Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Near-infrared spectroscopy identification method of cashmere and wool fibers based on an optimized wavelength selection algorithm

Yaolin Zhu [a], Long Chen [a,*], Xin Chen [a], Jinni Chen [a], Hongsong Zhang [b]

[a] *School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710000, China*
[b] *Shanghai Ranzi Industrial Co., Ltd, Shanghai, 201800, China*

ARTICLE INFO

Keywords:
Cashmere and wool
Near-infrared spectroscopy
Wavelengths selection
Grouping genetic algorithm
PLS-DA

ABSTRACT

Cashmere and wool fibers have similar chemical compositions, making them difficult to distinguish based on their absorption peaks and band positions in near-infrared spectroscopy. Existing studies commonly use wavelength selection or feature extraction algorithms to obtain significant spectral features, but traditional algorithms often overlook the correlations between wavelengths, resulting in weak adaptability and local optimum issues. To address this problem, this paper proposes a recognition algorithm based on optimal wavelength selection, which can remove redundant information and make the model effective in capturing patterns and key features of the data. The wavelengths are rearranged by computing the information gain ratio for each wavelength. Then, the sorted wavelengths are grouped based on equal density, which ensures that all wavelengths within each group have equal information and avoids over-focusing on individual groups. Meanwhile, the group genetic algorithm is used to find the wavelengths with highly informative and search optimal grouped combinations, in order to explore the entire spectrum wavelength. Finally, combined with a partial least squares discriminant analysis(PLS-DA) model, the recognition accuracy reached 97.3 %. The results indicate that, compared to traditional methods such as CARS, SPA, and GA, our method effectively reduces redundant information, selects fewer but more informative wavelengths, and improves classification accuracy and model adaptability.

1. Introduction

Cashmere is an expensive animal fiber with a scale structure that is mostly ring-shaped or oblique ring-shaped. The scales are neatly arranged [1], and the edge is not obvious. Therefore, cashmere has special properties such as soft texture and delicate luster. It is one of the best quality original materials in textile products worldwide and is known as "fiber gem" and "soft gold." [2]. The wool scales are mostly irregular, in larger tile-like shapes, with a cracked appearance. The scales are disorderly arranged, and the edges are clearly raised. Compared with cashmere fibers, wool is coarser and has a rougher texture. Currently, due to the scarcity of cashmere production, which accounts for only 0.2 % of the overall production of animal fibers in the world [3], some unscrupulous merchants use wool fibers to replace cashmere for huge profits, resulting in the frequent occurrence of cashmere adulterations [4]. China is a major producer and exporter of cashmere and wool, so it is very necessary to accurately and quickly identify the cashmere and wool fibers.

Some traditional identification methods of cashmere and wool fibers include the biology identification method [5], chemistry

[*] Corresponding author.
E-mail address: 18852533177@163.com (L. Chen).

https://doi.org/10.1016/j.heliyon.2024.e34537
Received 4 July 2024; Accepted 11 July 2024
Available online 14 July 2024

identification method [6], and the scanning electron microscopy method [7]. However, these conventional detection methods usually have some drawbacks. For instance, the biology identification method has a slow detection speed and a high cost. It also requires a certain level of professional knowledge from the inspector, which limits its application. The chemistry identification method usually involves sample preparation. It is destructive to samples and has a long detection period, so it is not suitable for rapid detection on some specific occasions. Therefore, there is a pressing need for a fast, accurate, and widely applicable identification method for cashmere and wool fibers.

Near-infrared spectroscopy technology is a recent research achievement that integrates various disciplines such as computer science, spectroscopy, and chemometrics [8]. It has the advantages of fast detection speed, high precision, low cost, and non-destructive sampling, and it is widely used in medical diagnostics, environmental monitoring, agriculture, the food industry, and the textile industry [9]. In recent years, many researchers have applied near-infrared spectroscopy to the identification of cashmere and wool fibers. Chen et al. [10] investigated the feasibility of classifying different types of animal fibers such as wool and cashmere with near-infrared spectroscopy and chemometric models. They used two common chemometric algorithms, PLS-DA and soft Independent modeling of class analogy (SIMCA), to build models with the pre-processed spectral data. The experiments showed that near-infrared spectroscopy, combined with signal processing and PLS-DA or SIMCA, can effectively distinguish between different types of animal fibers. Zhou et al. [11] applied principal component analysis (PCA), SIMCA, and linear discriminant analysis (LDA) to identify seven textile fibers, including cashmere, wool, polypropylene, and cotton. They found that five fibers except wool and cashmere could be well identified. To improve the recognition rate of cashmere and wool, which are highly similar, they selected six characteristic wavelengths and combined them with the LDA classifier. Liu et al. [12] proposed using PCA to reduce the dimensionality of the original spectral data, decompose the spectral principal component data and score data, and calculate the Mahalanobis distance to reflect the differences between samples in three-dimensional space. This approach achieved cluster analysis and discrimination between cashmere and wool fibers. Sun et al. [13] introduced a new identification method that combines near-infrared spectroscopy with representation learning-based classification of cashmere and wool, which are chemically very similar. This method can be directly applied to textile samples rather than just the textile fibers. Moreover, some researchers have integrated near-infrared spectral analysis with deep learning networks. For example, Wu et al. [14] used principal component analysis to reduce the dimensionality of the preprocessed cashmere and fine wool spectral data. They selected the first six principal components based on their reliability and fed them into the three-layer backpropagation artificial neural network. This method, which combines near-infrared spectroscopy with the PCA-BP algorithm, can establish identification models for cashmere and fine wool. Guo et al. [15] proposed a detection technique that combines near-infrared and Elman neural networks. They used a more effective hybrid genetic technique to dynamically optimize the network structure, weight, and self-feedback gain simultaneously. Experiments showed that this method could quickly build a high-precision cashmere and wool detection model. The above methods are effective for the identification of cashmere and wool. However, most of them analyze full-wavelength data and perform feature extraction using linear transformations such as PCA and LDA. However, they are affected by the redundant information in the full wavelengths during the analysis process. Moreover, the feature extraction process transforms the original data into a new low-dimensional space through linear transformations, resulting in only abstract principal component scores. This approach ignores the relationship between the wavelengths and reduces the interpretability of the data. To overcome this limitation, this paper selects the optimal wavelengths from the full wavelength by using a wavelength selection method for the subsequent analysis.

Motivated by the aforementioned work, this paper proposes a spectral feature wavelength selection method based on a combination of information gain ratio sorted grouping and an improved genetic algorithm. This method preserves the relationship between the original data as much as possible while screening the wavelengths to achieve dimensionality reduction. Compared to traditional wavelength selection algorithms, it addresses the issue of getting stuck in local optima and improves the efficiency of wavelength selection. Additionally,it is suitable for nonlinear models and avoids the problem that algorithms such as PCA and LDA are not appropriate for feature extraction of nonlinear data.

2. Methods and materials

2.1. Analysis of spectral

Typically, both cashmere and wool fibers possess a similar keratinous layer structure and are composed of various amino acids. However, cashmere typically contains a higher proportion of cysteine and arginine, while wool contains a higher proportion of serine. These differences directly impact the formation of hydrogen bonds and disulfide bonds between molecular chains. The keratin chains in cashmere have more disulfide bonds, resulting in denser and more compact molecular chains. Consequently, the surface of cashmere fibers appears more uniform and smooth. In contrast, wool has fewer disulfide bonds, leading to a looser molecular structure and a rougher fiber surface.

Table 1
The chemical bond corresponding to the absorption peak of cashmere and wool fibers.

Wavelength/nm	1187	1455	1507	1702	1734	1936	2047	2167	2291
Chemical bond	C–H	C=O N–H C–H	N–H NHCO	C–H	S–H	C=O N–H C–H	C=O N–H	N–H	C–H CH_2

When samples are exposed to external NIR light, the molecules inside undergo vibrational excitation and energy level transitions. Simultaneously, different chemical bonds and functional groups in different molecules exhibit unique reactions to light absorption and scattering. Therefore, the differences in the composition of amino acids in keratin between cashmere and wool fibers result in subtle differences in their near-infrared spectra. This paper lists the absorption peaks of cashmere and wool fibers in the near-infrared spectrum range and their corresponding chemical bonds, as shown in Table 1.

Among them, in the range of 1100∼1300 nm, absorbance is mainly related to the second overtone of CH, CH_2, and CH_3 [16], while CH_2 and CH_3 are not typical protein bonds. The absorption peak around 1455 nm is mainly related to the vibrational absorption of carbonyl groups (C=O) and the first overtone of N–H [17]. The amide group and N–H group exhibit obvious absorption peaks at around 1500 nm. The absorption peak at around 1700 nm is formed by typical C–H chemical bonds in cashmere and wool, as well as disulfide bonds (S–S) and hydrogen sulfide bonds (S–H) [18]. We use these characteristic peaks to distinguish cashmere and wool fibers.

2.2. Overview of the cashmere and wool identification processes

The range of near-infrared spectroscopy is typically between 700 and 2500 nm [19]. The selection of characteristic wavelengths is crucial for performing qualitative and quantitative analysis using near-infrared spectroscopy. This paper proposes a hybrid feature wavelength selection method based on a combined optimization, which can effectively eliminate the interference information and enhance the robustness of the qualitative model. This method combines a filter algorithm based on the information gain ratio with an improved grouping genetic algorithm. First, collect the spectra of cashmere and wool samples, apply spectral preprocessing on the collected spectra to reduce the noise and deviation that may affect the original spectra, and improve the reliability and accuracy of the data. Next, hidden information in the near-infrared spectra is extracted through various spectral transformations, while eliminating features with excessive redundant information through correlation analysis. Then, the feature data are input into a wavelength selection algorithm to extract characteristic wavelength bands from the spectra, reducing data dimensionality and interference information. Finally, three classifiers including PLS-DA [20], Decision Tree (DT) [21], and Support Vector Machine (SVM) [22] were used for classification, the recognition accuracy was compared, and the classifier with the highest recognition accuracy was chosen as the final classification model.

2.3. Spectral preprocessing

The near-infrared spectroscopy of cashmere and wool obtained from the near-infrared spectrometer may be affected by noise and spectral deviation due to the data collection process, external environmental factors, instrumentation, and the inherent characteristics of the samples. This makes it difficult to analyze cashmere and wool fibers using the original spectra directly. Therefore, to purify the spectral information and ensure the stability and reliability of the final results, it is necessary to perform a series of preprocessing operations on the original spectral data. Common preprocessing methods include background subtraction, scatter correction, noise reduction, and scaling, among others [23].

Since the preprocessing process needs to remove the noise influence while retaining the important features of the original spectrum, this experiment adopts Savitzky-Golay convolutional smoothing [24] based on the principle of least squares. It uses polynomials for data smoothing, which helps preserve the useful information in the signal, eliminating random noise, and making the spectrum smoother.

The Standard Normal Variate (SNV) [25] is employed to remove spectral deviation caused by fiber surface scattering and external optical path changes. The transformation formula is shown in Eq. (1).

$$X_{SNV} = \frac{X - \overline{X}}{\sqrt{\frac{\sum_{k=1}^{m}(X_k - \overline{X})^2}{(m-1)}}} \tag{1}$$

where $\overline{X} = \frac{\sum_{k=1}^{m} X_k}{m}$, m represents the number of wavelength points.$k = 1,2, …,m$.

Since the reflectance or transmittance intensity is lower than the background intensity, leading to negative absorbance values, and also due to significant differences in absorbance intensity observed in wool and cashmere samples from different origins in the spectral data, Min-Max normalization [26] is employed to map the original spectral data into the range of [0–1], thus eliminating the dimensional inconsistency. The transformation function is shown as Eq. (2).

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where $X_{\max}$ and $X_{\min}$ are respectively represent the maximum and minimum values of the sample spectral data. Each time a new sample is inputted, it requires to update the maximum and minimum values of the entire sample data to ensure that the normalized data does not exceed the set range.

2.4. Spectral transformation

Spectral transformation can unveil hidden information and features within a spectrum. By transforming the spectrum, the original data can be converted into a new set of parameters or features, facilitating sample classification, quantitative analysis, and more. Methods such as the derivative, reciprocal, and wavelet transform help extract features across different frequencies or wavelengths from the spectrum [27], enabling a deeper and more comprehensive spectral analysis. In this study, we opted for three types of transformation methods as detailed in Table 2. These include mathematical transformations, such as envelope removal, logarithm, square root, and cube root; derivative transformations, including first-order differentiation and second-order differentiation; and wavelet transformations using discrete wavelet transforms [28]. In this experiment, we choose the CoifN (N = 1) wavelet basis function, which has shown effective performance in identifying specific features. It exhibits good performance for analyzing spectral curves at 10 scales (2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10})

Note:Here, A_x indicates the absorbance value corresponding to the X-band. W_y denotes the y-th decomposition scale.

Correlation analysis measures the degree of association between two or more variables that have a certain relationship [29]. There are various types of correlation coefficients, such as the Pearson, Spearman, and Chebyshev coefficients [30]. In this study, we used the Pearson correlation coefficient for correlation analysis.

We treated the data from each spectral transformation as a new feature of the original spectrum. We performed correlation analysis between the spectral data that underwent mathematical, derivative, and discrete wavelet transformations and the original spectral data. The lower the correlation between different features of cashmere and wool, the less redundant information in that feature combination. The analyzed feature data, used for subsequent qualitative analysis, can significantly improve identification accuracy.

2.5. Spectrum feature wavelength selection

Spectral analyses typically involve multiple wavelengths, each reflecting different sample information. However, not all wavelengths are relevant for the analysis task. Feature wavelength selection, a dimensionality reduction technique, can efficiently remove or reduce low-contributing redundant wavelengths, thus lowering computational complexity and increasing the stability and reliability of the analysis.

Wavelength selection algorithms can be classified into three categories: filter algorithm, wrapper algorithm, and hybrid algorithm [31]. The information gain ratio [32] is a classic filter algorithm that computes the information gain ratio of different wavelengths. A higher information gain ratio of a wavelength indicates more information content. Feature subsets are then filtered by setting a threshold. However, this algorithm does not produce a feature subset with high recognition accuracy. Moreover, the threshold setting is somewhat arbitrary. Another researcher suggests using a search algorithm to screen feature wavelengths, such as a genetic algorithm [33]. This kind of method performs better than the filter algorithm, but it has higher computational complexity, which increases the running time. Moreover, with traditional genetic algorithms, when a “superior individual” emerges within the population with much higher fitness than the others, it can cause premature convergence to a local optimum and prevent finding the global optimal solution [34]. The hybrid algorithm proposed in this paper is based on information gain ratio grouping and a population-based genetic algorithm. First, the information gain ratio of each wavelength is calculated, and the original wavelengths are reordered in descending order. Then, the sorted wavelengths are grouped according to the principle of equal density, ensuring that the information content within each group is consistent, and preventing the subsequent selection process from getting stuck in an optimal solution. Finally, an improved genetic algorithm was used to search for characteristic wavelengths in both intraspecific and interspecific, considering the relationships among wavelengths within each group and between different groups, thereby ensuring the algorithm can search the entire wavelength range. Since the information gain ratio is used in the filtering stage of the hybrid algorithm, the C4.5 algorithm [35] is chosen as the evaluation algorithm in the genetic algorithm, as it is based on the information gain ratio.

The hybrid feature wavelength selection algorithm based on information gain ratio sorting grouping and grouping evolutionary genetic algorithm is implemented as follows: Suppose the sample set $S = \{X_1,X_2, \ldots,X_n\}$ each sample X_i contains m categories $F = \{F_1,F_2, \ldots,F_m\}^T$ Suppose the category F_m has N values, then the sample set S is divided into N sample subsets $f_1,f_2, \ldots,f_N$ according to F_m. Therefore, the average information entropy of the sample set S for classification is calculated as

$$H(S) = -\sum_{i=1}^{N} P(f_m)\log_2 P(f_m) \tag{3}$$

Table 2
The methods and expressions of spectral transformation.

Transformation type	Conversion method	Variable formula
Mathematical transformation	Envelope Removal	$C(A_x)$
	Logarithms	$Lg\,A_x$
	Square Root	$\sqrt{A_x}$
	Cubic Root	$\sqrt[3]{A_x}$
Derivative transformation	First-order Differential	A_x'
	Second-order Differential	A_x''
Wavelet transform	Discrete Wavelet Transform	A_x, W_y

where $P(f_m) = |S_i|/|S|(m = 1,2, \ldots,k)$, $|S_i|$ and $|S|$are the number of S_i and S samples, respectively.

Assuming that the characteristic wavelength F_i has t different values, the sample set S can be divided into t sample subsets $S_1,S_2, \ldots, S_t$ based on the value of F_i, so that $f_1,f_2, \ldots,f_N$ are divided into $N \times t$ subsets, each subset f_{nm} indicates that F_i belongs to the collection of samples of the *m-th* class in the case of $i = t$. Then, the average information entropy of the sample set S for classification after the division of the corresponding feature wavelength F_i is:

$$H\left(\frac{S}{F_i}\right) = -\sum_{i=1}^{t} P(f_m)\left[-\sum\nolimits_{p=1}^{N} P(f_{nm})\log_2 P(f_{nm})\right] \tag{4}$$

where $P(f_m) = \sum\nolimits_{P=1}^{N} |f_{nm}|/|S|, P(f_{nm}) = |f_{nm}|/|S|(n = 1,2,\ldots,t)$. According to the amount of sample information and the amount of conditional information after the division of the non-categorical feature wavelength F_i, The amount of information gain $G(S,F_i)$ for the sample set S can be found as

$$G(S,F_i) = H(S) - H(S,F_i) \tag{5}$$

Since the information gain ratio of F_i dividing S is equal to the ratio of the information gain amount to the split information amount, we can get

$$GR(S,F_i) = \frac{G(S,F_i)}{Sp(S,F_i)} \tag{6}$$

Among them, amount of split information $Sp(S,F_i) = -\sum\nolimits_{l=1}^{t}(|S_l|/|S|)\log_2(|S|/|S|)$. The information gain ratios of different characteristic wavelengths F_i are calculated according to the above eq. (3) ~ eq. (6), and sorted according to the calculated ratios to obtain the sorted feature set $GR = \{GR(F_1),GR(F_2), \ldots,GR(F_i)\}$ Group the sorted feature sets according to the principle of the same information density in each group. Suppose the sample set $S = \{S_1,S_2, \ldots,S_N\}$ the feature wavelength is $F = \{F_1,F_2, \ldots,F_i\}$ the *n-th* feature of the sample set S is $F_n(1\leq n \leq i)$, the *m-th* feature is $m < n$. Then the information density of the characteristic wavelength F_n as follows:

$$\rho_n = \frac{GR(F_n)}{\sum\limits_{l=1}^{i} GR(F_n)}, n = 1,2,\ldots i \tag{7}$$

Where $GR(F_n)$ is the information gain ratio of the *n-th* wavelength. Through Eq. (7), the information density ρ_n and ρ_m between the two characteristic wavelengths can be calculated separately then the density distance between the wavelengths can be defined as

$$D = \frac{GR(F_n) - GR(F_m)}{\sum\limits_{l=1}^{i} GR(F_l)}, n,m = 1,2,\ldots,i \tag{8}$$

The goal of feature grouping is to combine components with high density. We aim to achieve a low intergroup density and a high intragroup density, ensuring strong cohesion within each group while maintaining sparse connections between groups. From Eq. (7) and Eq. (8), it can be seen that the information density increases as the information gain ratio increases, and the information density distance between wavelengths becomes smaller, then the higher the probability that two wavelengths will be grouped together in one group.

Before applying the genetic algorithm to the grouped feature groups for further feature wavelength selection, feature encoding needs to be done first. In this experiment, there are only two situations: the wavelength is selected or not selected, so the binary encoding method is chosen in the encoding process, that is, the selected wavelength is defined as *1*, and the unselected wavelength is defined as *0*. In this paper, feature encoding is performed on each feature wavelength in the sorted and grouped feature groups to obtain a binary string of $N \times GR$ where GR is the sorted feature group, where $D_i = \{D_1,D_2, \ldots,D_N\},i = 1,2, \ldots,N$. D_i is the set of the whole individual, and it is regarded as the individual space, then the size of the individual space is defined as

$$C_{n_i}^{1} + C_{n_i}^{2} + \ldots + C_{n_i}^{n_i}, i = 1,2,\ldots,N \tag{9}$$

The population size in the genetic algorithm determines the size of the search space and the search efficiency of the algorithm. Generally speaking, the larger the initial population size, the more search space and efficiency, but an excessively large population size will also increase the computational cost, resulting in a longer running time of the algorithm. In this experiment, a linear multiple of the encoding length is chosen as the initial population size. Genes are randomly generated as the initial population by a random function.

To overcome the drawback of premature convergence in the traditional genetic algorithm, the improved genetic algorithm in this paper adopts the process of group evolution within and outside the population and uses two different evaluation functions to evaluate the fitness of the population.

For the individual $d_j = \{d_1,d_2, \ldots,d_{ni}\}$ in the population, $j = 1,2, \ldots,n_i$, $i = 1,2, \ldots,N$, in the genetic algorithm, each individual the fitness score is searched. The higher the fitness value of the individual is, the higher the probability that the individual will be selected. In this paper, the classification accuracy of the C4.5 algorithm is used as the fitness function of the evaluation of individuals.

$$f(d_j) = \frac{1}{K}\sum_{l=1}^{K} A_l \tag{10}$$

Among them, d_j is the *j-th* individual in the *n-th* group, and A_l represents the classification accuracy of any class in the k-class classification problem.

For individuals outside the population $d_j = \{d_1, d_2, \ldots, d_N\}, i = 1,2, \ldots, N,$ since the goal of feature wavelength selection is to select fewer feature wavelengths, thereby reducing redundant information, the feature groups with more individuals are penalized to some extent, the probability of these individuals being selected is lowered, and the feature groups with fewer individuals get more chances of reproduction, so a C4.5 algorithm is added with a penalty factor C.

$$f(d_j) = \frac{1}{K}\sum_{l=1}^{K} A_l - C\frac{\sum\limits_{j=1}^{n_i}\overline{d_l}}{\sum\limits_{i=1}^{N}\sum_{j=1}^{n_i} d_j}, j = 1,2,\ldots,n; i = 1,2,\ldots,N \tag{11}$$

The genetic algorithm mainly consists of the selection process, crossover process, and mutation process. Common selection methods include tournament selection, ranking selection, and random selection. In this experiment, to preserve the individual with the highest fitness value, that is, the best individual, the roulette selection strategy is adopted.

Suppose the size of the *i-th* individual is n_i, $i = 1,2, \ldots, N$ then $d_j = \{d_1, d_2, \ldots, d_j\}, j = 1,2, \ldots, n; i = 1,2, \ldots, N,$ the fitness of the individual is $f(d_j)$, then the probability of being selected in the roulette is

$$P(d_j) = \frac{f(d_j)}{\sum\limits_{j=1}^{n_i} f(d_j)}, j = 1,2,\ldots,n_i; i = 1,2,\ldots,N \tag{12}$$

The above formula shows that the probability of an individual being selected increases with the increase of the fitness value, and the higher the probability of being selected, also indicates that the individual is optimal, which is then used in the evolution process.

In the subsequent evolution process, crossover operation and mutation operation are also included. In this experiment, single-point crossover and bit mutation strategies are adopted respectively. Randomly select a crossover point, cut the two parent individuals at the crossover point to form two fragments respectively, exchange the two fragments, and obtain two new offspring individuals. Add the two offspring individuals to the next generation population and continue to repeat the above operations. As for the bit mutation operation, it is to randomly select the gene position of the individual that needs to be mutated, and then perform bitwise operations according to the method of changing the binary code *0* to *1* or *1* to *0* to form a new offspring individual.

3. Experimental results and analysis

3.1. Sample and instrument

This experiment selected 18 types of wool and cashmere fiber samples from different regions (such as Yulin in Shaanxi, Chifeng in Inner Mongolia, Ordos in Inner Mongolia, Baishan in Jilin, Qinghe in Hebei, etc.), including 14 types of wool fiber samples and 4 types of cashmere fiber samples. Each wool fiber sample was collected three times, while each cashmere fiber sample was collected ten times. During each collection process, spectra were obtained from both the front and back sides of the fiber samples, and their average spectra. In the end, 126 wool spectra data and 120 cashmere spectra data were obtained, ensuring as much as possible that the dataset maintained equal quantities of fiber data for the two different categories. All fiber samples in this experiment were provided by the Yulin Fiber Quality Monitoring Center. The equipment used was the RZNIR7900 near-infrared spectrometer provided by Shanghai Ranzi Industrial Co., Ltd. The instrument used the B603-AOTF near-infrared spectrometer as the instrument platform, the Windows

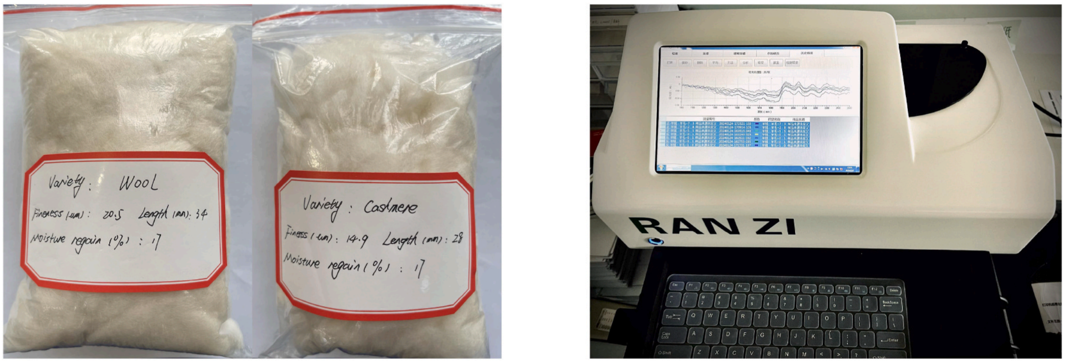

Fig. 1. The images of fiber samples and the near-infrared spectrometer.

system as the operating platform, and the C# programming language for development. It was mainly used to realize the acquisition and quantitative and qualitative analysis of the near-infrared spectral data of the analysis sample. The samples and equipment used in the experiment are shown in Fig. 1.

3.2. Experimental results

To reduce the problem of inaccurate data collection due to external environmental factors such as uneven sample thickness and light leakage caused by the inability to completely cover the lighting sheet during the spectrum collection process, when collecting spectra for each sample, it is required to manually extract portions of fibers multiple times from samples of this kind of fiber and performing a tabletting operation to create a felt with a thickness of approximately 6 mm. In the process of collecting spectral data, the near-infrared light of 1000–2500 nm was used for scanning, and the collection resolution was set to 1 nm, so that each spectral data finally contained 1500 band points. The original spectral data were preprocessed with a combination of Savitzky-Golay convolutional smoothing, standard normal transform, and max-mix normalization, and the preprocessed spectral maps are shown in Fig. 2. (a) represents part of the original spectra collected from cashmere and wool fibers, (b), (c), and (d) correspond to partial spectral data obtained after processing with the three preprocessing methods used in this study, and the final data are shown in Fig. 3 (d). This data will also serve as standard data for subsequent spectral transformations and other operations.

The standardized spectral data obtained from the above processing were subjected to nine different types of spectral transformations to serve as new feature parameters for the spectral data. The correlation analysis between the nine new features and the

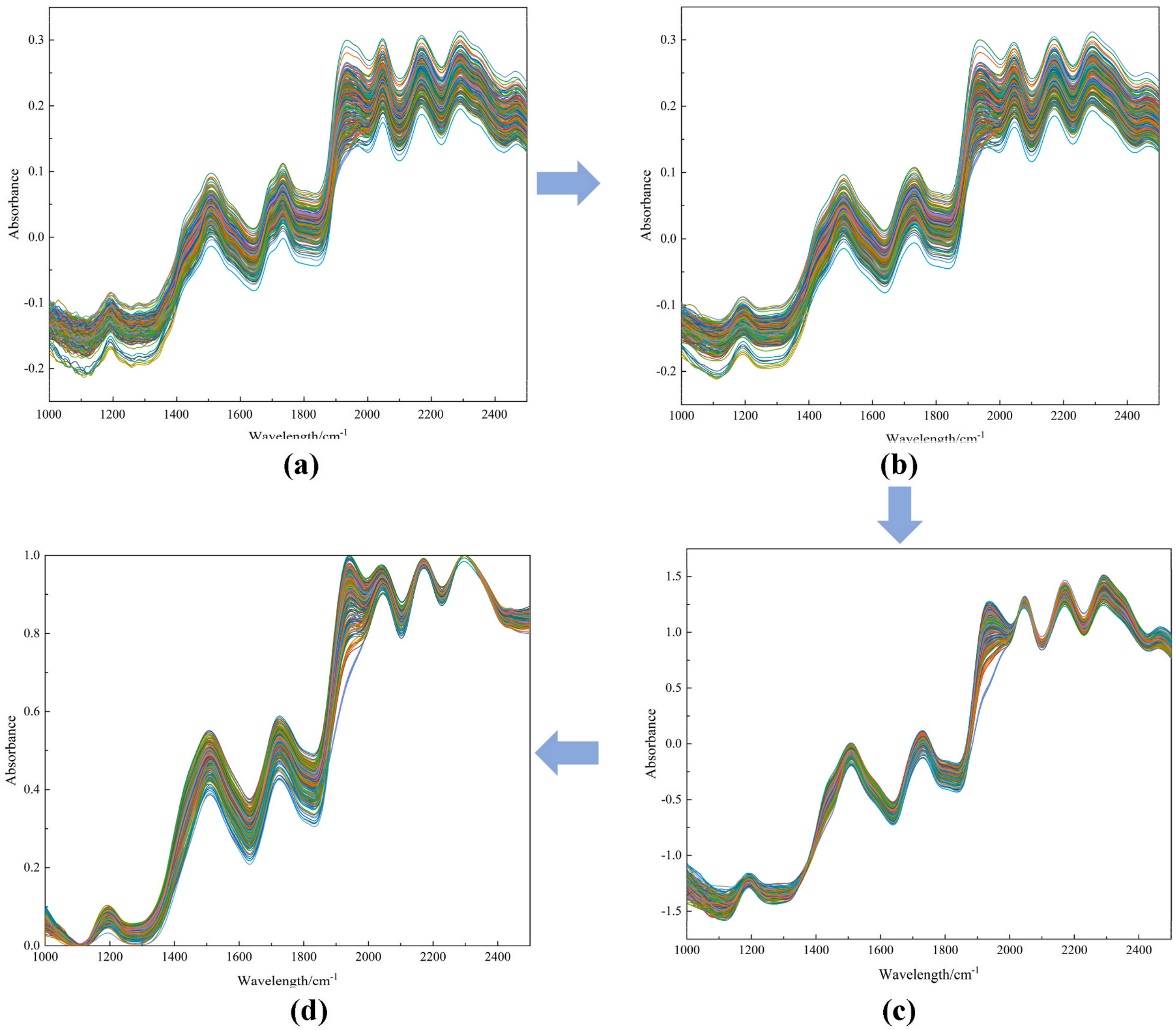

Fig. 2. Spectral preprocessing process: (a) original spectrum, (b) Savitzky-Golay, (c) Standard Normal Variate, (d)Max-Min Normalization.

original spectral data is carried out to ensure that the selected feature combinations contain the least redundant information. As can be seen from Table 3, the correlation between the original data and the data after processing of the first-order differential and second-order differential spectral transforms is extremely low (0.005<r < 0.03). The combination of first-order differential, second-order differential and square root, cube root and wavelet coefficients 3–5 is better, and the correlation coefficient between the spectral data processed by them is only between 0.01 and 0.03. While the square root, cube root and wavelet coefficients 3–5 are strongly correlated with each other, the correlation coefficient is above 0.8, which shows a strong correlation, indicating that the combination of the change contains more redundant information, and it can not be used as a feature combination. Therefore, from the overall analysis, the data processed by first-order differential, second-order differential, and wavelet coefficients 3–5 spectral transform were finally selected as the final spectral feature data.

The filtered feature-combined data was then input to a wavelength selection algorithm for screening. In this experiment, the full wavelength data was divided into 50 groups based on equal density, with each group containing 30 spectral wavelength points. These groups were then input to the grouping genetic algorithm for further search. For first-order differential features, we ultimately obtained 650 feature wavelengths. For second-order differential features, we obtained 900 feature wavelengths. For wavelet scales 3–5, we obtained 750, 650, and 700 feature wavelengths respectively.

The final selection results of wavelength selection for the above five spectral transformations are shown in Fig. 3a, b, c, d, and e. The first-order differential and second-order differential operations can highlight the slope and curvature transforms in the spectrum and enhance the features of the spectrum at the peaks and valleys, so that the wavelength selection process concentrates more on the absorption peak position of the spectrum. For the wavelet transform can simultaneously extract the features in both time and frequency domains, the selected feature wavelengths will be affected by both, so the final selection results are not only concentrated in the absorption peak position, but also more uniformly distributed in the entire wavelength range. The wavelengths that are jointly selected after the five transformations are also labeled, as shown in Fig. 3(f). These repeatedly selected wavelengths were mainly concentrated in the ranges of 1650–1750 nm, 1920–1990 nm, 2030–2100 nm, and 2140–2240 nm, which respectively corresponded to typical chemical bonds in cashmere and wool, such as S–H, $CONH_2$, $CONH_2(H)$, RNH_2, and CHO, demonstrating that the wavelength selection process was guided by the chemical information of cashmere and wool fiber samples, and the precise selection of chemically significant wavelengths helps to improve the accuracy of spectral data analysis.

For cashmere and wool fiber spectral data, combined with the proposed feature extraction and optimized wavelength selection algorithms, experiments were conducted using three classifiers, PLS-DA, DT, and SVM, for training and comparative analyses to find an optimal classifier to construct an efficient classification model. During the entire training process, the dataset was randomly divided into a 70 % training set and a 30 % testing set. The model was trained on the training set and its performance was evaluated on the testing set. The final tested classification performance is shown in Table 4. To avoid the randomness inherent in the random forest model, which might lead to occasional low classification accuracy results lacking persuasiveness, the highest classification accuracy among the ten experimental runs was selected as the final classification result for comparative analysis.

The results show that for the cashmere and wool spectral data features with high correlation and high feature dimension, the recognition accuracy of the PLS-DA model is significantly better than that of the DT and SVM models.

Secondly, to verify that the feature combination screened by correlation analysis in this paper can be used in near-infrared spectrum recognition, we put the screened single feature and feature combination into the PLS-DA model and compared them with the recognition accuracy of the above-mentioned original spectral data. The evaluation indicators are shown in Table 5.

It can be seen from Table 5 that the new spectral features obtained after the three types of spectral transformations and then the new feature combinations formed after the screening of the correlation analysis can obtain better recognition accuracy in the partial least squares discriminant analysis model than any single feature. Combined with Table 4 and Tables 5 and it can be found that the characteristic data obtained after spectral transformation is better than the original data.

Finally, to verify that the wavelength selection algorithm proposed in this paper can screen out the wavelengths with low contribution in the original wavelengths, reduce the amount of calculation, and increase the model's dependability and stability, the original data, the data after the wavelength selection and the feature combination data and the screened feature combination data after the wavelength selection are put into the classifier for verification, and the specific parameters are shown in Table 6.

The results show that the recognition accuracy of the PLS-DA model is improved when the original spectral data passes through the

Table 3
Matrix of coefficients of determination between different spectral transformations of cashmere and wool.

	Original	Envelope	Logarithm	D1	D2	SQR	Cbrt	Wf3	Wf4	Wf5
Original	1									
Envelope	0.16058	1								
Logarithm	0.85997	0.12847	1							
D1	0.01014	−0.22707	0.01342	1						
D2	−0.03806	0.17411	−0.02525	−0.8818	1					
SQR	0.97934	0.14358	0.93564	0.01256	−0.0354	1				
Cbrt	0.95762	0.13792	0.96199	0.01326	−0.0332	0.99573	1			
Wf3	0.91514	0.00383	0.68664	0.01462	−0.0389	0.84945	0.81105	1		
Wf4	0.94621	−0.00679	0.74894	0.01476	−0.0393	0.90128	0.86881	0.97503	1	
Wf5	0.97476	0.07991	0.80882	0.01838	−0.0395	0.94514	0.91841	0.93608	0.97908	1

Note: D1: First-order differential, D2: Second-order differential, SQR: Square root, Cbrt: Cube root, Wf: Wavelet coefficient.

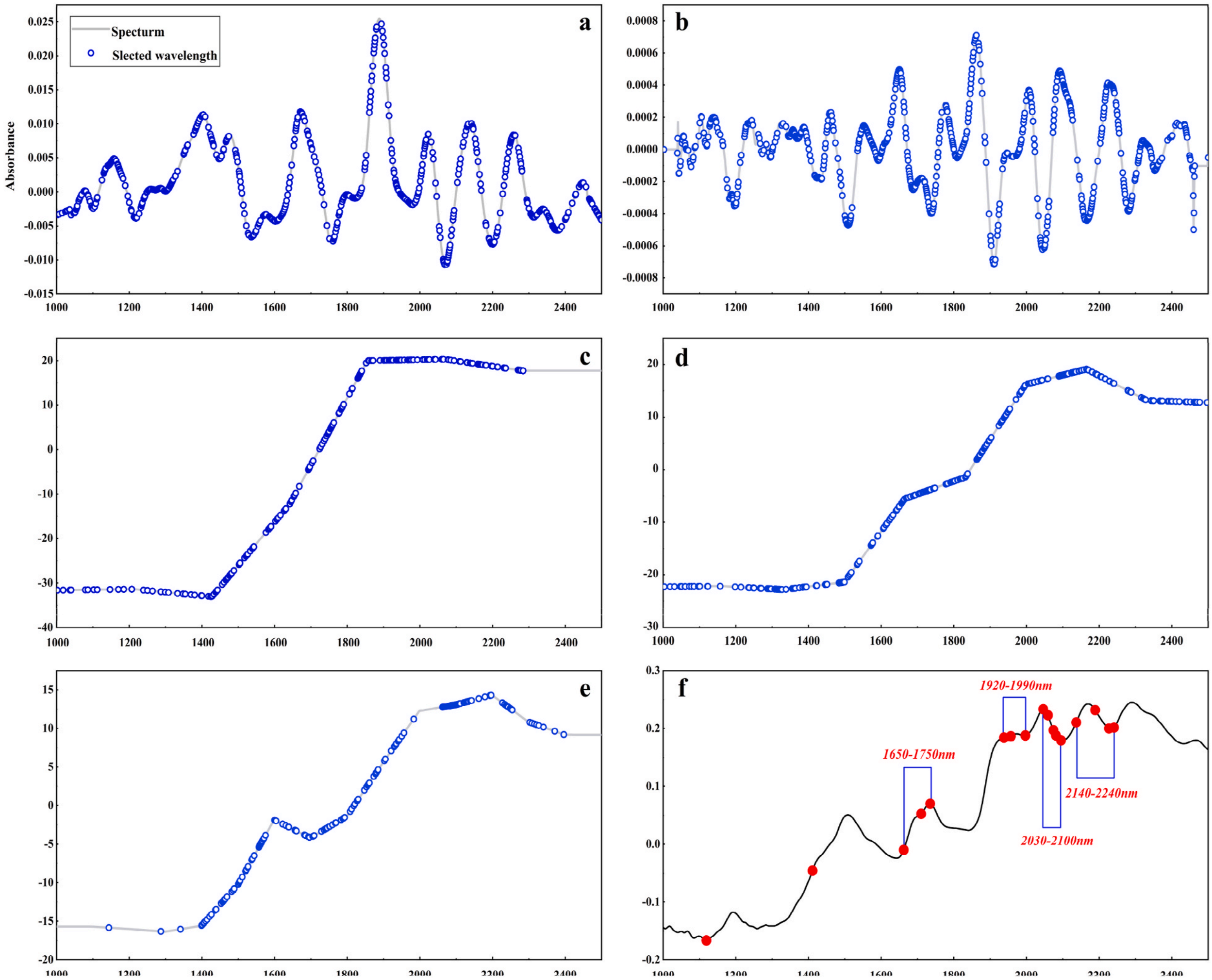

Fig. 3. Extracting the spectral feature wavelengths of cashmere and wool fibers based on the optimized wavelength algorithm.

Table 4
Recognition accuracy of original spectral data on different classifier models.

	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
PLS-DA	91.89	97.37	88.09	92.5
DT	85.14	85.46	85.14	85.16
SVM	86.49	84.21	88.89	86.49

Table 5
The recognition accuracy of single versus combined features in PLS-DA models.

PLS-DA	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
First-order differential	91.83	92.85	92.36	92.60
Second-order differential	90.54	92.68	90.48	91.57
Wavelet coefficients3	78.38	78.05	82.05	80
Wavelet coefficients4	86.49	82.5	91.67	86.84
Wavelet coefficients5	83.78	78.57	91.67	84.62
Multi-feature combinations	95.95	97.44	95	96.2

Table 6
Comparison of recognition accuracy before and after processing by feature wavelength selection algorithm.

		Accuracy(%)	Precision(%)	Recall(%)	F1(%)
original spectral data	No wavelength selection	91.89	97.37	88.09	92.5
	wavelength Selection	93.24	97.30	90	93.51
Combined feature data	No wavelength selection	95.95	97.44	95	96.2
	wavelength Selection	97.3	100	0.95	97.44

wavelength selection algorithm proposed in this paper, thus verifying the effectiveness of the wavelength selection algorithm in this experiment.

To further intuitively analyze the data distribution of the spectral data processed by the method proposed in this experiment, we perform dimensionality reduction on the original high-dimensional data and the processed data by PLS-DA and plot the first three principal components after dimensionality reduction in a three-dimensional space map, as shown in Fig. 4. X, Y, and Z represent the scores of each sample on principal component 1, principal component 2, and principal component 3, respectively. Fig. 4(a) and (b) show the principal component scores of the original spectral data and combined feature data after wavelength filtering, respectively. It can be seen from the figure that the original spectral data of the two categories of cashmere and wool overlap each other in a large area, so the two categories cannot be effectively identified. After being processed by the method proposed in this paper, a clustering phenomenon appeared between the two categories of cashmere and wool, and some abnormal data in the original data that were free on the edge were also screened out.

By randomly permuting the input data of the model and reconstructing the model to calculate its predictive performance, the fit of the model is assessed. The results of the permutation test for partial least squares discriminant analysis are shown in Fig. 5, where the Y-axis represents the frequency of permutations, and the X-axis represents the Q2 scores. The Q2 score of the original model is marked with a red dashed line in the graph, while the bar chart represents the Q2 scores of the permutation models. From the figure, it can be seen that the Q2 score of the original model is closer to 1 and significantly higher than most of the permutation scores, indicating that the model's predictive performance is not due to random chance. This demonstrates that the model has captured the true underlying structure of the data, rather than noise, thereby confirming the high predictive accuracy and stability of the model.

To validate the effectiveness of the proposed algorithm, the improved algorithm was compared with CARS, SPA, and GA wavelength selection algorithms using the original data. Five PLS-DA classification models were constructed and evaluated based on classification accuracy, F1 score, algorithm computation time, and the number of selected wavelengths, as shown in Table 7, where the optimization algorithm proposed in this paper is referred to as IGRS-GA. Higher classification accuracy and F1 scores indicate better performance and generalization ability of the models. Moreover, fewer selected wavelengths and less time consumption imply simpler models and higher efficiency in modeling.

From Table 6 and Tables 7 and it is evident that the classification performance of the final models obtained by the CARS, SPA, and GA algorithms was 90.54 %, 85.14 %, and 86.49 %, all lower than the classification accuracy of 91.87 % achieved using the full spectrum. In contrast, the classification accuracy obtained using the optimized wavelength selection algorithm proposed in this paper has been improved to 93.24 %, demonstrating the advantages of the optimization algorithm in the process of wavelength selection. The wavelength distributions selected by different wavelength selection methods (CARS, GA, SPA, IGRS-GA) were visualized. As shown in Fig. 6, significant differences in wavelength selection are evident across these methods. Traditional wavelength selection algorithms (CARS, SPA, GA) tend to select more wavelength features that evenly cover the entire wavelength range (1000 nm–2500 nm).

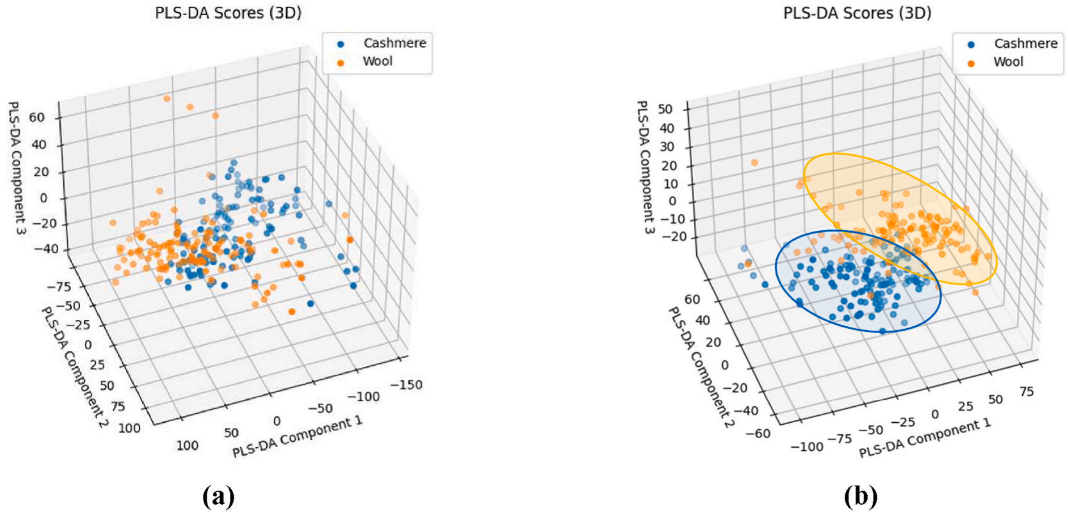

Fig. 4. Three-dimensional spatial maps of the first three principal components after PLS-DA dimensionality reduction.

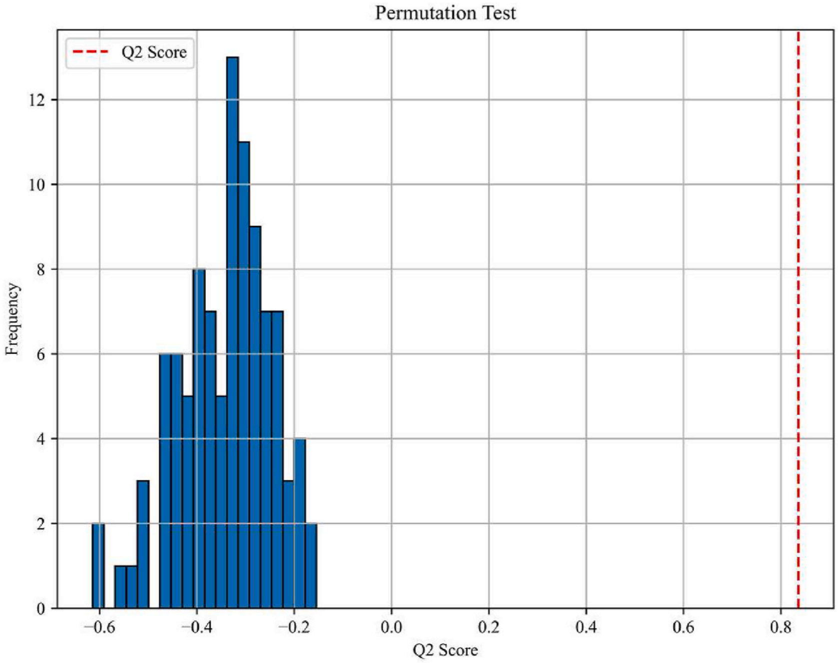

Fig. 5. The results of model permutation test.

Table 7
The performance of different wavelength selection methods on the cashmere and wool dataset.

Method	Accuracy(%)	F1(%)	Variable	T/s
CARS	90.54	88.89	725	14.389
SPA	85.14	82.54	357	8.295
GA	86.49	84.38	700	20.502
IGRS-GA	**93.24**	**93.51**	**340**	**12.403**

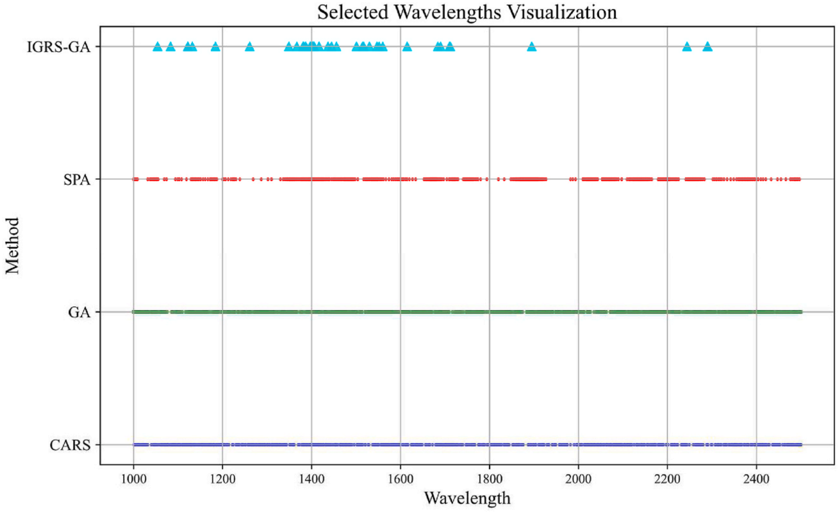

Fig. 6. Visualization of selected wavelengths by different wavelength selection methods.

Although this ensures comprehensiveness of the information, it also brings a large amount of redundant data. The excessive number of wavelengths increases the complexity of data processing, leading to lower computational efficiency. In contrast, the method proposed in this study (IGRS-GA) selects wavelengths more concentrated in the ranges of 1000 nm–1200 nm, 1400 nm–1600 nm, and 2200 nm–2400 nm. By focusing on specific bands, IGRS-GA can better extract key information. Due to the selection of fewer wavelengths, it avoids redundant data, significantly reduces computational resources and time, and improves the accuracy and efficiency of the analysis. This further validates the effectiveness of the algorithm proposed in this paper.

At the same time, this paper also compares the recognition accuracy of two existing cashmere and wool fiber recognition methods, and the experimental results show that the recognition rate of the feature selection algorithm proposed in this paper is relatively high, as shown in Fig. 7. The lower recognition accuracy of Lu et al., Xing et al. and Zhu et al. using image recognition method may be because the feature extraction is not sufficiently adequate, the feature subset does not cover all the features well, and the process of acquiring the fiber images using the micro electron microscope is cumbersome and complicated, and the quality of the acquired images is not high, which affects the accuracy of the model. Leng et al. introduced the deep learning into NIR spectroscopy and added an attention mechanism to improve the performance of the model, but it may be affected by the number of training datasets as well as the characteristics of high dimensionality and complexity of the spectral data, which results in the model not being able to learn effective features and patterns, leading to a lower than expected performance of the model. Quispe et al. combined FTIR spectroscopy with a machine learning model to find a reliable method for animal fiber recognition, but their program did not perform any research such as feature selection or feature extraction, resulting in a final model that may be interfered by redundant features and unable to learn key features, leading to a lower classification accuracy. Therefore, in this paper, the method of using near-infrared spectroscopy combined with an optimized wavelength selection algorithm is simple and feasible, and can better extract key features to achieve efficient and accurate identification of cashmere and wool fibers.

4. Conclusion

This paper proposes a near-infrared spectroscopy identification method for cashmere, wool, and modified wool fibers, employing an optimized spectrum feature wavelength selection algorithm. Initially, the method combines three preprocessing techniques to address issues in the original spectral data, purifying the information and yielding spectral images suitable for feature and wavelength selection. Next, nine spectral transformation methods are utilized to extract hidden information, and Pearson correlation analysis selects features with strong correlation. After screening, five low correlation features are retained for combination, effectively reducing data dimensionality. Information gain ratio ranks wavelengths, which are then grouped by equal density principles. A group genetic algorithm evolves both intra and inter-group parts to search for optimal feature wavelengths. Finally, different classification models train and classify feature subsets containing these wavelengths. Results indicate the PLS-DA model achieves the highest accuracy at 97.3 %. To validate effectiveness, classification results of original spectral data are compared with transformed, combined, and screened data. The method successfully reduces redundant information, improving computation speed and recognition accuracy. Furthermore, comparisons with existing fiber recognition methods based on computer vision and spectral analysis reveal relatively high recognition rates and model stability without overfitting. In the future, we will collect more cashmere and wool fibers from different origins to increase the diversity of samples and improve the applicability of the model. We will also consider combining spectral analysis with deep learning to find a more accurate and efficient fiber recognition model.

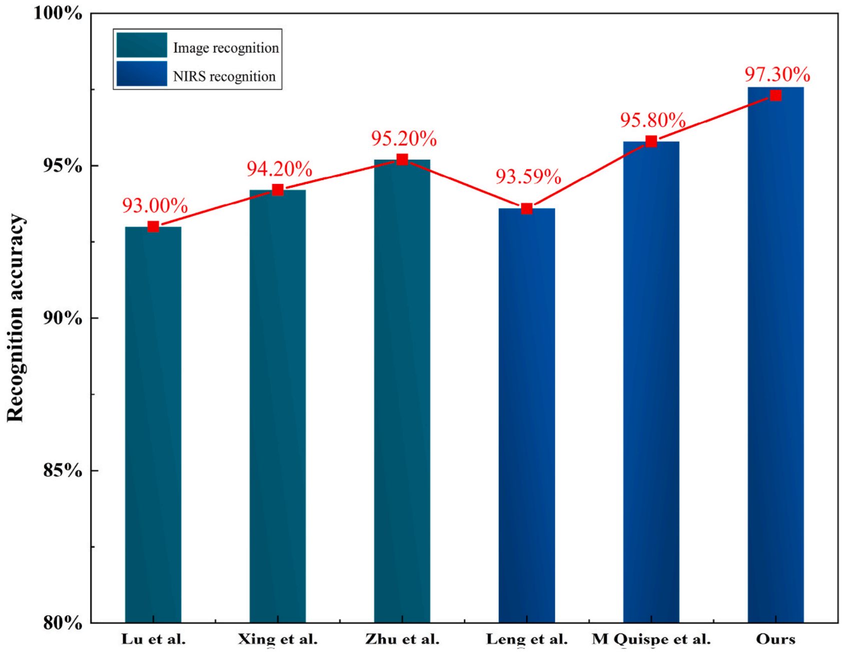

Fig. 7. Comparisons of different benchmark algorithms with image recognition and NIR recognition.

Funding

This work was supported by the natural science basic research key program funded by Shaanxi Provincial Science and Technology Department [grant number 2022JZ-35 and grant number 2023-JC-ZD-33], Science and Technology plan project of Xi'an City [grant number 23DCYJSGG0008-2023], The Yulin Science and Technology Bureau project [grant number YLKG-2023-04], The National Natural Science Foundation of China (General Program), the Chinese Postdoctoral Science Foundation [grant number 2020M683562], the Natural Science Foundation of Shaanxi Province of China [grant number 2022JM-331], the Key Research and Development Program of Shaanxi Provincial Science and Technology [grant number 2023-YBGY-142], the Scientific Research Project of Shaanxi Provincial Education Department [grant number 23JC031], and the Science and Technology Program of Xi'an, China [grant number 23GXFW0027].

CRediT authorship contribution statement

Yaolin Zhu: Conceptualization, Methodology, Software. **Long Chen:** Data curation, Formal analysis, Investigation, Writing – original draft. **Xin Chen:** Project administration, Writing – review & editing. **Jinni Chen:** Supervision, Visualization. **Hongsong Zhang:** Funding acquisition, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the Yulin Fiber Quality Monitoring Center for providing the fiber samples as well as Shanghai Ranzi Industrial Co., Ltd. for providing the near-infrared spectrometer.

References

[1] W.X. Ma, Y.L. Wu, C. Pu, Sericin grafting onto cashmere fiber and its properties investigation, J. Nat. Fibers 19 (6) (2022) 2176–2188, https://doi.org/10.3969/j.issn.1673-1921.2014.02.030. Liu Xianming. Five factors affecting the quality of cashmere and wool[J]. Breeding Technical Consultant, 2014(2):28-28.

[2] M. Jin, J. Lu, X. Fei, et al., Genetic signatures of selection for cashmere traits in Chinese goats[J], Animals 10 (10) (2020) 102768, 1905.Xing W, Liu Y, Deng N, et al. Automatic identification of cashmere and wool fibers based on the morphological features analysis[J]. Micron, 2020, 128.

[3] Franck, Robert R., ed., Silk, mohair, cashmere and other luxury fibres. Elsevier, 2001.

[4] C. Tan, H. Chen, Z. Lin, et al., Category identification of textile fibers based on near-infrared spectroscopy combined with data description algorithms, Vib. Spectrosc. 100 (2019) 71–78.

[5] Y. Zhu, H. JiaYI, Y. Li, et al., Image identification of cashmere and wool fibers based on the improved Xception network, Journal of King Saud University-Computer and Information Sciences 34 (10) (2022) 9301–9310.

[6] B.A. McGregor, Physical, Chemical, and Tensile Properties of Cashmere, Mohair, Alpaca, and Other Rare Animal fibers[M]//Handbook of Properties of Textile and Technical Fibres, Woodhead Publishing, 2018, pp. 105–136.

[7] F.J. Wortmann, W. Arns, Quantitative fiber mixture analysis by scanning electron microscopy: Part I: blends of mohair and cashmere with sheep's wool, Textil. Res. J. 56 (7) (1986) 442–446.

[8] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives–A review, Anal. Chim. Acta 1026 (2018) 8–36.

[9] X. Chu, Y. Huang, Y.H. Yun, et al., Chemometric Methods in Analytical Spectroscopy technology[M], Springer, Singapore, 2022.

[10] Hui Chen, Zan Lin, Chao Tan, Classification of Different Animal Fibers by Near-Infrared Spectroscopy and Chemometric Models, Microc, 2018, https://doi.org/10.1016/j.microc.2018.10.011.

[11] J. Zhou, L. Yu, Q. Ding, et al., Textile fiber identification using near-infrared spectroscopy and pattern recognition, Autex Res. J. 19 (2) (2019) 201–209.

[12] X.R. Liu, L.P. Zhang, J.F. Wang, et al., Use of visible and near infrared reflectance spectroscopy to identify the cashmere and wool, Spectrosc. Spectr. Anal. 33 (8) (2013) 2092–2095.

[13] X. Sun, H. Yuan, C. Song, et al., A novel drying-free identification method of cashmere textiles by NIR spectroscopy combined with an adaptive representation learning classification method, Microchem. J. 149 (2019) 104018.

[14] Guifang Wu, Dengsheng Zhu, H.E. Yong, Spectrosc. Spectr. Anal. 28 (6) (2008) 4, https://doi.org/10.3964/j.issn.1000-0593.2008.06.014.

[15] F. Guo, S. Shang, M. Qi, Study of detection method of cashmere and wool based on near-infrared spectroscopy and Elman neural network[C]. Artificial Intelligence and Computational Intelligence: Third International Conference, AICI 2011, Taiyuan, China, September 24 -25, 2011, Proceedings, Part III 3, Springer Berlin Heidelberg, 2011, pp. 362–369.

[16] R. Iwamoto, A. Nara, T. Matsuda, Near-infrared combination and overtone bands of the CH 2 sequence in CH 2 X 2, CH 2 XCHX 2, and CH 3 (CH 2) 5 CH 3 and their characteristic frequency zones, Appl. Spectrosc. 60 (4) (2006) 450–458.

[17] F. Bangelesa, E. Adam, J. Knight, et al., Predicting soil organic carbon content using hyperspectral remote sensing in a degraded mountain landscape in Lesotho, Applied and environmental soil science 2020 (2020) 1–11.

[18] J. Zhou, R. Wang, X. Wu, et al., Fiber-content measurement of wool–cashmere blends using near-infrared spectroscopy, Appl. Spectrosc. 71 (10) (2017) 2367–2376.

[19] R.S. Gupta, R.R. Deshmukh, A.V. Kshirsagar, Analysis of pharmaceutical solid dosage forms using machine learning and chemometrics techniques, Information Systems and Emerging Technologies (2022) 583.

[20] L.C. Lee, C.Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps, Analyst 143 (15) (2018) 3526–3539.

[21] Y.Y. Song, L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai archives of psychiatry 27 (2) (2015) 130.

[22] S. Suthaharan, S. Suthaharan, Support Vector machine[J]. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 2016, pp. 207–235.

[23] P. Mishra, A. Biancolillo, J.M. Roger, et al., New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC, Trends Anal. Chem. 132 (2020) 116045.

[24] H. Chen, Q. Song, G. Tang, et al., The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models, Int. Sch. Res. Notices 2013 (2013).
[25] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (5) (1989) 772–777.
[26] X. Jintao, Y. Liming, L. Yufei, et al., Noninvasive and fast measurement of blood glucose in vivo by near-infrared (NIR) spectroscopy, Spectrochim. Acta Mol. Biomol. Spectrosc. 179 (2017) 250–254.
[27] L. Shen, M. Gao, J. Yan, et al., Hyperspectral estimation of soil organic matter content using different spectral preprocessing techniques and PLSR method, Rem. Sens. 12 (7) (2020) 1206.
[28] G. Wang, W. Wang, Q. Fang, et al., The application of discrete wavelet transform with improved partial least-squares method for the estimation of soil properties with visible and near-infrared Spectral Data, Rem. Sens. 10 (6) (2018) 867.
[29] S.C. Kafle, Correlation and regression analysis using SPSS, Management, Technology & Social Sciences (2019) 126.
[30] J.M. Droz, Fast computation of p-values for the permutation test based on Pearson's correlation coefficient and other statistical tests, arXiv preprint arXiv: 1807.10046 (2018).
[31] K. Wan, J. Wang, B. Li, et al., Object feature selection under high-dimension and few-shot data based on three-way decision, Vis. Comput. 39 (6) (2023) 2261–2275.
[32] H. Uğuz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowl. Base Syst. 24 (7) (2011) 1024–1032.
[33] X. Chen, Z. Dong, J. Liu, et al., Hyperspectral characteristics and quantitative analysis of leaf chlorophyll by reflectance spectroscopy based on a genetic algorithm in combination with partial least squares regression, Spectrochim. Acta Mol. Biomol. Spectrosc. 243 (2020) 118786.
[34] J. Andre, P. Siarry, T. Dognon, An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization, Advances in engineering software 32 (1) (2001) 49–60.
[35] Z. Xiaoliang, Y. Hongcan, W. Jian, et al., Research and application of the improved algorithm C4. 5 on decision tree[C]//2009 International Conference on Test and Measurement, IEEE 2 (2009) 184–187.