientific Nicolet iS5 Fourier transform infrared spectrometer with accompanying iD1 smart transmission accessory and iD5 attenuated total reflectance accessory (Thermo Fisher Scientific, Waltham, MA, USA). Software analysis was conducted using the TQ Analyst 9 and OMNIC 8.2 (Thermo Fisher Scientific, Waltham, MA, USA), as well as MATLAB2019 (MathWorks, Natick, MA, USA).

2.3. Sample Collection and Processing. The blood is collected by the clinical nurse, and the fast serum tube is selected for temporary storage and transferred to the laboratory at 4° cold chain. After confirming that there is no hemolysis, after confirming the receipt, let it stand at room temperature for 30 minutes. Centrifuge at 3500 r/min for 5 minutes to separate the serum and store at −80°.

2.4. Work Flow Establishment and Serum Spectrum Collection. Open Omnic software (Omnic 8.2, Thermo Nicolet Corporation, Waltham, MA, USA) creates a new workflow and sets experimental parameters. The number of scans set in "Number of scans" is 16. In the "Resolution" option, ensure that the resolution of the collected spectra is 4 cm^{-1}. In the "Data Format" option, confirm that Absorbance is the data format of the spectrum. The data interval is 0.482. The range of scanning wavenumber is 4000~400 cm^{-1}. The detector is DTGS/KBr, and the beam splitter is KBr. We remove the samples in a −80°C freezer. Considering that room temperature and relative humidity may affect serum status, we controlled the room temperature at 24°C and 44% humidity. We equilibrated the samples for 2 hours in this environment. Then, turn on the instrument (Thermo Scientific Nicolet iS5, USA) to preheat for 0.5 h. Prior to testing, we cleaned the instrument sample pool and performed air zeroing. We then injected 5 μL of the sample into the sample pool using a sampling gun. To eliminate the influence of background substances such as water, a solvent blank was independently prepared [27]. Spectra were collected according to the set parameters, and spectral data within the range of 550–1800 cm^{-1} were retained for subsequent analysis. The background spectrum was collected as shown in Figure S1-A, and the collected sample spectrum (S1-B) was subtracted from the background spectrum to obtain the final sample spectrum, as shown in Figure S1-C. To effectively avoid errors caused by instrument instability, each spectrum was normalized uniformly, as shown in Figure S1-D.

2.5. Spectral Data Analysis and Removing Outlier Samples. OMNIC 8.2 (OMNIC 8.2, Thermo Nicolet Corporation, Waltham, MA, USA) was used for spectrometer control. TQ Analyst 9 (Thermo Fisher Scientific, Waltham, MA, USA) was used for primary analysis of the data. Open the raw spectrum data with OMNIC and save it as a CSV text file that Matlab can recognize. Import the spectrum file into Matlab to save a new workspace. Eliminate abnormal spectra caused by abnormal conditions such as sample addition, operation, and program error reporting during the experiment.

2.6. Optimization Algorithm

2.6.1. Classification of Training and Test Sets. The key to the study is the normal division of the dataset. The classification in this study follows the following principles: the training set should be mutually exclusive with the test set as much as possible, and the distribution of training and test data should be consistent. Combined with the characteristics of the sample data itself, we choose k-fold cross validation, and the specific steps are as follows:

Step 1: Divide the data set into training set and test set completely randomly. Set aside the test set as validation data for the trained model.

Step 2: Randomly divide the training set into K mutually exclusive subsets of similar size.

Step 3: Each time use 1 of the k copies as the verification set, and all the others as the training set.

Step 4: After k training and validation, we get k different models.

Step 5: Evaluate the effects of k models, select the hyperparameters with the best effect, and take the average of k verification results as an index to measure the accuracy of the model.

Step 6: Use the optimal hyperparameters, and then use all k data sets as the training set to retrain the model, and output the final model.

2.6.2. Decision Tree. The decision tree regards a spectrum with n wave points as a point in n-dimensional space. The process of classification is to find a hyperplane in n-dimensional space or higher dimensional space and divide these spectra. Decision trees are supervised learning. Based on the known classification results, a prediction model is obtained by learning the training set samples, performing feature selection, generating a decision tree, and pruning the decision tree. This decision tree is able to give the correct classification for new spectral data. Decision trees are a common machine learning algorithm.

2.6.3. Discriminatory Analysis. Discriminant analysis is used to find which category is most similar to the unknown category spectrum among the known categories. The analysis process is to establish a discriminant function by estimating the change of the number of each wave point in the analysis area with a batch of spectral samples that have been

clearly classified, so as to minimize the misjudgment of the spectrum. On this basis, a given new spectral sample is discriminated.

2.6.4. Bayesian Classification. Bayesian classification is a statistical classification method. Its formula is $P(Ai|B) = P(Ai) * P(B|Ai)/\sum_{j=1}^{n} P(Aj)P(B/Aj)$, where $P(Ai|B)$ is the posterior probability distribution, and $P(Ai)$ is the prior probability distribution. The prediction process is based on the first training to build a classifier on the spectral data. Predict the probability that a spectrum belongs to a certain class by solving the posterior probability distribution.

2.6.5. Support Vector Machine (SVM). SVM is a class of supervised learning. The basic idea of learning is to solve the separating hyperplane that can correctly divide the training data set and maximize the geometric interval. Based on this, binary classification is performed on the data. Its learning strategy is to maximize the interval between different categories of spectra, which can be formalized as an optimization algorithm for solving convex quadratic programming.

2.6.6. k-Nearest Neighbor (KNN). The idea of KNN is that in the feature space, if most of the k-nearest samples near a sample belong to a certain category, the sample also belongs to this category. The implementation process is that in a training spectral data set, input unknown spectral data, and find the K samples closest to the spectrum in the training data set. Most of these K samples belong to a certain class, and the input instance is classified into this class.

2.7. Metrics for Evaluation. After the quantitative model is established, its performance needs to be evaluated. The main inspection indicators are accuracy, errorrate, macro precision (macro-P), macro recall (macro-R), harmonic mean ($F1$), and predictive recall (R). The main calculation formula is as follows:

$$
\begin{aligned}
P - \text{Accuracy}(\%) &= \frac{\sum_{i=1}^{n} TPi}{\sum_{i=1}^{n} TPi + FPi}, \\
P - \text{Errorrate}(\%) &= \frac{\sum_{i=1}^{n} FPi}{\sum_{i=1}^{n} TPi + FPi} \times 100, \\
\text{macro} - P &= \frac{1}{n} * \sum_{i=1}^{n} \frac{TPi}{TPi + FPi}, \\
Ri &= \frac{TPi}{TPi + FNi}, \\
\text{macro} - R &= \frac{1}{n} * \sum_{i=1}^{n} Ri, \\
F1 &= \frac{2 \times \text{macro} - P \times \text{macro} - R}{\text{macro} - P + \text{macro} - R}.
\end{aligned}
\tag{1}
$$

Note: TPi (True Positive): the true classification i is correctly predicted as classification i; FPi (False Positive): the wrong prediction of other true classifications is classification i; FNi (False Negative): the true classification i is incorrectly predicted as other Classification. Ri (Recall): the number of samples correctly predicted as category i accounts for the actual number of samples of category i. P-Accuracy: the accuracy of the test set sample prediction. P-Errorrate: the test set sample prediction error rate.

3. Results and Discussion

3.1. Sample Conditions and Basic Characteristics of Mid-Infrared Spectra. All collected samples were assayed by mid-infrared spectroscopy as described in Materials Method 4. A total of 351 spectral samples were analyzed for this analysis. According to the sample source, the samples were divided into normal control group, breast cancer group, and lung cancer group. We randomly split the collected spectra into a training set (301) and a test set (50). Please refer to Table 1 for details such as grouping situation and gender ratio. There was no gender difference between the training set and the prediction set ($X^2 = 0.567$, $P = 0.451$, $P > 0.05$). There was no age difference between the training set and the prediction set ($t = 0.91$, $P = 0.927$, $P > 0.05$). Through statistical analysis, the results of random grouping can be used in subsequent experiments. The quality of a model depends on objective evaluation criteria. Existing standards are basically applicable to a single assessment, which leads to the randomness of the assessment. We need to increase the number of training times to solve the contingency problem. Training a model repeatedly on the same dataset does not solve this problem. Therefore, the preferred method is to divide the data set differently, train multiple models, and evaluate comprehensively. The common division methods are as follows: holdout cross validation, leave one out cross validation, and k-fold cross validation. Holdout cross validation is to statically divide the data set into training set, verification set, and test set. The model corresponding to the test set may have a large gap compared with the trained model, the fidelity is low, and there is no better solution. This method is suitable for a large number of samples (ten thousand, million). Leave one out cross validation is to divide the training set one by one. Each validation set has only one sample, and m times of training and prediction are required. This will increase the complexity of training and increase the computational cost. This method is generally used when the data set is scarce. Static holdout cross validation is sensitive to data division. k-fold cross validation is a dynamic validation method [28]. This method is not affected by the random sample division method, and the trained model is very similar to the overall data set model, which is an ideal division scheme for small data sets. In summary, combined with the characteristics of the sample data itself, we choose k-fold cross validation ($K = 5$). The specific process is shown in Figure 1. Figure 2 is the original mid-infrared spectrum of the collected samples. The overall spectrum is uniform and consistent, which can better reflect the physical and chemical properties of serum. Each

Table 1: Sample information and its grouping situation.

	Number	Male/female	Normal group (N)	Breast cancer (B)	Lung cancer (L)	Age (Mean ± SD)	CV
Training dataset	301	146/205	131	85	85	51.19 ± 11.51	0.22
Test dataset	50	18/32	27	13	10	51.04 ± 10.95	0.21

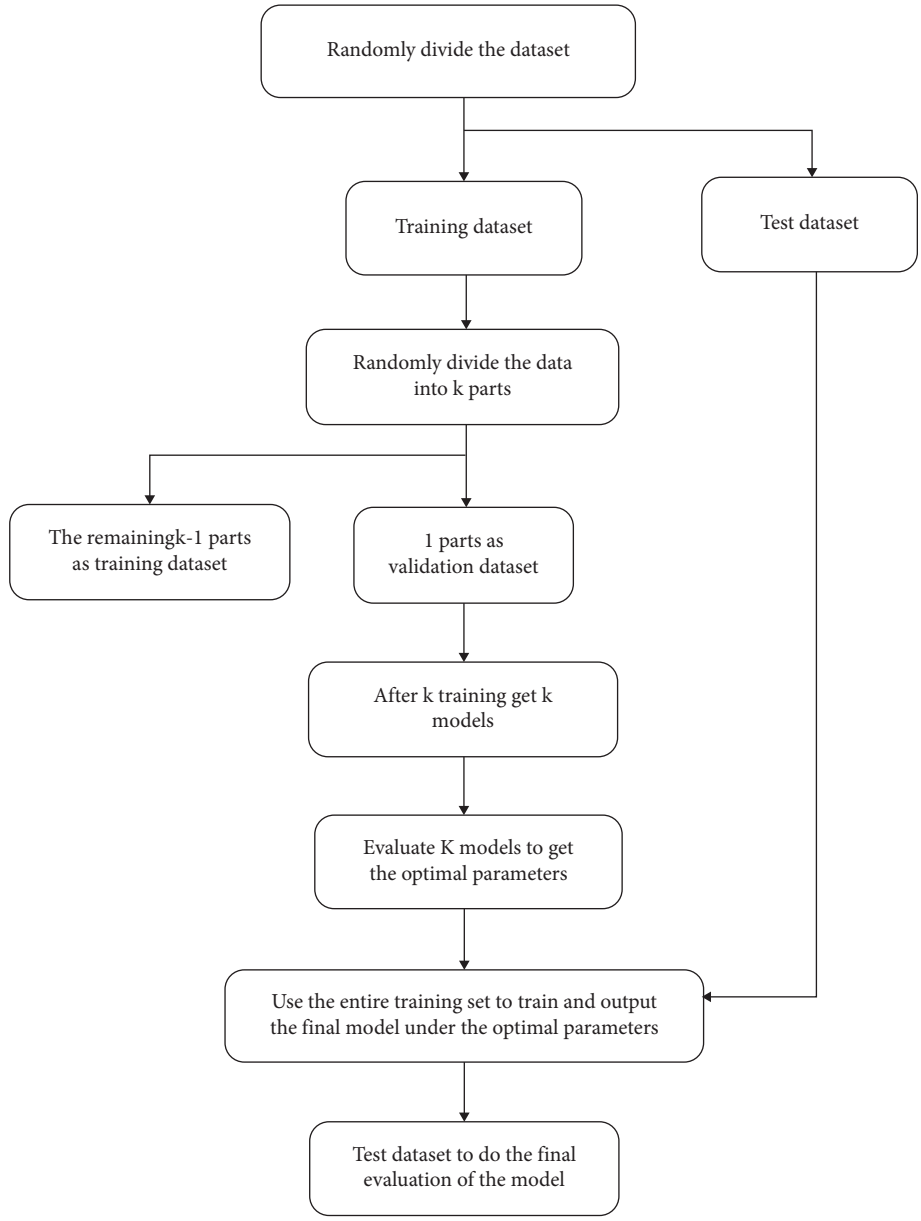


Figure 1: Dataset partitioning process.

spectrum consists of 2595 wave points; that is, each spectrum has 2595 features.

3.2. Difference Spectral Analysis. Based on the collected mid-infrared spectra, the training set spectra were divided into 131 normal control spectra, 85 breast cancer spectra, and 85 lung cancer spectra according to the sample source. Convert the final IR spectra to 2595 wave point data using the method described in Materials Methods 5. In order to eliminate the inherent errors in instrument measurement and analyze the spectra better, we first normalized each spectrum to obtain three sets of new data. After that, we calculated the average value of the 2595 wave point data that constitute the spectrum in the same group one by one. Finally, we reconstructed the obtained average result into an average spectrum, and the result is shown in Figure 3(a). In organic components, when the vibration frequency or rotation frequency of atoms constituting chemical bonds or functional groups is the same as that of infrared light, molecules

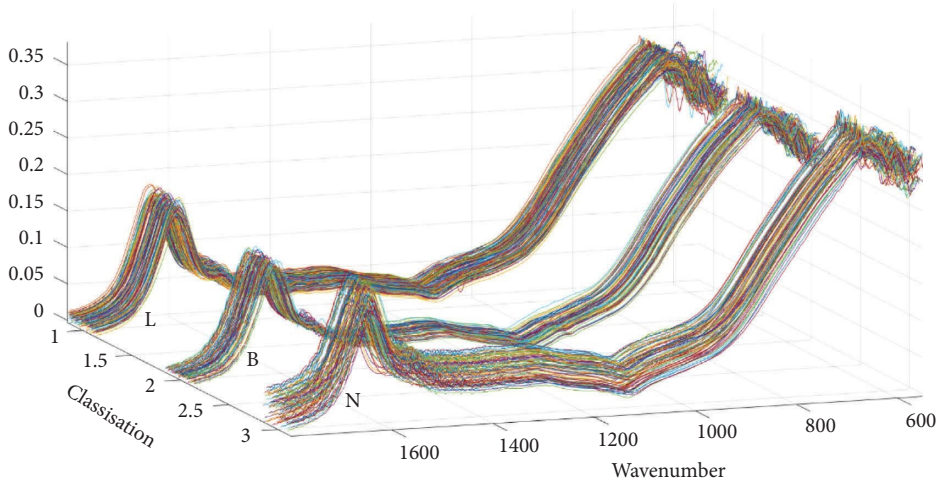
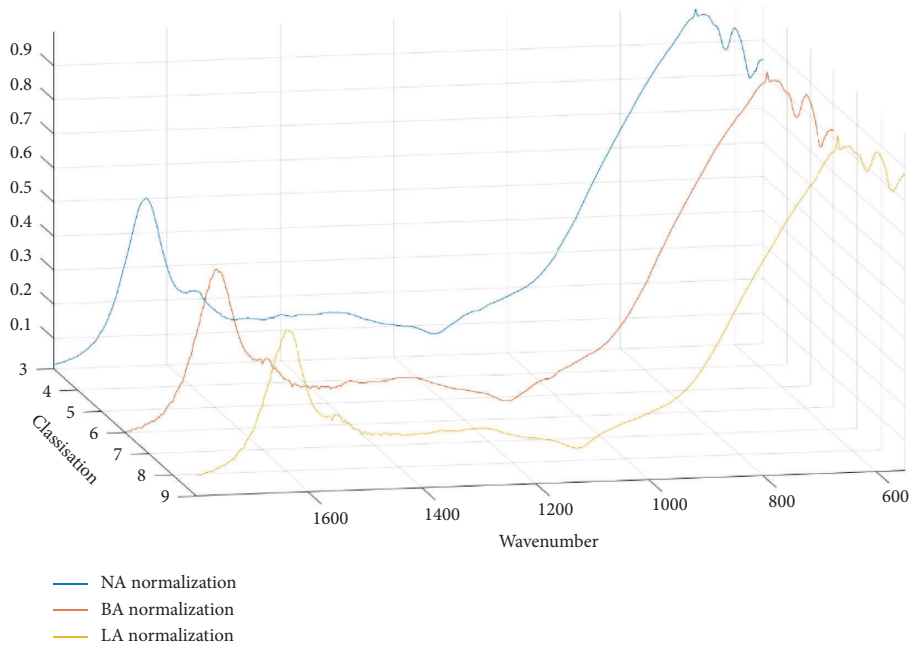
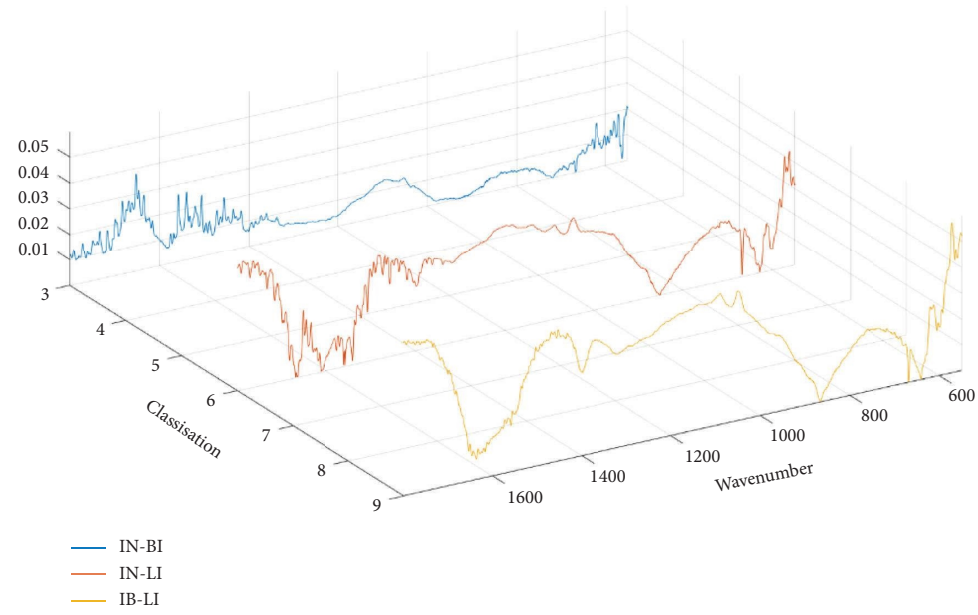


Figure 2: Mid-infrared spectroscopy ($550\,cm^{-1} \leq$ wavenumber $\leq 1800\,cm^{-1}$). N: normal group spectroscopy; B: breast cancer group spectroscopy; L: lung cancer group spectroscopy.



(a)

Figure 3: Continued.



(b)

Figure 3: Analysis of the overall characteristics of the spectrum (550 cm^{-1} ≤ wavenumber ≤ 1800 cm^{-1}): (a) average spectrum after spectral normalization and (b) the difference between the mean spectra.

will undergo vibration and rotation energy level transitions after absorbing infrared radiation [21]. Based on the above principles, theoretically specific functional groups or biochemical components have corresponding absorption in the frequency range of specific infrared light (*S*2). We subtracted the mean spectra of each group pairwise to obtain their absolute values and reconstructed them into three difference spectra. They are normal control-breast cancer (|N-B|), normal control-lung cancer (|N-L|), and breast cancer-lung cancer (|B-L|), and the difference spectrum is shown in Figure 3(b). The spectral intervals with large differences between different average spectra are marked, and the detailed results are shown in Table 2. The spectral differences between breast cancer and normal population were mainly concentrated in the ranges of 1492.15–1583.27 cm^{-1} and 1597.25–1721.64 cm^{-1}. This region is mainly the spectral region of C=C, C=O, C=N, N-H, and O-H absorptions. This result may be related to the absorption of the amide structure of the protein with the α-helical structure. It reflects the difference in protein content and its secondary structure, which is consistent with the findings of Zelig and Yang [29, 30]. The spectral difference between lung cancer and normal people not only reflects the amide I region (1398.62–1543.74 cm^{-1}) [31–33] corresponding to protein absorption but also reflects the phosphate group region of RNA and DNA (1172.99–1319.55 cm^{-1}). Spectral differences between breast and lung cancers were also significant in the range 1318.59–1515.77 cm^{-1}, a region containing the absorption of key structures of lipids and phospholipids [30, 34]. In general, the spectra of the different populations are significantly different or can be effectively distinguished. This suggests that spectroscopy has the potential for important clinical applications. From a clinical perspective, modern clinical practice requires a simple, fast, and accurate auxiliary diagnostic method. The sample drying process is cumbersome and does not meet the practical needs of clinical practice. Related research reports in the past year have also shown that satisfactory results can be obtained from serum samples [15, 17–19]. It should be noted that although biochemical components require relatively high concentrations to prevent their biochemical characteristics from being masked, this study's results show that the differences between spectra are mainly concentrated in the wavenumber range (1700 cm^{-1}~1500 cm^{-1}) that reflects key protein structures. This means that distinguishing different categories based on spectra may still depend on differences in protein abundance. This result also coincides with the marked differences in protein content between cancer patients and the normal population in clinical practice. At the same time, this is consistent with similar studies in the same category [29, 30, 32, 33]. This further demonstrates the important clinical application potential of spectroscopy.

3.3. Model Results and Partial Model Performance Verification. On the basis of differential spectrum analysis, different classification models are used for training in combination with the characteristics of different groups of spectra. For the complex spectral data composed of 2595 wave points, it is necessary to reduce the dimensionality of the data. This study uses the commonly used PCA dimensionality reduction. Since the feature score after dimensionality reduction is determined by the principal component score after dimensionality reduction, this will directly affect the model training effect. Therefore, the determination of the principal components is an important

Table 2: The interval with more significant difference in difference spectrum.

	N-B		N-L		B-L	
	Wavenumber range	Difference value	Wavenumber range	Difference value	Wavenumber range	Difference value
Interval1	1305.57–1401.03 cm^{-1}	0.0077 ± 0.0016	1398.62–1543.74 cm^{-1}	0.0298 ± 0.0085	1024.49–1254.95 cm^{-1}	0.0458 ± 0.0028
Interval2	1408.26–1482.99 cm^{-1}	0.0121 ± 0.0030	1632.93–1656.53 cm^{-1}	0.0164 ± 0.0042	1318.59–1401.03 cm^{-1}	0.0396 ± 0.0033
Interval3	1492.15–1583.27 cm^{-1}	0.0142 ± 0.0048	1677.28–1702.84 cm^{-1}	0.0203 ± 0.0058	1414.53–1515.77 cm^{-1}	0.0466 ± 0.0026
Interval4	1597.25–1721.64 cm^{-1}	0.0199 ± 0.0064	1172.99–1319.55 cm^{-1}	0.0368 ± 0.0031	1672.95–1752.01 cm^{-1}	0.0474 ± 0.0087

parameter for model establishment. In our research, we mainly use models including decision trees, discriminant analysis, SVM, kNN, and related subalgorithms, a total of 24 models. Different algorithm models and different principal component numbers are selected, and the accuracy results of the training model are shown in Figure 4. Output part of the model and use the test set that did not participate in the model training to make predictions. This was used to check the performance of the trained model, and the relevant results are shown in Table 3. When further analyzing the results, an overall improvement in the training performance of the model was observed with the increase of principal components. However, the accuracy of fine Gaussian SMV and linear discriminant increases first and then gradually decreases, which means that there is a limit to model optimization, and model performance will not increase infinitely with the increase of principal components. Second, the higher the model accuracy, usually the better the model performance. However, we believe that models trained by cross validation may not be able to completely avoid overfitting. Therefore, in order to more objectively evaluate the performance of the trained model, a set of data from the same source but not involved in training is needed as a test set for further verification. The results show that boosted tree has the worst training effect among the 24 algorithms, and the test set is all identified as "N," which means that not all algorithms can adapt to such spectral data, so it is necessary to choose an appropriate algorithm to obtain performance relatively best model. It should be noted that due to the reduction of the spectral range, some information is lost, and the performance of the model is affected to a certain extent. Although it was found during the training process that the accuracy of the model increases with the increase of principal components, pursuing only the improvement of accuracy inevitably requires more principal components, resulting in an increase in computation and complexity of the model. Therefore, we believe that before making a decision, these issues need to be considered comprehensively to balance the model's predictive performance, accuracy, computational complexity, and other parameters. The experimental results (Table 3) show that the training model and prediction model of quadratic discriminant, cubic SMV, and fine KNN have achieved relatively good results. In the case of similar accuracy, we consider the computational cost and efficiency of the model and believe that the fine KNN model with a principal component of 3 is superior to other algorithms. In other similar studies, researchers included approximately 29–74 patient samples, with accuracy rates ranging from 72% to 95.7% [29, 31, 34, 35]. In fact, for machine learning, sample size is a key factor affecting the quality of model training. After we increased the training set and test set, it also reflected a relatively good effect. In the future, we plan to further expand the sample size to correct the model and lay a preliminary experimental foundation for further clinical applications in the future.

3.4. Preferred Model. Comprehensive analysis of the fine KNN model based on 3 principal components has relatively the best performance. Through the prediction of the test set, *P*-Accuracy is 98.00%, and Macro-*P*, Macro-*R* and Macro-*F*1 are 98.80%, 97.43%, and 98.06%, respectively. PCA dimensionality reduction is performed on the data, and the data are distributed in three-dimensional space. It can be seen that the spectral data points of different categories show a discrete trend, and the spectral data points of the same category show a trend of aggregation, as shown in Figure 5(a). The spectral data points are shown in the 2D plane that was correctly classified during training. The results show that each spectral data is correctly classified, and the training effect is good. The results are shown in Figure 5(b). Export the trained model for prediction on the test set. The predicted spectral results of each group were almost always correctly classified. The results show that the model is not overfitting, and the results are shown in Figure 5(c). We usually use the receiver operating characteristic curve (ROC) to measure the method's ability to identify diseases. The area under the ROC curve (AUC) is between 0.5 and 1. In the case of $AUC > 0.5$, the closer the AUC is to 1, the better the diagnostic effect. AUC has lower accuracy when it is 0.5–0.7. AUC has a certain accuracy when it is 0.7–0.9. The accuracy is higher when the AUC is above 0.9. In this study, $AUC = 0.98$, indicating that the model has excellent accuracy, and the results are shown in Figure 5(d). The KNN method is a simple but effective classification method that classifies examples based on the intuitive premise and similar data points in close proximity in the feature space [36]. In this study, the fine KNN performed very well in the spectral discrimination of related diseases, which has important potential for clinical application. Cancer patients are usually tested with x-rays. Additional evaluation with ultrasonography or magnetic resonance imaging is usually required for high-risk patients with suspicious radiographic findings [37]. This requires medical institutions to provide corresponding equipment and relevant imaging specialists. Equipment costs and personnel costs are unaffordable in some areas. Alternatives such as thermography, transillumination, and positron emission tomography have not been shown to be more sensitive or specific than radiography. Therefore, the mid-

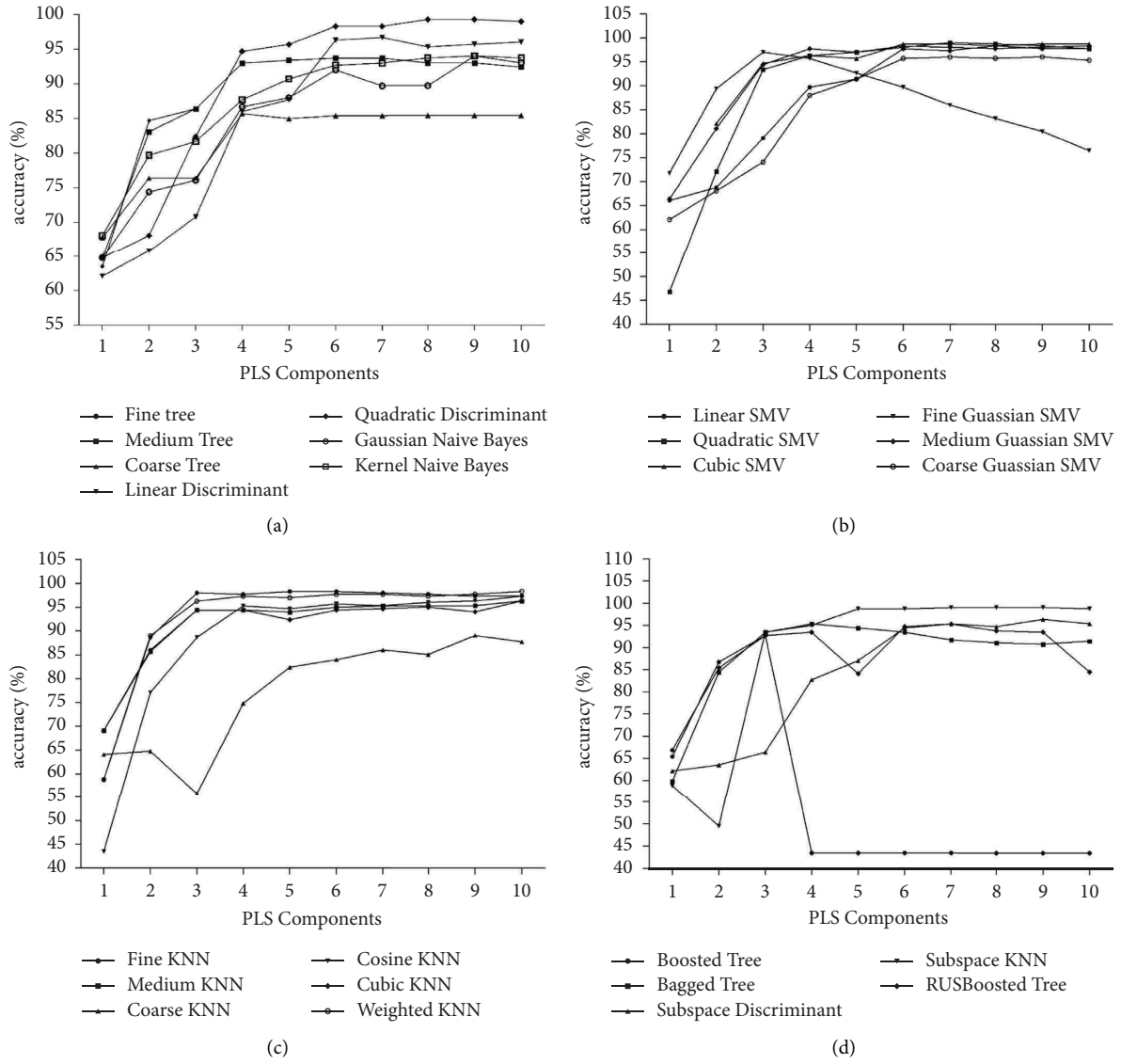


Figure 4: The training results of each model: (a) decision trees models, discriminant analysis models, Naive Bayes models, (b) SMV models, (c) KNN models, and (d) other models.

Table 3: Partial model performance verification.

	PLS	Accuracy (%)	*P*-Errorrate (%)	*P*-Accuracy (%)	Macro-*P* (%)	Macro-*R* (%)	Macro-*F*1 (%)
Kernel naïve Bayes	1	68.10	28.00	72.00	76.31	73.20	73.98
Coarse tree	3	76.40	18	82.00	85.00	87.55	84.52
Subspace discriminant	5	87.00	10.00	90.00	89.27	91.16	90.13
Cubic SMV	7	99.00	2.00	98.00	98.80	97.43	98.06
Quadratic discriminant	9	99.30	2.00	98.00	98.80	97.43	98.06
Fine KNN	3	98.00	2.00	98.00	98.80	97.43	98.06

infrared spectroscopy with simple operation, rapid detection, no damage to the sample, and no reagent addition is a promising alternative. Mid-infrared spectrometers, on the other hand, can produce unique spectral "fingerprints" by detecting the absorption of infrared light by organic compounds at energies (wavenumbers) that correspond to the properties of the bonds between their atoms. In fact, water in serum accounts for about 90%, and the solvation of water and the change of cluster structure have a great influence on the structure of solute, and the two influences each other. Therefore, the essence of this spectral "fingerprint" is the infrared spectrum of solvent water that contains a large amount of information about the solute in the solution, which is also the theoretical basis of aquaphotomics [38–40]. So we think that by measuring serum spectra, it is theoretically possible to identify biochemical differences in

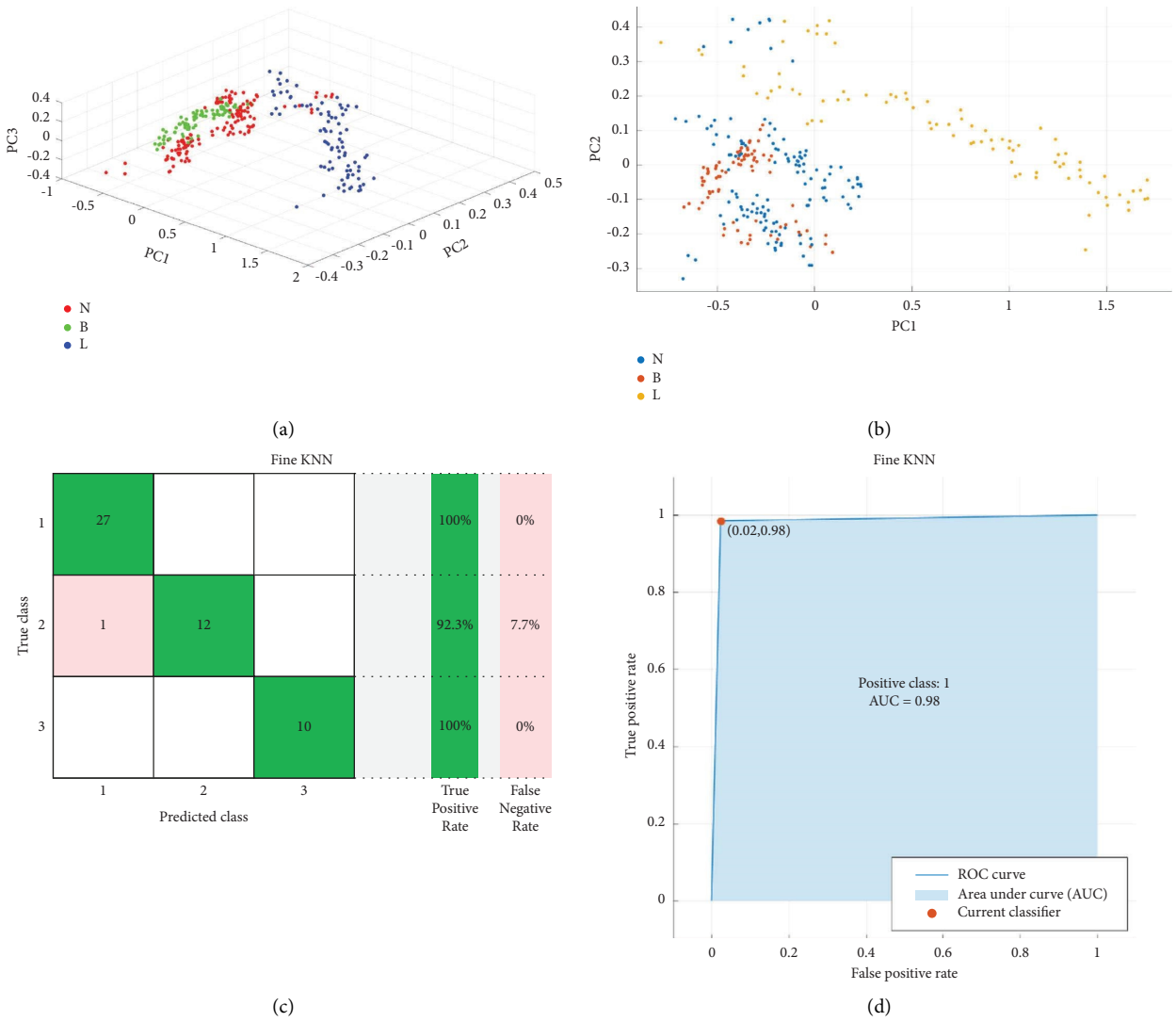


Figure 5: Fine KNN training station: (a) three-dimensional display of the spatial distribution of the principal component dimension reduction spectrum (PC1, PC2, PC3 as examples); (b) scatter plot: two-dimensional plane displays principal component dimension reduction spectral training and cross-validation results (PC1, PC2, PC3); (c) confusion matrix; (d) receiver operating characteristic curve.

patient samples that are associated with disease. On this basis, combining the spectral "fingerprint" with the fine KNN model may be able to assist clinical screening of high-risk breast and lung cancer patients in the general population at a lower cost and higher sensitivity in the future.

4. Conclusion

The burden of cancer in the global society is increasing. Breast cancer and lung cancer are the top two cancers worldwide. New cases account for the majority of cancer patients worldwide. At present, breast cancer and lung cancer are the two tumors with the largest health burden in the world. Early diagnosis is an important prognostic factor. Timely detection, differentiation, and treatment are of great significance for alleviating the status quo of these two diseases. Infrared spectroscopy is generally considered a simple, reagent-free, inexpensive, noninvasive, and nondestructive technique. Therefore, it has been widely used in pharmaceutical, food, environmental, and forensic industries. In fact, infrared technology has begun to show great potential in clinical application.

This study established a simple, noninvasive rapid screening method for the auxiliary diagnosis of breast cancer and lung cancer. First, in a pilot experiment, we developed the model using TQ Analyst 9 (Thermo Fisher Scientific, Waltham, MA, USA) using the traditional method. The pre-experimental results show that the model established directly with the original spectral data as a variable is not effective and cannot meet the research needs. To do this, we convert the spectrum into a matrix of numbers. We train the spectral data based on different algorithms and output the corresponding prediction models. At the same time, combined with the characteristics of the data itself, k-fold cross validation and random algorithm are used to divide the

training set and test set to solve the problems of chance and overfitting. From the difference spectrum results, there are significant differences in the mid-infrared spectra between different types of sera. These differences may mainly reflect protein abundance expression. This has an important prompting effect for the subsequent exploration of cancer classification. We compare decision trees, discriminant analysis, SVM, kNN and other 24 models, and fine KNN, which is superior to other models in terms of training model accuracy, calculation cost, and prediction accuracy, trains the data, and outputs the prediction model. In summary, this study established a rapid screening method for the auxiliary diagnosis of breast cancer and lung cancer based on serum mid-infrared spectroscopy. Under the same conditions, the fine KNN model is superior to the classification algorithm model. In addition, compared with traditional cancer detection methods, this nondestructive rapid detection method has obvious advantages, or it can realize rapid screening of serum breast cancer and lung cancer. This may provide new ideas for the development of new detection methods in the medical field. It also allows us to see the important potential of identifying many different cancers simultaneously from a single serum. This is also the direction of our later efforts, but it requires more clinically diagnosed samples of different categories to train the model.

Data Availability

The data supporting the current study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Kejing Zhu (first author) contributed to design the experimental methods, did actual investigation and research, analyzed the experimental data, and wrote the first draft of the paper. Jie Shen did actual investigation and research. Wen Xu performed paper review and revision. Keyu Yue contributed experimental data analysis and experimental results visualization. Liying Zhu discussed the experimental results and formulated further research directions. Yulin Niu (corresponding author) performed acquisition of research funds and collected research resources. Qing Wu (co-corresponding author) performed the supervision and guidance of the research project. Wei Pan (co-corresponding author) contributed the generation of the research concept, the verification of the experimental design, and the review and revision of the paper.

Acknowledgments

The authors are thankful for the financial support by the Guiyang Science and Technology Plan Project ([2019] 9-1-39) and Basic Research Program of Guizhou Province (Guizhou Science and Technology Combination Foundation-ZK [2023] General 380).

Supplementary Materials

Supplementary File 1. Serum spectrum collection; Supplementary File 2. Correspondence between wave number and matter. (*Supplementary Materials*)

References

[1] S. G. Kandlikar, I. Perez-Raya, P. A. Raghupathi et al., "Infrared imaging technology for breast cancer detection–Current status, protocols and new directions," *International Journal of Heat and Mass Transfer*, vol. 108, pp. 2303–2320, 2017.

[2] Z. Shi, W. Cai, X. Feng et al., "Radiomics analysis of Gd-EOB-DTPA enhanced hepatic MRI for assessment of functional liver reserve," *Academic Radiology*, vol. 29, no. 2, pp. 213–218, 2022.

[3] B. O. Anderson, A. M. Ilbawi, E. Fidarova et al., "The Global Breast Cancer Initiative: a strategic collaboration to strengthen health care for non-communicable diseases," *The Lancet Oncology*, vol. 22, no. 5, pp. 578–581, 2021.

[4] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. Fidler-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," *Lancet Global Health*, vol. 8, no. 8, pp. e1027–e1037, 2020.

[5] M. Arnold, E. Morgan, H. Rumgay et al., "Current and future burden of breast cancer: global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022.

[6] J. Ferlay, M. Colombet, I. Soerjomataram et al., "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International Journal of Cancer*, vol. 144, no. 8, pp. 1941–1953, 2019.

[7] K. Chaitanya Thandra, A. Barsouk, K. Saginala, J. Sukumar Aluru, and A. Barsouk, "Epidemiology of lung cancer," *Współczesna Onkologia*, vol. 25, no. 1, pp. 45–52, 2021.

[8] E. Kaznowska, K. Łach, J. Depciuch et al., "Application of infrared spectroscopy for the identification of squamous cell carcinoma (lung cancer). Preliminary study," *Infrared Physics and Technology*, vol. 89, pp. 282–290, 2018.

[9] P. Zubor, P. Kubatka, K. Kajo et al., "Why the gold standard approach by mammography demands extension by multiomics? Application of liquid biopsy miRNA profiles to breast cancer disease management," *International Journal of Molecular Sciences*, vol. 20, no. 12, p. 2878, 2019.

[10] D. E. Bloom, A. Boersch-Supan, P. McGee, and A. Seike, "Population aging: facts, challenges, and responses," *Benefits and Compensation International*, vol. 41, no. 1, p. 22, 2011.

[11] A. Travo, C. Paya, G. Déléris, J. Colin, B. Mortemousque, and I. Forfar, "Potential of FTIR spectroscopy for analysis of tears for diagnosis purposes," *Analytical and Bioanalytical Chemistry*, vol. 406, no. 9-10, pp. 2367–2376, 2014.

[12] J. Titus, E. Viennois, D. Merlin, and A. Unil Perera, "Minimally invasive screening for colitis using attenuated total internal reflectance fourier transform infrared spectroscopy," *Journal of Biophotonics*, vol. 10, no. 3, pp. 465–472, 2017.

[13] A. G. Theakstone, C. Rinaldi, H. J. Butler et al., "Fourier-transform infrared spectroscopy of biofluids: a practical approach," *Translational Biophotonics*, vol. 3, no. 2, Article ID e202000025, 2021.

[14] A. A. Aziz, V. Selvaratnam, Y. F. B. A. Fikri, M. S. A. Sani, and T. Kamarul, "Diagnosis of osteoarthritis at an early stage via infrared spectroscopy combined chemometrics in human serum: a pilot study," *Processes*, vol. 11, no. 2, p. 404, 2023.

[15] T. C. P. Veettil and B. R. Wood, "A combined near-infrared and mid-infrared spectroscopic approach for the detection and quantification of Glycine in human serum," *Sensors*, vol. 22, no. 12, p. 4528, 2022.

[16] D. L. Woernley, "Infrared absorption curves for normal and neoplastic tissues and related biological substances," *Cancer Research*, vol. 12, no. 7, pp. 516–523, 1952.

[17] Y. Ma, F. Liang, M. Zhu, C. Chen, C. Chen, and X. Lv, "FT-IR combined with PSO-CNN algorithm for rapid screening of cervical tumors," *Photodiagnosis and Photodynamic Therapy*, vol. 39, Article ID 103023, 2022.

[18] Y. Wang, H. Qian, X. Shao et al., "Multimodal convolutional neural networks based on the Raman spectra of serum and clinical features for the early diagnosis of prostate cancer," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 293, Article ID 122426, 2023.

[19] A. Sala, J. M. Cameron, C. A. Jenkins et al., "Liquid biopsy for pancreatic cancer detection using infrared spectroscopy," *Cancers*, vol. 14, no. 13, p. 3048, 2022.

[20] F. Le Naour, C. Sandt, C. Peng et al., "In situ chemical composition analysis of cirrhosis by combining synchrotron fourier transform infrared and synchrotron X-ray fluorescence microspectroscopies on the same tissue section," *Analytical Chemistry*, vol. 84, no. 23, Article ID 10260, 10266 pages, 2012.

[21] H. Xin, D. Wang, X. Qi, G. Qi, and G. Dou, "Structural characteristics of coal functional groups using quantum chemistry for quantification of infrared spectra," *Fuel Processing Technology*, vol. 118, pp. 287–295, 2014.

[22] M. Hlavatsch and B. Mizaikoff, "Advanced mid-infrared lightsources above and beyond lasers and their analytical utility," *Analytical Sciences*, vol. 38, no. 9, pp. 1125–1139, 2022.

[23] D. Finlayson, C. Rinaldi, and M. J. Baker, "Is infrared spectroscopy ready for the clinic?" *Analytical Chemistry*, vol. 91, no. 19, pp. 12117–12128, 2019.

[24] A. S. Ahuja, "The impact of artificial intelligence in medicine on the future role of the physician," *PeerJ*, vol. 7, Article ID e7702, 2019.

[25] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[26] K. Zhu, S. Zhang, K. Yue et al., "Rapid and nondestructive detection of proline in serum using near-infrared spectroscopy and partial least squares," *Journal of analytical methods in chemistry*, vol. 2022, Article ID 4610140, 12 pages, 2022.

[27] H. Yang, S. Yang, J. Kong, A. Dong, and S. Yu, "Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy," *Nature Protocols*, vol. 10, no. 3, pp. 382–396, 2015.

[28] P. Jiang and J. Chen, "Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation," *Neurocomputing*, vol. 198, pp. 40–47, 2016.

[29] U. Zelig, E. Barlev, O. Bar et al., "Early detection of breast cancer using total biochemical analysis of peripheral blood components: a preliminary study," *BMC Cancer*, vol. 15, no. 1, pp. 408–410, 2015.

[30] B. Yang, C. Chen, C. Chen et al., "Detection of breast cancer of various clinical stages based on serum FT-IR spectroscopy combined with multiple algorithms," *Photodiagnosis and Photodynamic Therapy*, vol. 33, Article ID 102199, 2021.

[31] F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad, and N. Nikbakhsh, "Application of FT-IR spectroscopy on breast cancer serum analysis," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 187, pp. 87–91, 2017.

[32] J. Ollesch, D. Theegarten, M. Altmayer et al., "An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma," *Biomedical Spectroscopy and Imaging*, vol. 5, no. 2, pp. 129–144, 2016.

[33] X. Wang, X. Shen, D. Sheng, X. Chen, and X. Liu, "FTIR spectroscopic comparison of serum from lung cancer patients and healthy persons," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 122, pp. 193–197, 2014.

[34] J. Liu, H. Cheng, X. Lv et al., "Use of FT-IR spectroscopy combined with SVM as a screening tool to identify invasive ductal carcinoma in breast cancer," *Optik*, vol. 204, Article ID 164225, 2020.

[35] G. Clemens, B. Bird, M. Weida, J. Rowlette, and M. J. Baker, "Quantum cascade laser-based mid-infrared spectrochemical imaging of tissue and biofluids," *Spectroscopy Europe*, vol. 26, no. 4, pp. 14–19, 2014.

[36] H. Heidari, M. Arabi, and T. Warziniack, "Effects of climate change on natural-caused fire activity in western U.S. National forests," *Atmosphere*, vol. 12, no. 8, p. 981, 2021.

[37] R. J. Hooley, L. Andrejeva, and L. M. Scoutt, "Breast cancer screening and problem solving using mammography, ultrasound, and magnetic resonance imaging," *Ultrasound Quarterly*, vol. 27, no. 1, pp. 23–47, 2011.

[38] E. B. van de Kraats, J. Munćan, and R. N. Tsenkova, "Aquaphotomics—origin, concept, applications and future perspectives," *Substantia*, vol. 3, no. 2, pp. 13–28, 2019.

[39] J. Muncan and R. Tsenkova, "Aquaphotomics—from innovative knowledge to integrative platform in science and technology," *Molecules*, vol. 24, no. 15, p. 2742, 2019.

[40] R. Tsenkova, J. Munćan, B. Pollner, and Z. Kovacs, "Essentials of aquaphotomics and its chemometrics approaches," *Frontiers in Chemistry*, vol. 6, p. 363, 2018.