Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 1, 2025

Rapid Textile Fibers Classification Technology Based on Near-Infrared Spectroscopy and Machine Learning

Xingyan Ruan [1], Li Luo [1, *], Xiangyang Yu [2], Yefan Cai [3], Weibin Hong [3]

[1] School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China
[2] School of Physics, State Key Laboratory of Optoelectronic Materials and Technology, Sun Yat-sen University, Guangzhou, Guangdong, 510275, China
[3] Guangzhou Guangxin Technology Co., Ltd., Guangzhou, Guangdong, 510300, China

*** Corresponding author:** Li Luo

Abstract: The accurate identification of textile materials plays a crucial role in production control, quality inspection, and market regulation. In recent years, near-infrared (NIR) spectroscopy has been widely applied in the identification and classification of textile fibers due to its rapid, non-destructive, and efficient characteristics. Therefore, this study utilizes a handheld near-infrared spectrometer and chemometrics to provide a convenient, efficient, non-destructive, and green analytical method for the qualitative analysis of textiles. The article collects spectral data of fiber samples using a handheld NIR spectrometer and analyzes them with chemometric methods. Data preprocessing techniques, feature extraction algorithms, comprehensive sampling methods, and classification algorithms are employed to establish classification models. After ten-fold cross-validation, the optimal SG_RF model is obtained, with a classification accuracy of 93.7% on the test set and an AUC_macro of 0.928, demonstrating excellent performance. The article validates the feasibility of using near-infrared spectroscopy combined with chemometric methods for textile fiber classification, offering a rapid, non-destructive, and efficient approach for textile fiber classification.

Keywords: Near-infrared Spectroscopy; Machine Learning; Fiber Classification; Accuracy; AUC_macro.

1. Introduction

The accurate identification of textile materials is a key process in production control, quality inspection, and market regulation within the textile industry. It plays a significant role in improving product performance, standardizing market order, and protecting consumer rights. The material of textiles directly influences their physical properties, processing methods, and market value, thus there is a broad demand for the rapid and accurate identification of fiber types and compositions. Traditional fiber identification methods, such as microscopic observation, combustion, and chemical dissolution methods[1], although offering certain advantages in terms of accuracy, are complex to operate, time-consuming, and often involve destructive testing or the use of chemical reagents. These drawbacks hinder large-scale testing and are not in line with environmental protection requirements.

Near-Infrared Spectroscopy (NIR) provides an innovative solution for the identification of textile materials due to its rapid, non-destructive, and efficient characteristics. In 2009, Jinzhou Cai[2] used NIR spectroscopy to perform qualitative clustering analysis of 75 samples, including pure cotton, pure polyester, cotton/polyester, and cotton/nylon blends. Principal component analysis (PCA) was applied to extract characteristic spectra, and Mahalanobis distance was used for clustering analysis, yielding excellent classification results. This study verified the feasibility of using NIR spectroscopy for qualitative analysis of textiles. In the same year, Danhong Wang[3] employed discriminant analysis, PCA, and Mahalanobis distance to rapidly identify fibers such as Tencel, cotton, viscose, and copper-ammonium. The results demonstrated that this method provided a reliable and simple means for classifying these fibers. In 2016, Jianhua Fu[4] analyzed 892 samples, including polyester, linen, cotton, silk, wool, polyester-cotton blends, and polyester-wool blends, and explored the application of short-wave and medium-wave NIR spectroscopy in textile composition detection. The study examined the impact of Savitzky-Golay (SG) smoothing on model performance and established a model for textile composition identification and content detection based on medium-wave NIR spectra. Additionally, the Sparse Principal Component Analysis (SPCA) method was proposed to overcome the impact of sample color on the model. A textile composition analysis and identification model based on short-wave NIR spectra was also developed. In 2022, Abdulrahman Aljannahi[5] and others used Fourier Transform Infrared (FT-IR) spectroscopy to classify 138 types of synthetic textile fibers. The study analyzed preprocessing techniques, including SG first derivative and Standard Normal Variate (SNV) methods, and established Principal Component Analysis (PCA) to observe unique patterns and cluster samples. A wealth of research has shown that NIR spectroscopy is a feasible method for rapid, non-destructive identification of fibers.

In recent years, with the miniaturization and intelligence of spectroscopic instruments, as well as advancements in data analysis algorithms, the application of NIR spectroscopy has gradually expanded from laboratory settings to field testing. Portable NIR spectrometers offer advantages such as ease of use, fast analysis, non-destructiveness, minimal sample preparation, multi-component simultaneous analysis, and low cost[6]. They have demonstrated excellent performance in real-time online analysis[7] and have been widely applied both domestically and internationally[8]. Moreover, the integration of artificial intelligence and big data analysis provides strong support for processing and interpreting complex spectral data, further enhancing the accuracy and applicability of the detection.

This study aims to use a handheld NIR spectrometer to collect near-infrared spectral data from fiber samples, combined with chemometric methods, to classify fiber types. The NIR spectral data of the samples are divided into training and test sets, and a classification model is built using SMOTETomek data oversampling, data preprocessing algorithms, feature processing algorithms, and classification algorithms. Ten-fold cross-validation is applied to the training set, and the optimal model is selected based on evaluation metrics. The selected model is then applied to the test set to obtain the evaluation metrics for the test set. The evaluation metrics used are classification accuracy and AUC_macro under multi-class conditions. In the fiber classification model SMOTETomek_SG_RF, the classification accuracy for the test set is 93.7%, and the AUC_macro is 0.928. Additionally, the impact of feature processing algorithms such as PCA and CARS on model performance was also evaluated. Ultimately, the research aims to provide a rapid, non-destructive method for textile fiber classification, thereby improving the efficiency and value of fiber identification.

2. Experiments and Methods

2.1. Experimental Samples

Fibers are defined as slender, flexible materials with a certain length and strength, and they can be classified based on their origin and properties[9]. According to the source of the fiber, samples can be divided into synthetic fibers and natural fibers. To establish a general methodology, this experiment uses polyester, linen-silk, cotton, leather, and wool as fiber samples for classification. These five types of fibers represent common fiber types in the market, covering natural fibers, chemical fibers, and animal leather. Each of these fibers has unique physicochemical properties, which reflect the typical characteristics of different fiber types. The experimental samples were provided by a well-known domestic textile company, consisting of 869 samples in total, including 348 polyester, 102 linen-silk, 226 cotton, 100 leather, and 93 wool samples. Each sample is a square fabric with approximately 15 cm sides, containing a variety of styles, patterns, colors, and structures. Figure 1 shows a sample example.

Fig 1. Example of a fiber sample

2.2. Experimental Setup and Data Acquisition

We used a handheld long-wave near-infrared reflection spectrometer (GX-S1325, Guangzhou Guangxin Technology Co., Ltd., China) to collect spectral data. The wavelength range was 1350-2500 nm, with a resolution of 16 nm. The sample coverage area was 10 mm, and the typical signal-to-noise ratio was 2000:1. The spectrometer was placed flat on a table, and the fabric sample was placed over the detector opening of the spectrometer, followed by a standard reflection plate to shield ambient light. To reduce random noise caused by instrument and temperature variations during spectral measurements, each sample position was measured three times, and the average of the three spectra was taken as the spectral data for that position. Each sample yielded one near-infrared spectrum, resulting in a total of 869 spectral data points. Figure 2 shows a schematic of the near-infrared spectral collection setup.

Fig 2. Near infrared spectroscopy data acquisition device

2.3. Data Analysis Methods

The data analysis algorithm was implemented under Windows 11 operating system using Visual Studio Code software and the programming language was Python 3.9.

2.3.1. Data Balance Processing--SMOTETomek Comprehensive Sampling

In the 869 samples tested, there were 348 polyester samples, 102 linen-silk samples, 226 cotton samples, 100 leather samples, and 93 wool samples, resulting in an imbalanced class distribution. To eliminate the impact of this imbalance, we used the SMOTETomek combined sampling method. This method combines SMOTE (Synthetic Minority Over-sampling Technique) oversampling and TomekLink undersampling to obtain a balanced training dataset, thereby improving the performance of the classification model[10]. The core idea is to first use the SMOTE oversampling method to increase the sample data, and then apply the TomekLink method to remove data points that are in a borderline state, thereby eliminating noise and boundaries.[11]

2.3.2. Data Preprocessing Algorithms

Near-infrared spectral data are often affected by noise, baseline drift, sample surface unevenness, stray light interference, and other factors. Therefore, data preprocessing is necessary to obtain more accurate spectral data. Commonly used spectral preprocessing methods include Savitzky-Golay (SG) smoothing, Multivariate Scatter Correction (MSC), Standard Normal Variate (SNV) transformation, first-order differentiation (FD), and their combined use. The SG algorithm is typically used to improve the signal-to-noise ratio of the spectrum[12]. MSC and SNV methods can eliminate overlay effects and multiplicative effects in the spectrum, and they are widely used scattering correction techniques in near-infrared spectroscopy[13]. The FD method can remove constant baseline drift[14].

2.3.3. Feature Processing Algorithm

Dimensionality reduction of the original data and feature band extraction can reduce the data processing workload and improve analysis efficiency while ensuring good analytical results. Feature processing algorithms are primarily divided into two categories: feature selection and feature extraction. Feature selection directly selects from the original variables to obtain the feature variables, while feature extraction involves rearranging and recombining the original variables to generate new feature variables. In this study, the CARS [15]algorithm from feature selection and the PCA[16] algorithm from feature extraction are compared.

2.3.4. Classification Algorithms

Support Vector Machine (SVM)[17], k-Nearest Neighbors (KNN)[18], and Random Forest (RF)[19] are commonly used classification algorithms. SVM can achieve satisfactory classification results by selecting appropriate kernel functions and parameters. The KNN method is often used to solve supervised classification problems; its principle is simple, and the results are closely related to the choice of the parameter k. RF is a classification method that trains samples using multiple decision trees and makes predictions. It introduces a mechanism of randomly selecting samples and features, making RF less prone to overfitting and providing strong noise resistance.

2.4. Model Evaluation

The confusion matrix[20] is commonly used for visualizing the results of a classification model. In a confusion matrix, each row represents the true class of the data, while each column represents the predicted class. The color blocks on the diagonal of the matrix indicate correct classifications, where the predicted class matches the true class. Accuracy[21] is one of the most commonly used evaluation metrics in classification models, and it represents the proportion of correctly classified samples to the total number of samples. The formula for calculating accuracy is as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{all\_data}$$

Where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives. When the class distribution of sample data is imbalanced, accuracy may not fully assess the model's performance. It is possible to have a high accuracy, but still misclassify all samples of the minority class. Therefore, other evaluation metrics should be considered in combination.

In imbalanced classification problems, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is a widely used method for evaluating the performance of classification models, as it is not affected by the distribution of positive and negative samples[22]. The ROC curve plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis, providing a visual representation of the model's performance at different thresholds. The closer the AUC value is to 1, the better the classification model's performance.

3. Results and Discussion

3.1. Spectral Features

We plot the data exported from the spectrometer as raw spectra. To eliminate the absolute intensity differences and highlight the relative variations in the data, we choose to normalize the raw spectra using the maximum value normalization method. Figure 3 shows the spectral diagram obtained after normalizing all spectral data. Figure 4 shows the average spectrum for each fiber type.

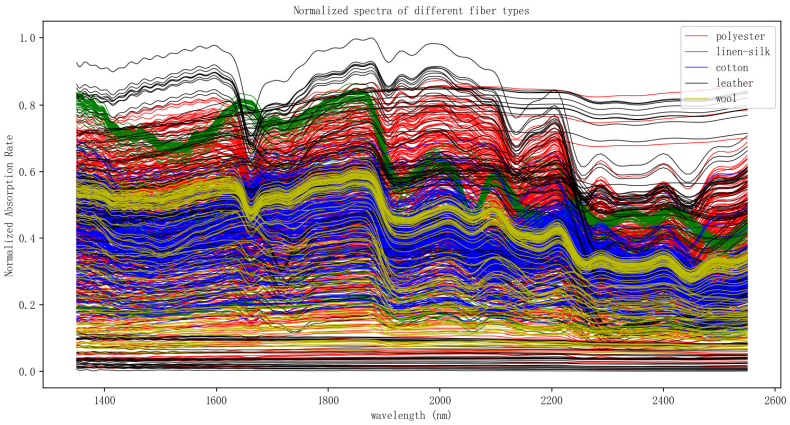

Fig 3. Normalized spectra of different fiber types

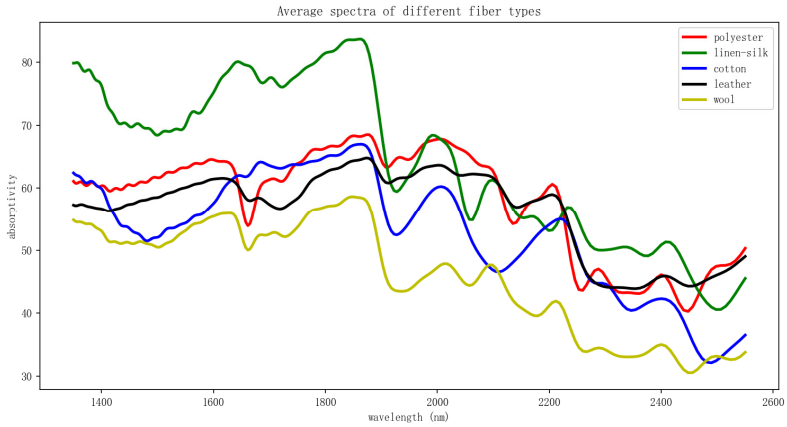

Fig 4. Average spectra of different fiber types

From the raw spectral curves, it can be observed that the baseline drift in the fiber sample spectra is quite significant, and the spectral fluctuations for samples with the same composition are large. Although these issues increase the difficulty of sample classification, properly handling spectral drift and fluctuations can enhance the robustness of the classification model, reduce noise interference, and uncover potential information, thereby improving the scientific validity and accuracy of the classification. From the average spectral curves, it is evident that the spectra of different fiber samples show variability, which lays the foundation for the classification model.

3.2. 10-fold Cross-validation

To evaluate the performance of different data preprocessing algorithms (SG, MSC, SNV, SGFD, MSCFD, SNVFD) and classification algorithms (SVM, KNN, RF), the 869 spectral data samples of the five categories were randomly split into training and test sets, with a training-to-test ratio of 8:2 for each category. To address the issue of class imbalance, the SMOTETomek combined sampling method was used. After applying the method, the number of samples for each fiber type became: 342 polyester, 348 linen-silk, 344 cotton, 346 leather, and 348 wools. The results of 10-fold cross-validation with and without sampling, before and after data preprocessing, were compared. The specific results are shown in Tables 1, 2, and 3. The classification algorithms were experimentally tuned to obtain the optimal parameter settings.

Generally, traditional AUC calculation methods are applicable to binary classification problems. However, in the fiber classification analysis experiment, there are five categories. To handle this multi-class scenario, the following method was adopted: each category was treated as a separate positive class, while the other categories were treated as negative classes, and the AUC for each category was calculated. This results in five ROC curves, and the average of these curves is used to obtain the final ROC curve. The area under this final ROC curve is the AUC_macro, which is used as an indicator to measure the performance of multi-class models. In different sampling methods and data processing scenarios, using the SG preprocessing algorithm and RF classification algorithm, the SMOTETomek combined sampling method before data preprocessing achieved the highest 10-fold cross-validation average AUC_macro of 0.9885, along with a 91.37% average 10-fold cross-validation accuracy.

Table 1. The 10-fold cross-validation results without sampling

Preprocessing Algorithm	*Classification Algorithms*	*10-fold cross-validation-average accuracy (%)*	*10-fold cross-validation-average AUC_macro*
none	SVM	72.09	0.9289
	KNN	75.26	0.9063
	RF	79.86	0.9300
SG	SVM	80.86	0.9679
	KNN	86.76	0.9724
	RF	90.79	0.9821
MSC	SVM	62.16	0.9246
	KNN	84.17	0.9580
	RF	62.88	0.9053
SNV	SVM	73.81	0.9336
	KNN	75.40	0.9000
	RF	77.76	0.9389
SGFD FD	SVM	79.71	0.9656
	KNN	85.61	0.9685
	RF	90.50	0.9814
MSCFD FD	SVM	59.71	0.9328
	KNN	84.31	0.9517
	RF	90.21	0.9847
SNVFD FDFD	SVM	83.17	0.9731
	KNN	85.61	0.9688
	RF	88.49	0.9797

Table 2. 10-fold Cross-validation Results of SMOTETomek Combined Sampling Before Data Preprocessing

Preprocessing Algorithm	*Classification Algorithms*	*10-fold cross-validation-average accuracy (%)*	*10-fold cross-validation-average AUC_macro*
none	SVM	72.09	0.9272
	KNN	75.26	0.9063
	RF	79.86	0.9335
SG	SVM	80.86	0.9684
	KNN	86.76	0.9725
	RF	91.37	0.9885
MSC	SVM	62.16	0.9222
	KNN	84.17	0.9580
	RF	66.47	0.9196
SNV	SVM	73.81	0.9336
	KNN	75.40	0.9000
	RF	77.27	0.9377
SGFD FD	SVM	79.71	0.9653
	KNN	85.61	0.9684
	RF	90.35	0.9833
MSCFD FD	SVM	59.71	0.9338
	KNN	84.31	0.9518
	RF	89.79	0.9862
SNVFD FDFD	SVM	83.16	0.9734
	KNN	85.61	0.9688
	RF	88.78	0.9789

Table 3. 10-fold Cross-validation Results of SMOTETomek Combined Sampling After Data Preprocessing

Preprocessing Algorithm	*Classification Algorithms*	*10-fold cross-validation-average accuracy (%)*	*10-fold cross-validation-average AUC_macro*
none	SVM	76.26	0.9371
	KNN	74.68	0.9064
	RF	79.28	0.9374
SG	SVM	84.90	0.9684
	KNN	87.20	0.9667
	RF	91.08	0.9835
MSC	SVM	78.85	0.9394
	KNN	85.61	0.9535
	RF	64.33	0.9017
SNV	SVM	76.55	0.9380
	KNN	76.26	0.9074
	RF	78.27	0.9372
SGFD FD	SVM	83.75	0.9669
	KNN	86.90	0.9650
	RF	90.79	0.9842
MSCFD FD	SVM	78.85	0.9406
	KNN	84.46	0.9532
	RF	91.08	0.9865
SNVFD FDFD	SVM	87.78	0.9697
	KNN	86.33	0.9630
	RF	89.06	0.9794

3.3. Test Set Analysis

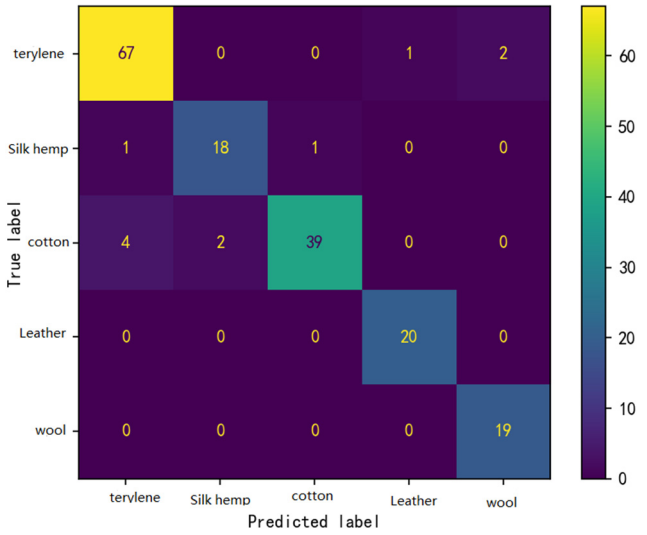

Fig 5. Confusion matrix of the SG_FR model predicting the test set

Based on the 10-fold cross-validation results, SMOTETomek combined sampling before data preprocessing and the SG_RF model were selected to analyze the test set. The classification confusion matrix of the test set is shown in Figure 5. The test set contains 70 polyester, 20 linen-silk, 45 cotton, 20 leather, and 19 wool spectral data. The results show that 68 polyester samples were correctly classified, with 1 polyester classified as leather and 2 polyester classified as wool; 18 linen-silk samples were correctly classified, with 1 linen-silk classified as polyester and 1 linen-silk classified as cotton; 39 cotton samples were correctly classified, with 4 cotton samples classified as polyester and 2 cotton samples classified as linen-silk; all 20 leather samples were correctly classified; all 19 wool samples were correctly classified, resulting in a classification accuracy of 93.7%. The ROC curve is shown in Figure 6, with an AUC_macro of 0.928.

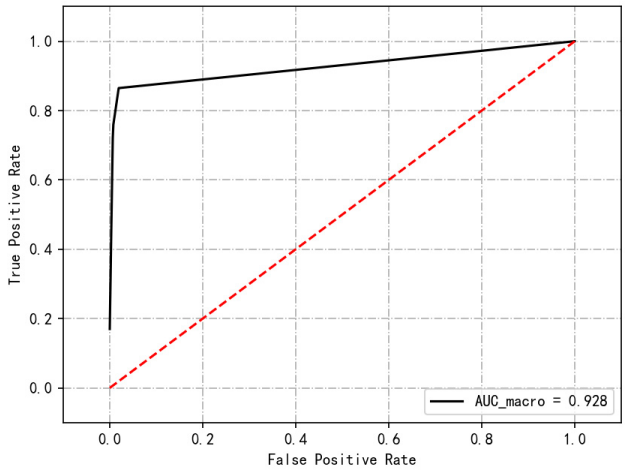

Fig 6. ROC curve of the SG_FR model predicting the test set

3.4. Feature Processing

The PCA algorithm was used for feature extraction on the fiber spectral data. We reduced the spectral dimensionality to 1-20 dimensions and analyzed the classification accuracy and AUC_macro of the test set data after combining SMOTETomek sampling and the SG_RF model. The results are shown in Figure 7. As the dimensionality increased, both the classification accuracy and AUC_macro initially increased, reaching a maximum at 9 dimensions, after which they stabilized. This indicates that increasing the dimensionality no longer captures additional spectral information to improve the classification results. At 9 dimensions, the classification accuracy was 0.902 and the AUC_macro was 0.927. Both values were lower than the classification results of the test set without feature extraction. These results suggest that the PCA feature extraction method is not well-suited for this set of spectra.

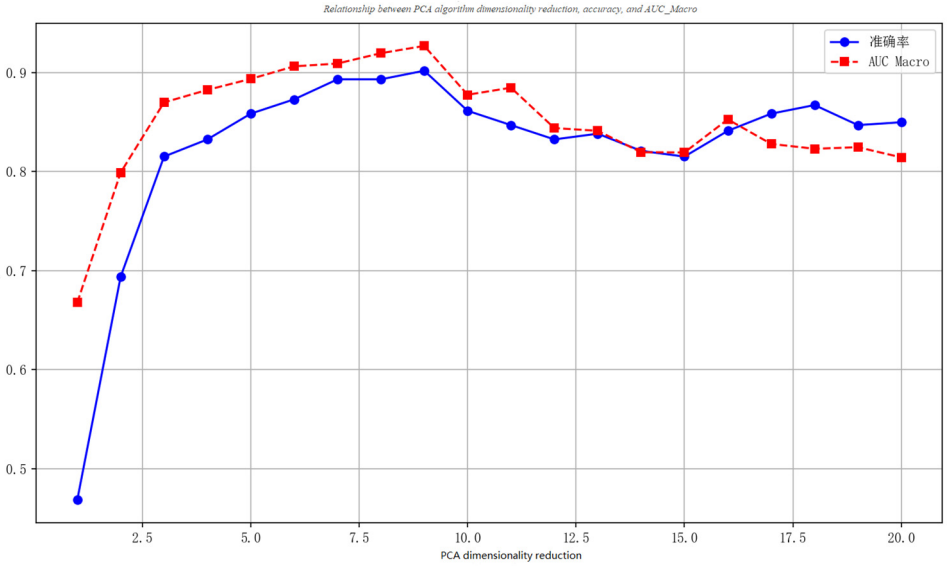

Fig 7. Relationship between PCA algorithm dimensionality reduction, accuracy, and AUC_Macro

The CARS algorithm was used for feature extraction on the fiber spectral data. We set the number of selected features to range from 1 to 20 and analyzed the classification accuracy and AUC_macro of the test set data after combining SMOTETomek sampling and the SG_RF model. The results are shown in Figure 8. As the number of selected features increased, both classification accuracy and AUC_macro initially increased, then stabilized. This indicates that further

feature selection no longer captured additional spectral information to improve classification results. When the number of selected features was 5, the accuracy reached 93.9%, and the AUC_macro exceeded 0.938, both of which were better than the classification results of the test set without feature extraction. Additionally, compared to the PCA algorithm, both classification accuracy and AUC_macro were slightly higher after feature extraction with CARS. These results indicate that the CARS feature extraction method is effective in improving classification performance and slightly outperforms the PCA feature extraction method.

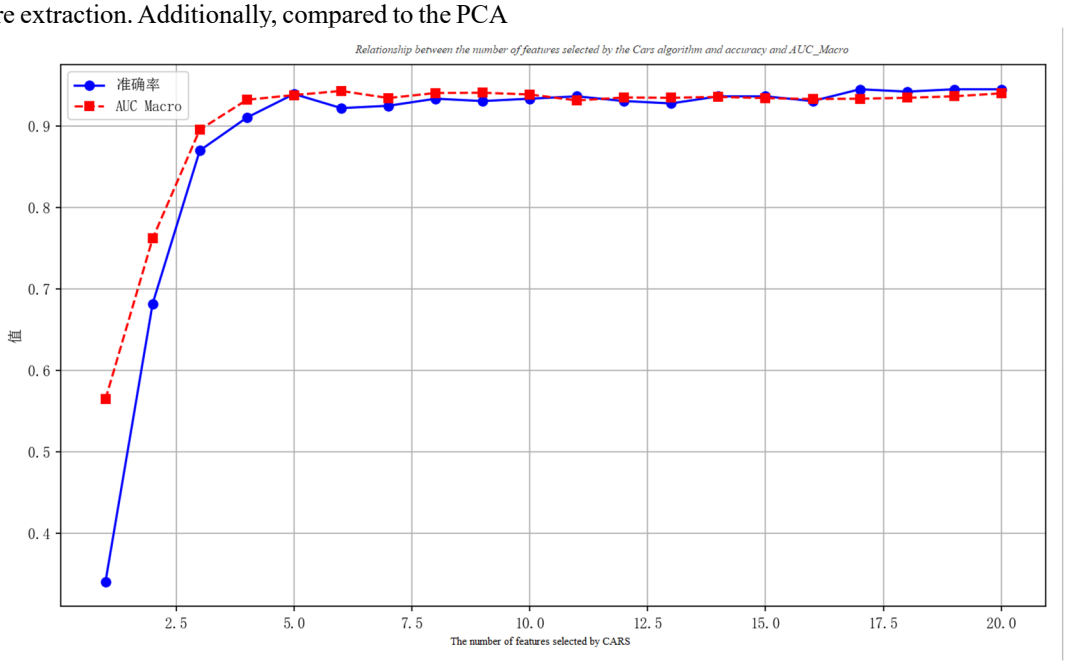

Fig 8. Relationship between the number of features selected by the Cars algorithm and accuracy and AUC_Macro

4. Summary

This study demonstrates the feasibility of using a handheld near-infrared spectrometer combined with SMOTETomek combined data sampling, data preprocessing algorithms, feature processing algorithms, and classification algorithms for fiber classification. By using the handheld near-infrared spectrometer to collect spectral data from fiber samples, a fiber classification model was established. In fiber classification analysis, due to the imbalance in fiber sample data, SMOTETomek combined data sampling was used to obtain a balanced training dataset, and AUC_macro and classification accuracy were employed to evaluate the model's performance. 10-fold cross-validation was used to determine the optimal preprocessing algorithm, classification algorithm, and sampling sequence. Finally, SMOTETomek combined sampling was performed before data preprocessing. After applying the determined SG_RF analysis model, the classification accuracy of the test set data reached 93.7%, with an AUC_macro of 0.928. When CARS feature extraction was applied with 5 selected features, the classification accuracy of the test set increased to 93.9%, and the AUC_macro improved to 0.938. When the number of selected features was increased to 20, the classification accuracy of the test set reached 94.5%, with an AUC_macro of 0.940. In conclusion, our research results show that the handheld near-infrared spectrometer, combined with relevant chemometric applications, can achieve a classification accuracy exceeding 94% and an AUC_macro value over 0.940 in fiber classification. This provides fast, convenient, and non-destructive technical support for the classification of fibers.

References

[1] Zhang Jingwen Yuan Bingnian, "A review of the application of qualitative analysis methods for fiber content in textiles".

[2] Chai Jinchao, Jin Shangzhong, "Application of near infrared spectroscopy in qualitative testing of textiles".

[3] Wang Danhong, Wu Wenxi, Lin Zhiwu, Tu Mandi, "Identification of four types of fibers including Tencel by near infrared spectroscopy".

[4] Fu Jianhua, "Application of Near Infrared Spectroscopy in Textile Testing".

[5] A. Aljannahi et al., "Forensic Analysis of Textile Synthetic Fibers Using a FT-IR Spectroscopy Approach," Molecules, vol. 27, no. 13, p. 4281, Jul. 2022, doi: 10.3390/molecules 2713 4281.

[6] P. Chapanya, P. Ritthiruangdej, R. Mueangmontri, A. Pattamasuwan, and W. Vanichsriratana, "Temperature Compensation on Sugar Content Prediction of Molasses by Near-Infrared Spectroscopy (NIR)," Sugar Tech, vol. 21, no. 1, pp. 162–169, Feb. 2019, doi: 10.1007/s12355-018-0635-x.

[7] Q. Du et al., "Adulteration detection of corn oil, rapeseed oil and sunflower oil in camellia oil by in situ diffuse reflectance near-infrared spectroscopy and chemometrics," Food Control, vol. 121, p. 107577, Mar. 2021, doi: 10.1016/j. foodcont. 2020. 107577.

[8] H. Yan and H. W. Siesler, "Identification Performance of Different Types of Handheld Near-Infrared (NIR) Spectrometers for the Recycling of Polymer Commodities," Appl. Spectrosc., vol. 72, no. 9, pp. 1362–1370, Sep. 2018, doi: 10. 1177/0003702818777260.

[9] Ye Fei, Liu Weihong, Yang Juanya, Chen Chaohong, Wang Zhenhua, Huo Zhengtong, Qu Ruide, Wang Xiaodong, "Multi-feature fusion fiber classification algorithm based on support vector machine".

[10] D. Li, Z. Liu, P. Xiao, J. Zhou, and D. Jahed Armaghani, "Intelligent rockburst prediction model with sample category balance using feedforward neural network and Bayesian optimization," Undergr. Space, vol. 7, no. 5, pp. 833–846, Oct. 2022, doi: 10.1016/j.undsp.2021.12.009.

[11] ZHE Wang, C Wu, K Zheng, X Niu, X Wang, “SMOTETomek-based resampling for personality recognition”.

[12] Z. Yang et al., “Fast determination of oxides content in cement raw meal using NIR spectroscopy combined with synergy interval partial least square and different preprocessing methods,” Measurement, vol. 149, p. 106990, Jan. 2020, doi: 10.1016/j.measurement.2019.106990.

[13] Y. Wu, S. Peng, Q. Xie, Q. Han, G. Zhang, and H. Sun, “An improved weighted multiplicative scatter correction algorithm with the use of variable selection: Application to near-infrared spectra,” Chemom. Intell. Lab. Syst., vol. 185, pp. 114–121, Feb. 2019, doi: 10.1016/j.chemolab.2019.01.005.

[14] Y. Bi et al., “A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation,” Anal. Chim. Acta, vol. 909, pp. 30–40, Feb. 2016, doi: 10.1016/j.aca.2016.01.010.

[15] G. Tang et al., “A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm,” The Analyst, vol. 139, no. 19, p. 4894, Jul. 2014, doi: 10.1039/C4AN00837E.

[16] Duhok Polytechnic University, B. M. Salih Hasan, A. M. Abdulazeez, and Duhok Polytechnic University, “A Review of Principal Component Analysis Algorithm for Dimensionality Reduction,” J. Soft Comput. Data Min., vol. 02, no. 01, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.003.

[17] Y. Liu, K. Wen, Q. Gao, X. Gao, and F. Nie, “SVM based multi-label learning with missing labels for image annotation,” Pattern Recognit., vol. 78, pp. 307–317, Jun. 2018, doi: 10.1016/j.patcog.2018.01.022.

[18] N. C. Dingari et al., “Development and comparative assessment of Raman spectroscopic classification algorithms for lesion discrimination in stereotactic breast biopsies with microcalcifications,” J. Biophotonics, vol. 6, no. 4, pp. 371–381, Apr. 2013, doi: 10.1002/jbio.201200098.

[19] A Liaw, “Classification and regression by randomForest,” 2002.

[20] S. Pazi, C. M. Clohessy, and G. D. Sharp, “A framework to select a classification algorithm in electricity fraud detection,” South Afr. J. Sci., vol. 116, no. 9/10, Sep. 2020, doi: 10.17159/sajs.2020/8189.

[21] J Huang, J Lu, CX Ling - Third IEEE International Conference …, 2003 - ieeexplore.ieee.org J Huang, J Lu, CX Ling, “Comparing naive Bayes, decision trees, and SVM with AUC and accuracy,” 2003.

[22] C Cortes, M Mohri - … neural information processing systems, 2004 - proceedings.neurips.cc C Cortes, M Mohri, “Confidence intervals for the area under the ROC curve”.