Classification of Leguminous Wood Species Based on Small Sample Hyperspectral Images

Hang Su,[a] Shuo Xu,[a] Zhongjian Wang,[a] Wenxin Zhao,[a] Yanan Wen,[b] and Lei Zhao [a,]*

Leguminous wood occupies an important position in the market of cultural and high-end wood. Accurate identification and classification of its species is crucial for the development of the industry. However, existing studies are still deficient in classification methods under small sample conditions. This paper uses hyperspectral image data and combines models such as support vector machine (SVM), random forest (RF), logistic regression (LR), and one-dimensional convolutional neural network (1-CNN). The synthetic minority oversampling technique (SMOTE) data enhancement technology was introduced to classify and recognize 18 common legume woods. After data processing, the classification accuracy of the traditional models was improved by about 5% on average, with the SVM model reaching 98.86%; the accuracy of the 1-CNN model was increased to 97.67% after adding the first-order derivative transform and Savitzky-Golay filtering, it reached 98.89% after further adding the SMOTE.

DOI: 10.15376/biores.20.3.6317-6337

Keywords: Hyperspectral image data; Leguminous wood classification; SMOTE Data Enhancement; 1-CNN

Contact information: a: College of Science and Information, Qingdao Agricultural University, Qingdao 266109; b: Qingdao Quenda Terahertz Technology Co., Ltd. China;
**Corresponding author: lei.zhao@qau.edu.cn*

INTRODUCTION

Due to its excellent physical properties and a wide range of application scenarios, legume wood occupies an important position, especially in the high-end furniture and artifacts market. This type of wood is widely used in the production of high-end furniture, handicrafts, and decorations due to its high hardness, durability, superior resistance to compression and bending, as well as resistance to degradation. However, the scarcity and high market value of legume timber has led to a large number of counterfeit and shoddy timber flooding the market, which not only affects the rights and interests of consumers, but it also poses a serious challenge to the fair competition and healthy development of the industry. According to the statistics of domestic e-commerce platforms, there are more than 20 types of high-end wood used in the market for making cultural games and decorations, of which leguminous wood occupies the vast majority of the share. Due to the lack of identification technology and insufficient standardized management, the existence of shoddy wood has caused huge economic losses and market confusion. Therefore, how to quickly and accurately identify the species of leguminous timber and avoid the inflow of counterfeit and shoddy products into the market has become a key problem that needs to be solved in the current timber market.

With the increasing application of hyperspectral technology in wood research, it has become a cutting-edge research direction for wood species identification by virtue of its non-destructive measurements, high-throughput analysis, and rich information dimensions. By analyzing the reflectance spectra of wood samples, hyperspectral technology can reveal the physical properties of wood such as color, texture, density, hardness, *etc*., and capture the subtle differences in the growth environment, processing and drying techniques, providing rich feature information for the accurate classification of wood species. However, how to fully mine and utilize this feature information to solve the classification problem under small sample conditions with limited sample size is still a technical challenge that needs to be investigated. Therefore, exploring effective hyperspectral data processing and classification methods is of great theoretical and practical significance to improve the accuracy and reliability of wood species identification.

Research Status

Hyperspectral imaging technology has proven to be a powerful tool in wood species identification, especially when combined with machine learning and deep learning models. Several studies have demonstrated the potential of this approach for achieving high classification accuracy. For instance, Zhu *et al.* (2019) used convolutional neural networks (CNNs) coupled with hyperspectral imaging for soybean variety identification, which could be extended to wood species recognition (Zhu *et al.* 2019). Pan *et al*. (2023) proposed a deep learning multimodal fusion framework using near-infrared spectroscopy, GADF, and RGB images, showing how multimodal data can enhance classification performance. Similarly, Marrs and Ni-Meister (2019) applied LiDAR and hyperspectral data for tree species classification, indicating the promise of integrating spatial and spectral data for improved accuracy.

However, these studies often face limitations related to the sample size and data complexity. For example, Aydemir and Bilgin (2017) addressed small sample sizes with a semi-supervised classification method, but their results still suggest that small datasets may not fully represent the variability of wood species, limiting the generalization of the model. Chen *et al.* (2024) used hyperspectral imaging combined with machine learning for *Dalbergia* species identification, but this approach might struggle when dealing with very limited samples. Additionally, Masoumi and Bond (2024) focused on predicting moisture content and swelling in thermally modified hardwoods, but the direct application of their model for wood species classification remains unclear.

Ravindran *et al.* (2021) and Gerasimov *et al*. (2016) applied hyperspectral and Raman spectroscopy methods for wood species identification, but their approaches primarily focus on standard spectral data, which could benefit from incorporating advanced preprocessing methods to better handle noisy data and improve classification accuracy in complex environments. Zhao *et al*. (2021) proposed a fuzzy reasoning and decision-level fusion technique for wood species recognition using visible and near-infrared spectral analysis, but it still requires further refinement to handle the complexities of heterogeneous datasets and limited data points. Fabijańska *et al*. (2021) employed residual CNNs for wood species classification from wood core images, demonstrating the power of deep learning, but they did not account for preprocessing methods like derivative transformations or filtering techniques that could enhance the input data quality, especially when dealing with small sample sizes.

Objective and Scope of This Study

The main objective of this study was to address the limitations of existing legume wood classification methods under small sample conditions, by combining hyperspectral image data, advanced data processing techniques, and various classification models to enhance the accuracy and reliability of legume wood classification. Specifically, the research objectives included the following aspects:

Solving the classification accuracy problem in small sample learning: To address the overfitting issue of traditional classification methods under small sample data, synthetic minority oversampling technique (SMOTE) data augmentation was employed to increase the diversity of training samples, thereby improving the model's generalization ability.

Optimizing feature extraction and modeling of hyperspectral image data: A one-dimensional convolutional neural network (1-CNN) was combined with traditional machine learning methods, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). Data preprocessing techniques such as Savitzky-Golay filtering and first-order derivative transformation are introduced to maximize the potential of hyperspectral data and enhance the performance of the classification model.

While hyperspectral imagery inherently contains both spectral and spatial information, this study specifically focuses on exploiting spectral signatures for material discrimination. The experimental design prioritized spectral resolution (0.3353 nm) over spatial context for two key reasons: (1) The target samples exhibited homogeneous texture characteristics under macroscopic observation, reducing the immediate necessity for spatial feature extraction; (2) Our preliminary tests using SVM classification achieved 92% accuracy without spatial processing, indicating sufficient discriminative power from spectral features alone.

This targeted approach aligns with established methodologies in spectroscopic analysis where spectral fingerprints provide primary identification criteria (Lima *et al.* 2022). Nevertheless, we acknowledge the potential benefits of integrating spatial-textural features for complex heterogeneous materials, which constitutes a critical direction for our subsequent research.

This study focused on the classification and identification of 18 common legume wood species, covering typical legume wood species in the wood market, primarily used in high-end furniture and cultural craft markets. The research employed hyperspectral image data, combined with traditional machine learning methods like SVM, RF, LR, and modern deep learning methods such as 1-CNN, to optimize classification through data augmentation (SMOTE) and hyperspectral data preprocessing (*e.g.*, Savitzky-Golay filtering).

Despite recent advances in wood spectral analysis, three critical challenges remain unaddressed: (1) effective denoising across ultra-broad spectral ranges (400 to 2500 nm) without losing discriminative features, (2) coordinated optimization of sample imbalance and dimensionality curse in small-sample scenarios, and (3) generalization of preprocessing benefits across divergent classifiers. To bridge these gaps, this study delivers threefold innovations:

- First, a cascaded denoising pipeline integrating Savitzky-Golay filtering (for temporal noise suppression) with first-derivative transformation (for spectral slope enhancement), specifically tailored for wide-band hyperspectral characteristics.
- Second, a parallelized SMOTE-PCA co-optimization framework that simultaneously addresses class imbalance and feature redundancy through complementary

dimensionality operations—SMOTE expanding sample diversity in original space while PCA extracting compact representations.

- Third, comprehensive validation across four classifier archetypes (SVM, RF, LR, 1D-CNN), demonstrating for the first time that preprocessing-induced accuracy gains (avg. +5%) are model-agnostic, thus providing a universal solution for spectral data scarcity.

These innovations collectively establish a new paradigm for small-sample hyperspectral analysis, with particular efficacy in leguminous wood identification where chemical homogeneity and sample paucity coexist.

EXPERIMENTAL

Sample Preparation

According to the definition of leguminous wood in the International Code of Botanical Nomenclature (ICN), this work took 18 species of leguminous wood as the research object. Detailed information on these woods is shown in Table 1. In order to prevent homogeneity, the same wood samples were purchased from different merchants and on different dates, thus ensuring that the same wood samples did not come from the same tree or all came from the same area.

Table 1. Sample Data of Leguminous Woods

No.	Scientific Name	Main Characteristics	Main Distribution Area
1	*Guibourtia*	High density, corrosion-resistant, commonly used in high-end furniture and flooring	Tropical Africa
2	*Guibourtia conjugata*	Fine grain, durable, suitable for decorative crafts	Tropical Africa
3	*Pterocarpus erinaceus* Poir.	Clear texture, hard wood, commonly used for rosewood furniture materials	West Africa
4	*Streblus* sp.	Clear texture, lightweight wood, suitable for general furniture manufacturing	Southeast Asia and South Asia
5	*Dalbergia cultrata* Graham	Heavy and fine, commonly seen with black-brown stripes, used for high-end furniture	South Asia
6	*Dalbergia nigra* Allem.	Hard and dense material, dark color, commonly used for musical instruments and decorations	Brazil and the tropical rainforests of South America
7	*Pterocarpus soyauxii* Taub.	Wood color is warm and suitable for carving and decorative use	Tropical Africa
8	*Swartzia* spp.	High hardness, high density, strong corrosion resistance, often used for high-end crafts and flooring	South America
9	Golden rosewood	Golden color, tough wood, suitable for making decorative items	Southeast Asia and Tropical Africa
10	*Millettia*	High hardness of wood, suitable for indoor decoration and flooring	Tropical Asia and Africa
11	Côte d'Ivoire rosewood	Deep red color, beautiful wood grain, commonly used to make musical instruments and high-end furniture	West Africa

No.	Scientific Name	Main Characteristics	Main Distribution Area
12	Burma padauk	Stable wood, fine texture, often used for rosewood furniture and decorative materials	Myanmar and Southeast Asia
13	Mexican rosewood	Rich color, suitable for carving and small decorative items	Mexico and Central America
14	Black ebony	High density, high hardness, fine and smooth material, often used for high-end musical instruments and decorations	Tropical Africa
15	*Pterocarpus santalinus*	Deep red wood, high hardness, fine texture, suitable for traditional crafts and Buddhist beads	India, Southeast Asia
16	*Pterocarpus indicus*	Stable wood, bright color, used for decorative furniture and flooring	Southeast Asia and Tropical Asia
17	Peltogyne	Distinct purple tone, dense and durable wood, suitable for high-end furniture and decorations	South America, especially the Amazon Rainforest in Brazil
18	*Pterocarpus tinctorius* Welw.	Deep red wood, high hardness, commonly used for carving and traditional crafts	Tropical Africa

Before data collection, the length, width, and height of all wood blocks were unified to 6 cm × 4 cm × 2 cm, and the long side corresponded to the cross section of the wood. Among the cut samples, two samples of each type of wood were taken for processing. During the selection process, pure samples without cracking, insect infestation, or oil contamination were selected. Before measurement, sandpaper with gradually finer grain sizes (240, 400) was used, with grain sizes of 600, 800, 1,000, 1,500 used for polishing.

The data collection platform was Resonon Pika L03030988 hyperspectral imager. The spectrum extraction and analysis software is SpectrononPro. In this study, SpectrononPro software was used to process and analyze hyperspectral image data. SpectrononPro is a professional hyperspectral image processing software that is widely used in remote sensing, agriculture, geology, ecology, environmental monitoring, and other fields. It provides a variety of data preprocessing functions such as atmospheric correction, geometric correction, and radiometric correction to ensure the accuracy and consistency of data; at the same time, the software has spectral analysis tools such as spectral curve extraction, spectral matching, and spectral mixing analysis to help in-depth exploration of the sample feature; Fig. 1. shows the appearance and working schematic of the hyperspectral instrument.

Data Preprocessing and Feature Extraction

The wavelength range of the spectral data collected using the spectrometer was between 350 and 1050 nm. The spectral resolution was 0.3353 nm, and its dimension was 1050. To enhance spectral separability, critical preprocessing steps including regional averaging, smoothing, baseline correction, and noise reduction were systematically implemented. Direct classification of raw spectra risks triggering the 'curse of dimensionality' and compromises computational efficiency. Thus, dimensionality reduction through spectral feature optimization is necessary.

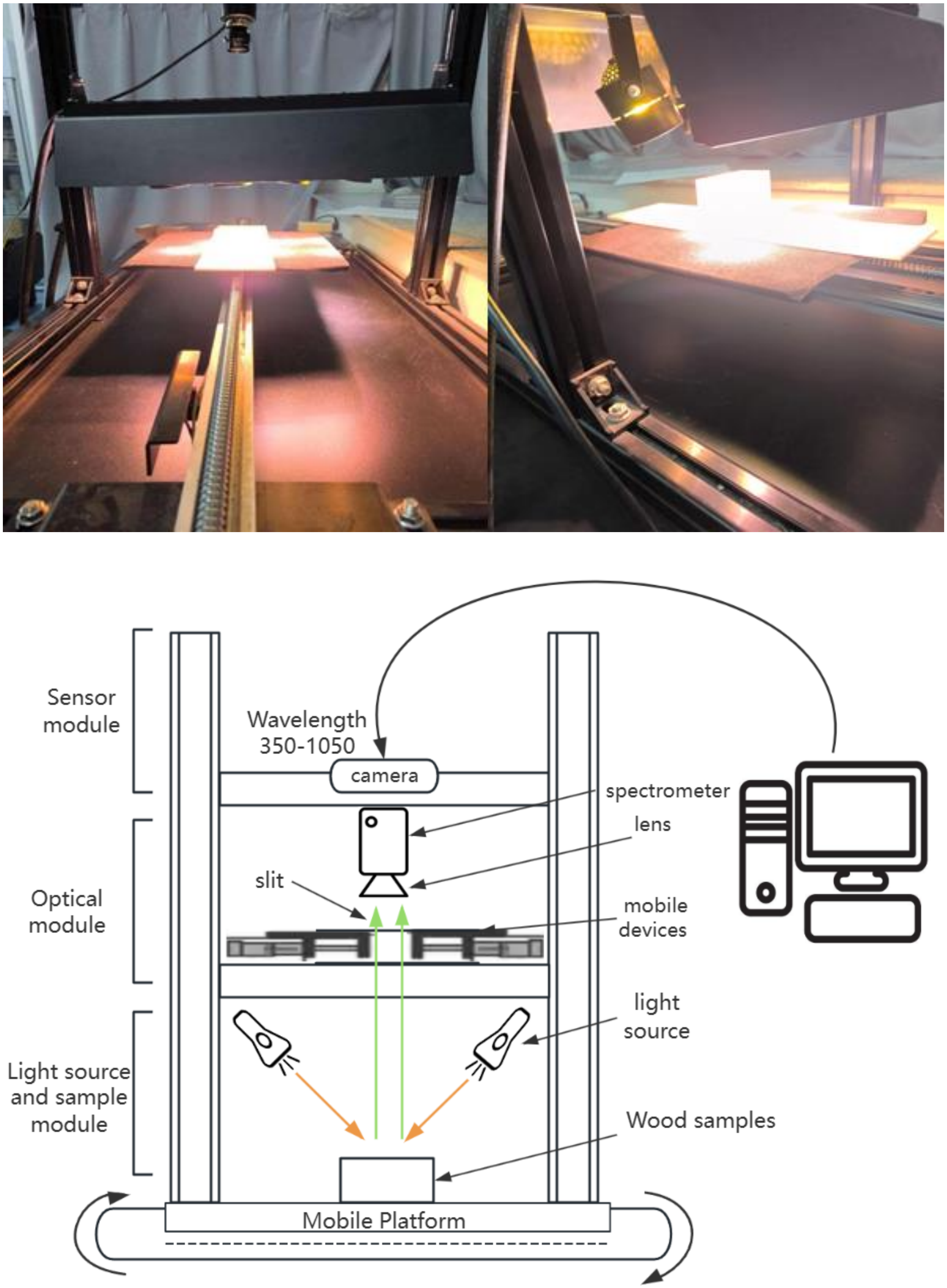

Fig. 1. Hyperspectral instrument and working schematic

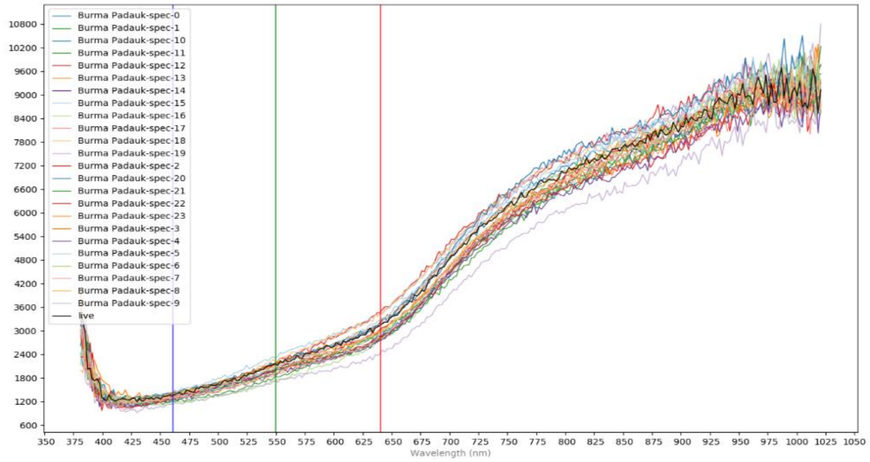

Fig. 2. Raw spectral reflectance curve

Figure 2 illustrates the original spectral reflectance curves containing 1050 dimensional features, demonstrating inherent spectral variability across samples. Figure 3. presents the optimized spectral signature processed through SpectrononPro's automated workflow. This procedure generates a representative average spectrum by integrating regional spectral features within designated wavelength intervals, effectively reducing data dimensionality while preserving discriminative information. The contrast between the multi-curve representation in Fig. 2 and the unified curve in Fig. 3 visually demonstrates how preprocessing transforms high-dimensional raw data into a compact spectral profile suitable for efficient pattern recognition.

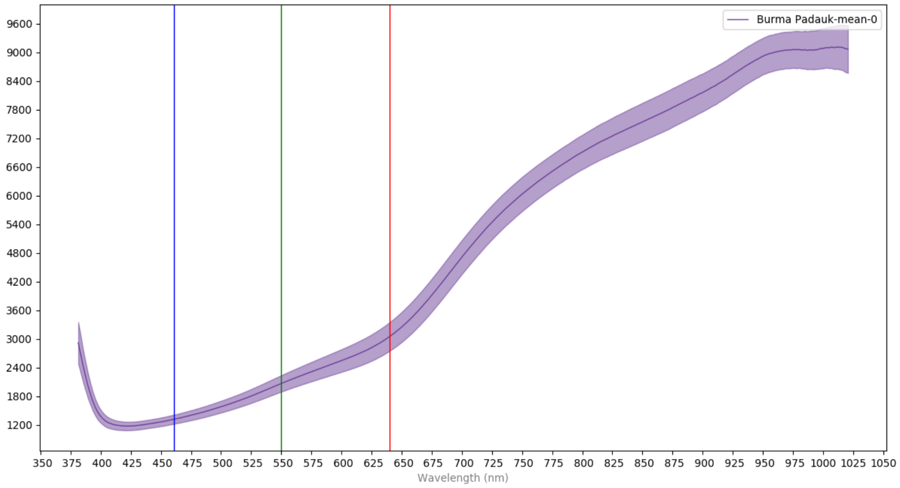

Fig. 3. Spectral curve after data processing

To eliminate the impact of feature dimensional differences, data normalization was done. The normalization process transforms the data into zero mean and unit variance. Let the original data matrix be X , and the normalization formula is as follows:

$$Z = \frac{X - \mu}{\delta} \tag{1}$$

where μ and δ represent the mean and standard deviation of each column of data, respectively. This method was applied to standardize the data, ensuring that all features were on the same scale and effectively reducing the imbalance in the influence of different features on the model. For specific scenarios, the data were also normalized based on the range between the maximum and minimum values, as shown in the following formula:

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

The first step of PCA is to standardize the data to eliminate the dimensional differences of different features. The aforementioned formula was used to standardize the data matrix X, obtaining the standardized matrix Z. After standardization, the covariance matrix S of the data was computed to quantify the correlation between features. The formula for the covariance matrix is:

$$S = \frac{1}{n} Z^T Z \tag{3}$$

where Z^T is the transpose of the standardized data matrix, and n is the number of samples.

Next, eigenvalue decomposition of the covariance matrix S was carried out to obtain the eigenvalues λ_i and corresponding eigenvectors μ_i. The formula is:

$$S\mu_i = \lambda_i \mu_i \tag{4}$$

where λ_i represents the variance of the principal components, and μ_i represents the direction of the principal component.

The principal components with larger eigenvalues represent the directions of the largest variance in the data. Based on the cumulative explained variance ratio, the first k principal components were selected to ensure that at least 95% of the information was retained. The cumulative explained variance ratio is given by:

$$Cumulative\ \mathrm{E}\text{xplained}\ Variance\ \mathrm{R}\text{atio} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{5}$$

where m is the total number of features.

Based on the experimental results, the first k eigenvectors U were selected to form the new basis.

Finally, the standardized data Z were projected onto the new principal component space to obtain the reduced-dimensional data matrix Y:

$$Y = ZU \tag{6}$$

where U is the matrix containing the first k eigenvectors. This step significantly reduced the dimensionality of the data and improved the efficiency of subsequent model training.

The results of the first two principal components, PC1 and PC2, extracted by PCA show that the main information of the data was effectively captured:

Principal Component 1 (PC1) captured the primary trend with the largest variation in reflectance, related to the overall spectral information, and reflected the global features of the samples.

Principal Component 2 (PC2) represents the secondary variation direction orthogonal to PC1, capturing subtle features under different wavelength combinations.

The application of PCA significantly reduced the dimensionality of the data while retaining 95% of the variance information, providing a solid foundation for model training and analysis.

Identification of Leguminous Tree Species Using Initial Data

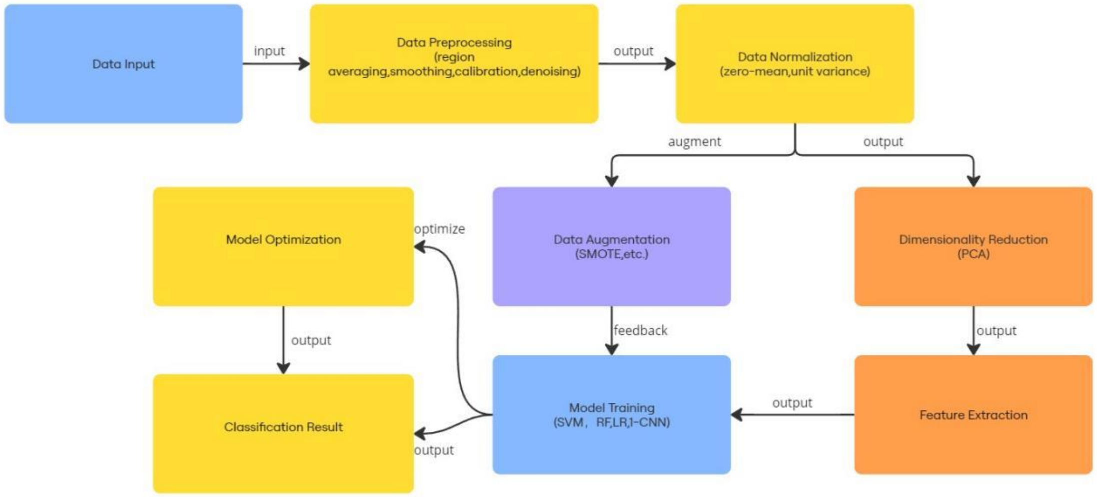
Data Input
input
Data Preprocessing (region averaging,smoothing,calibration,denoising)
output
Data Normalization (zero-mean,unit variance)
augment
output
Model Optimization
optimize
Data Augmentation (SMOTE,etc.)
Dimensionality Reduction (PCA)
output
feedback
output
Classification Result
Model Training (SVM， RF,LR,1-CNN)
output
Feature Extraction
output

Fig. 4. Hyperspectral data processing and classification flow chart

Figure 4 shows a flowchart for processing tree species spectral data. It should be noted that PCA dimensionality reduction and SMOTE processing are performed simultaneously rather than sequentially. The SMOTE method solves the problem of class imbalance by synthesizing minority class samples to enhance the model's ability to represent spectral features. PCA solves the problem of high-dimensional data redundancy by extracting low dimensional principal components of spectral data through orthogonal transformation, alleviating the curse of dimensionality. The two are not simply upstream and downstream relationships, but independent optimizations for data distribution (category balance) and feature space (dimensional redundancy). Parallel processing can avoid coupling interference in sequential operations. If SMOTE is first followed by PCA, then the synthesized high-dimensional samples may lose key discriminative features during the dimensionality reduction process, weakening the data augmentation effect. If PCA is first followed by SMOTE, the low dimensional space after dimensionality reduction is difficult to accurately depict the original spectral distribution, resulting in synthesized samples deviating from the true feature space. The parallel branch design allows the original standardized data to enter two independent processing channels simultaneously. The SMOTE branch generates synthetic samples that conform to the spectral data distribution in the original high-dimensional space, ensuring class balance. PCA branch extracts low dimensional discriminative principal components, eliminates noise and redundancy, and improves computational efficiency.

In this study, four different models were used to classify the hyperspectral images of 18 species of leguminous trees. To ensure the reliability of the results, we first performed a preliminary evaluation of the accuracy for each model. The classification accuracy of each model was recorded without using the SMOTE method.

Random Forest Model: The accuracy of the random forest model reached 88%. This model performed well in classifying most of the tree species, but there were still some errors when handling certain categories.

Support Vector Machine Model (SVM): The accuracy of the SVM model was 92%. The SVM performed better than the random forest model and was able to handle the complex patterns in the hyperspectral data more effectively.

Logistic Regression Model: The classification accuracy of the logistic regression model was 93%. This model performed excellently, accurately classifying most of the samples, but its sensitivity to imbalanced data might have affected the overall performance.

The class imbalance in our dataset inherently reflects the objective imprint of economic principles governing the collectible wood market within research samples. As core commodities in China's premium timber sector, leguminous rosewoods (Dalbergia/Pterocarpus spp.) face triple supply-chain constraints:

Stringent international trade quotas under CITES for endangered species such as *Dalbergia odorifera* severely restrict global circulation volumes. Custom records indicate that annual imports of regulated species account for less than 12% of non-restricted varieties, legally constraining the naturally diminished sample pool accessible to researchers.

Market pricing mechanisms exacerbate sample disparity. Mass-producible species such as *Dalbergia cochinchinensis* (priced below 600,000 CNY/ton) dominate 76% of manufacturers' procurement, while premium-grade *D. odorifera* (exceeding 1.2 million CNY/ton) becomes hoarded by investment entities, creating a paradoxical "circulation without accessibility" scenario for scientific inquiry. This price hierarchy functionally filters samples, compelling research reliance on readily available mid-tier species.

High-value wood sampling incurs disproportionate expenses: destructive testing of auction-grade materials requires substantial security deposits, while non-destructive micro-sampling techniques (to preserve material integrity) prolong processing time by 2.8× per specimen. These combined economic and technical barriers objectively compress sampling scales for rare species.

As per 2023 China Collectibles Market Report, the circulation of sandalwood exhibits the Matthew effect:

Dalbergia nigra accounts for 38.7% of the total.

Guibourtia accounts for 29.1% of the total.

Pterocarpus soyauxii accounts for only 5.3%.

1D Convolutional Neural Network Model (1-CNN): The initial classification accuracy of the 1-CNN model was 92.25%.

In hyperspectral image classification, due to the class imbalance in the dataset, minority class samples often lead to the model being biased towards the majority class, thereby affecting classification accuracy. To address this issue, this study adopted three methods: first-order derivative transformation, Savitzky-Golay filtering, and SMOTE (Synthetic Minority Over-sampling Technique) to process the data, aiming to improve classification accuracy.

The first-order derivative transformation is a differential processing technique applied to spectral data to emphasize the trend of changes and characteristic points on the spectral curve. In hyperspectral data, many important spectral features (such as absorption peaks and reflection peaks) manifest themselves as changes in the spectral curve. By calculating the rate of change for each spectral band, the response in these significantly changing parts can be enhanced, helping the model better identify subtle differences between tree species. Let $S(\lambda)$ represent the value of the spectral curve at wavelength λ, then the formula for the first-order derivative is:

$$\frac{dS(\lambda)}{d\lambda} = \frac{S(\lambda+\Delta\lambda)-S(\lambda)}{\Delta\lambda} \quad (7)$$

where $\Delta\lambda$ is the wavelength step size. This method highlights abrupt changes or variation regions in the spectral data, helping the classification model focus on key features.

The Savitzky-Golay filtering is a commonly used smoothing technique, especially suitable for noise removal while retaining the main features of the signal. In hyperspectral data, noise can interfere with the model training process, leading to incorrect classification results. The Savitzky-Golay filter smooths the data by performing local polynomial fitting on the spectral data, thus reducing the noise's influence. This filtering method is a filtering method based on local polynomial least squares fitting in the time domain. Its biggest feature is that it can keep the shape and width of the signal unchanged while filtering out noise. The basic principle is to fit the data with a sliding window, using local polynomial approximations to smooth the data. The algorithm idea is to suppress noise through motion smoothing.

Figure 5 explains the principle of how the sliding window smoothes the function graph.

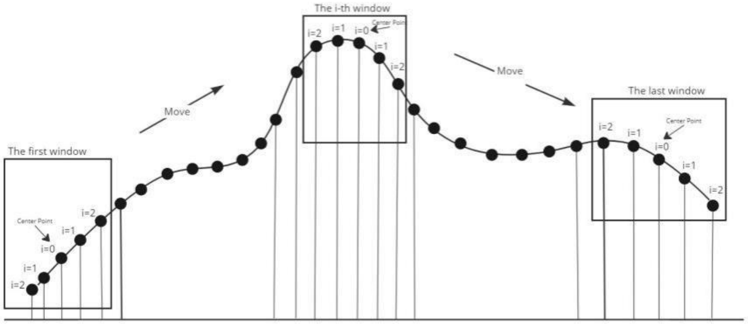

Fig. 5. Schematic diagram of sliding window noise suppression

The key parameters of SG filtering are the number of window points (non physical width), and the window slides along the standardized data point number (step size=1 point). During processing, it does not rely on wavelength calibration but only focuses on the numerical relationship between adjacent data points. The 5-point window (the optimal value in this experiment) balances the noise suppression requirements with the ability to preserve spectral peak valley features.

The formula is as follows:

$$y(t) = \sum_{k=-m}^{m} b_k x(t+k) \tag{8}$$

where y(t) is the smoothed data, x(t+k) is the original data point, b_k is the filter coefficient, and m is the window size.

This method allows the Savitzky-Golay filter to smooth spectral data while effectively preserving spectral feature information, improving the quality of the data. Figure 6 shows the process of window sliding and local fitting during the filtering process.

Figure 7 presents a comparative analysis of the spectral reflectance curves for selected wood species samples before and after the application of the Savitzky-Golay (SG) filtering technique. The left subplot displays the original spectral curves, illustrating the inherent variability and noise present in the raw hyperspectral data. Each blue line represents the reflectance spectrum of an individual sample, capturing the detailed fluctuations across the wavelength range from 350 to 1050 nm. In contrast, the right subplot showcases the filtered spectral curves, depicted as red dashed lines.

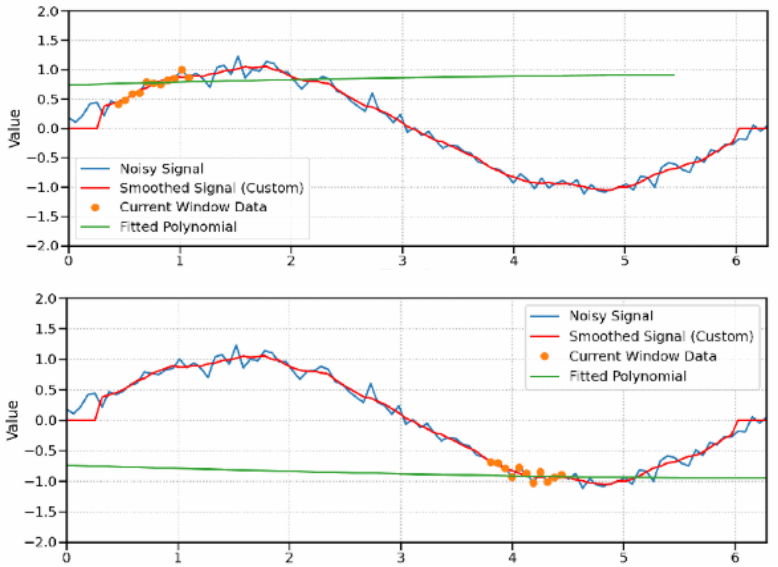

Fig. 6. Process diagram of window sliding and local fitting (The horizontal axis represents the standardized time point)

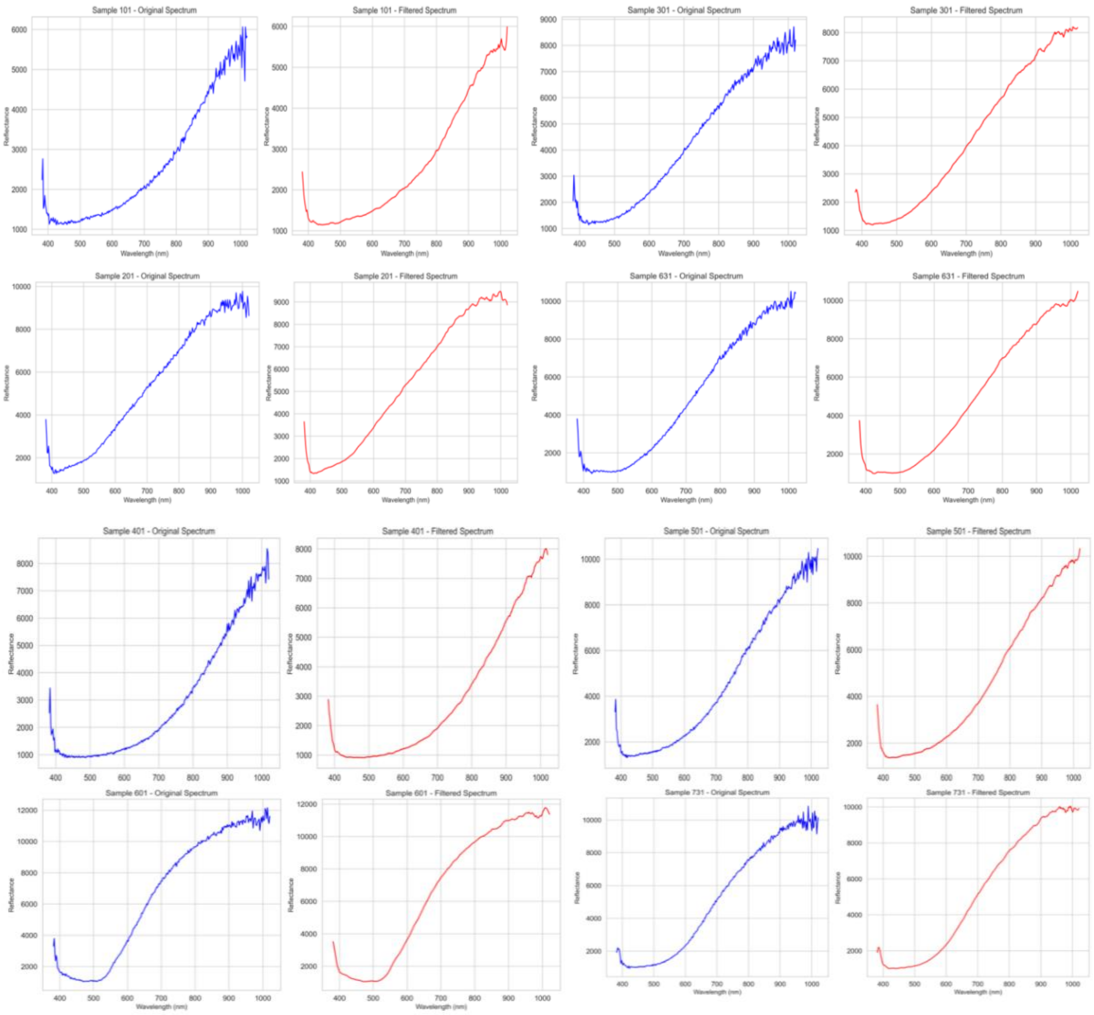

Fig. 7. Spectral curves comparison before and after Savitzky-Golay filtering

The SG filter effectively smooths out high-frequency noise while preserving essential spectral features, resulting in more continuous and less erratic reflectance profiles. This enhancement facilitates better visualization and subsequent analysis by mitigating the impact of random noise without distorting significant spectral characteristics. The side-by-side comparison underscores the efficacy of the Savitzky-Golay filter in improving data quality, thereby enabling more accurate and reliable interpretations of the hyperspectral information.

Figure 8 illustrates the quantitative impact of Savitzky-Golay filtering on the spectral reflectance data by plotting the difference between the filtered and original reflectance values for each selected sample. Each plot represents the reflectance difference (Filtered - Original) across the wavelength range of 350 to 1050 nm for an individual wood species sample. The purple line indicates the magnitude and direction of changes introduced by the filtering process. Positive values signify an increase in reflectance post-filtering, while negative values denote a decrease. This visualization highlights the areas where the SG filter has smoothed out noise and retained or enhanced significant spectral features. By isolating the reflectance differences into separate plots for each sample, the figure avoids overlapping data points, ensuring clarity and facilitating a more precise assessment of the filter's effects. The consistent pattern of reduced variability and enhanced spectral smoothness across samples demonstrates the robustness of the Savitzky-Golay filter in refining hyperspectral data, thereby supporting more accurate classification and analysis of wood species based on their spectral signatures.

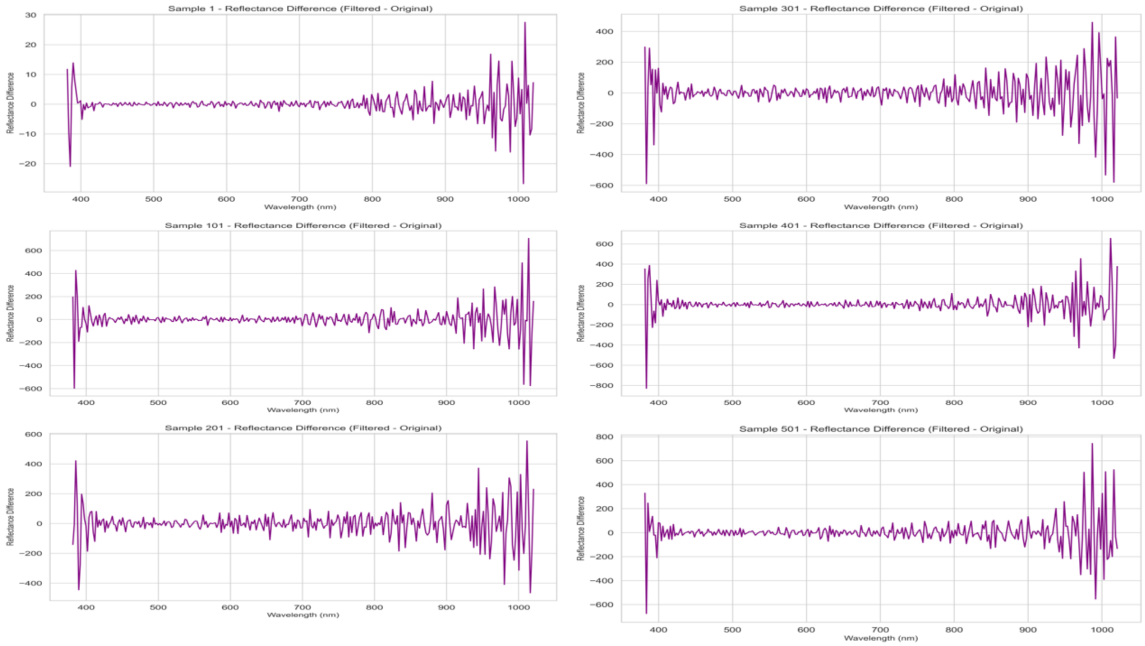

Fig. 8. Reflectance difference after Savitzky-Golay filtering

SMOTE (Synthetic Minority Over-sampling Technique) is a widely used data augmentation method in small-sample learning to address the class imbalance problem. SMOTE increases the proportion of minority class samples in the dataset by synthesizing new samples. Specifically, the SMOTE method interpolates within the feature space of the minority class samples to generate new samples. Let x_i represent a minority class sample and x_j represent its nearest neighbor sample, the new synthetic sample x_{new} can be generated by the following formula:

$$x_{new} = x_i + \delta \cdot (x_j - x_i) \tag{9}$$

where δ is a parameter randomly chosen within the interval [0, 1], controlling the position of the newly generated sample. By averaging the minority class sample with its neighbor samples, the SMOTE method effectively increases the number of minority class samples while preserving the diversity and features of the samples.

The left side of the figure below shows a PCA scatter plot of wood species before applying SMOTE, showing the original data distribution and class separation. The right side of the PCA scatter plot after applying SMOTE shows the expansion of the minority class and the improvement of class balance.

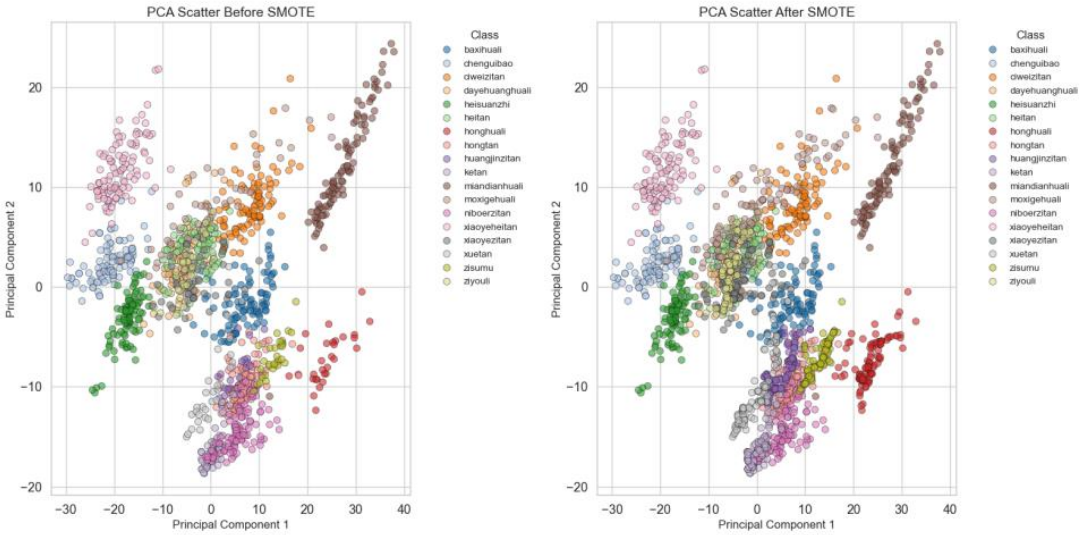

Fig. 9. PCA scatter plot before and after SMOTE

The combination of these methods helps to address the issue of scarce minority class samples, thereby enhancing the model’s learning and generalization ability. The first-order derivative transformation and Savitzky-Golay filtering help to reinforce important features in the spectral data, while the SMOTE method increases the number of minority class samples and balances the sample distribution across classes, improving the model's performance on the minority class.

In this study, to address this issue, this study applied SMOTE (Synthetic Minority Over-sampling Technique) to process the data, aiming to improve classification accuracy. For the 1D CNN model, there was a combining of first-order derivative transformation, Savitzky-Golay filtering, and SMOTE synthesis to enhance the model’s ability to recognize spectral features and improve classification accuracy for minority class samples.

RESULTS AND DISCUSSION

This section presents the classification results for the different machine learning models applied to hyperspectral images of leguminous tree species. Classification performance was evaluated for Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (Logistic Regression), and 1D Convolutional Neural Network (1D

CNN), both before and after applying data augmentation techniques such as SMOTE and first-order derivative transformations.

The core contribution of this study lies in proposing an innovative preprocessing pipeline for small-sample spectral data. Consequently, all classification models were implemented using standard configurations to focus on evaluating the effects of data optimization:

SVM: Implemented through scikit-learn (v1.2) SVC class with default radial basis function (RBF) kernel, regularization parameter C=1.0, and gamma parameter set to 'scale' (*i.e.*, 1/(n_features * X.var())).

Random Forest (RF): Utilized scikit-learn's RandomForestClassifier with default settings: 100 decision trees, Gini impurity as the splitting criterion, and unlimited maximum tree depth.

Logistic Regression (LR): Employed scikit-learn's LogisticRegression with L2 regularization, lbfgs solver, and maximum iterations set to 1000.

1D-CNN: Constructed using Keras framework, consisting of:

An input layer (accepting raw spectral sequences)

One convolutional layer (64 filters, kernel size=3, ReLU activation)

A global average pooling layer (replacing fully-connected layers to reduce parameters)

An output layer (softmax activation with neurons matching category count)

All conventional models (SVM/RF/LR) employed scikit-learn's default parameters, which have demonstrated robust performance in multiple spectral analysis benchmark tasks. For the 1D-CNN implementation, a compact architecture was adopted that is commonly used in spectral analysis (He *et al.* 2020), modifying only the input dimensions to accommodate data characteristics. To ensure reproducibility, all models were initialized with identical random seeds (seed=42). Future work will incorporate advanced optimization techniques such as Bayesian optimization or grid search to further enhance model performance.

Results of Traditional Machine Learning Models

In the initial experiments, Random Forest, Support Vector Machine (SVM), and Logistic Regression models were used for classification tests. These tests did not apply SMOTE or other data enhancement techniques. The experimental results after adding the SMOTE processing method are as follows:

Random Forest (RF): With the introduction of SMOTE processing, the accuracy of the random forest model increased from 88.00% to 92.26%. Although the random forest model performed well in classifying most tree species, there were still some misclassifications, especially between tree species with similar spectral characteristics. Nevertheless, the application of SMOTE effectively improved the overall performance of the model, making it more robust when dealing with unbalanced data.

Support Vector Machine (SVM): After SMOTE processing, the accuracy of the support vector machine was increased from the initial 92.00% to 98.86%. SVM can effectively capture nonlinear relationships and complex patterns in high-dimensional spectral data, and performs better than random forests, especially when distinguishing similar tree species, and it can more accurately identify their spectral differences. Through the application of SMOTE, SVM has improved the problem of class imbalance and reduced the misclassification rate.

Logistic Regression: The accuracy of the logistic regression model increased from 93.00% to 97.26% after SMOTE processing. Logistic regression performs well on linear classification tasks and accurately classifies most samples. However, while the high accuracy indicates good results in general, the model's sensitivity to class imbalance still exists. After applying SMOTE, the model's performance was significantly improved when dealing with imbalanced data, especially on tree species with fewer samples.

Figure 10 presents the confusion matrix for the three machine learning models, highlighting the significant reduction in misclassification rates and the improved performance across all tree species, particularly in the minority class.

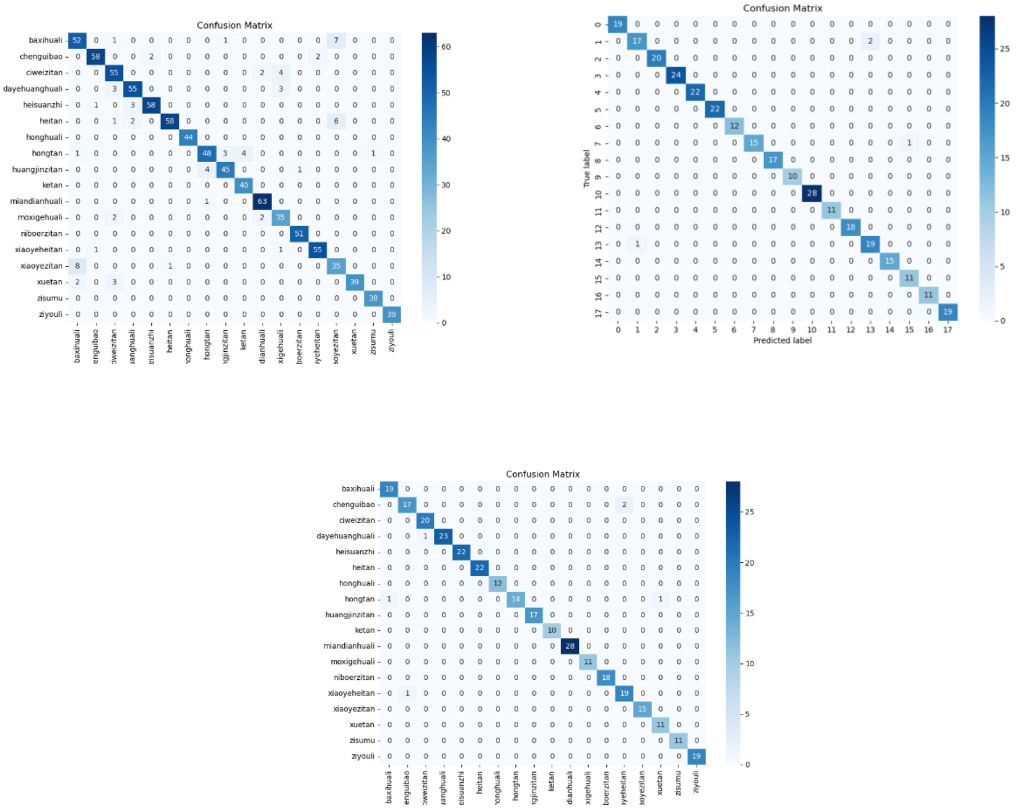

Fig. 10. Random Forest and SVM and Logistic Regression Confusion matrix

Results of the 1D Convolutional Neural Network (1D CNN)

In contrast to the traditional models, the 1D Convolutional Neural Network (1D CNN) model was first evaluated with the initial data and achieved an accuracy of 92.25%. While this result was already quite strong, further improvements were made by applying feature transformation techniques.

First-Order Derivative Transformation: After applying the first-order derivative transformation to the hyperspectral data, the classification accuracy of the 1D CNN model was increased to 97.67%. The first-order derivative transformation highlights changes in spectral features, such as absorption and reflection peaks, which significantly improved the model's ability to distinguish between species with subtle spectral differences.

Combination of First-Order Derivative and SMOTE: When both the first-order derivative transformation and SMOTE were applied together, the accuracy of the 1D CNN model reached 98.89%. This substantial improvement can be attributed to the combination of enhanced spectral features through the derivative transformation and the balanced class distribution achieved by SMOTE. This dual approach allowed the 1D CNN model to learn more effectively from the minority class samples and to better capture the spectral differences between tree species.

The results of the 1D CNN model demonstrate the powerful impact of both feature engineering (through first-order derivatives) and data augmentation (*via* SMOTE) in enhancing classification performance. The combination of these techniques enabled the model to achieve the highest classification accuracy among all models tested.

Figure 11 presents the confusion matrix for the 1D CNN model, highlighting the significant reduction in misclassification rates and the improved performance across all tree species, particularly the minority class.

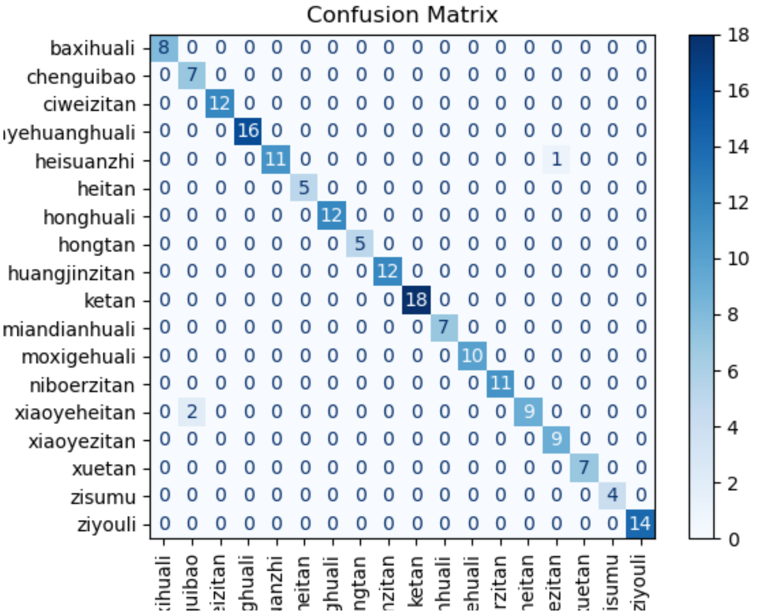

Fig. 11. 1D CNN Confusion matrix

Impact of SMOTE and First-Order Derivative Transformation

The application of SMOTE and first-order derivative transformation significantly influenced the classification results across all models. For traditional machine learning approaches (RF, SVM, and Logistic Regression), SMOTE— as highlighted by Blagus and Lusa (2013), who demonstrated its efficacy in handling high-dimensional class-imbalanced data—helped balance class distributions, thereby improving model generalization and performance on minority classes. Concurrently, the first-order derivative transformation, when applied to the 1D CNN model, enhanced its capability to capture spectral changes and refine sensitivity to key features in hyperspectral data, aligning with the feature enhancement principles underlying such preprocessing techniques.

Table 2. Classification Accuracy of Different Models

Model	Configuration	Accuracy (%)
Random Forest	Without SMOTE	88.00
Random Forest	With SMOTE	92.26
SVM	Without SMOTE	92.00
SVM	With SMOTE	98.86
Logistic Regression	Without SMOTE	93.00
Logistic Regression	With SMOTE	97.26
1D CNN	Initial	92.25
1D CNN	With First-Order Derivative	97.67
1D CNN	With First-Order Derivative +Savitzky-Golay filtering + SMOTE	98.89
This table summarizes the classification accuracy of various machine learning models under different configurations, including the application of SMOTE and preprocessing techniques.		

Together, these techniques address the challenges posed by class imbalance and subtle spectral differences, which are particularly common in hyperspectral image classification tasks. Table 2 summarizes the introduction of all the models experimented above and the improvement in accuracy.

Role of SMOTE in Small Sample Learning

In hyperspectral image classification tasks, class imbalance is a prevalent issue, especially when certain tree species have fewer samples. Traditional classification models tend to be biased towards majority classes, resulting in poorer performance for minority classes. SMOTE addresses this by over-sampling the minority classes, thereby balancing the class distribution.

According to Chawla *et al.* (2002), SMOTE generates feature-space interpolated instances by interpolating between existing minority class samples in the feature space. This not only increases the number of minority class samples but also introduces new sample diversity, reducing the risk of overfitting. By generating feature-space interpolated instances within the minority class's feature space, SMOTE enables classifiers to better learn the characteristics of these underrepresented classes, thereby improving their recognition capabilities.

Advantages of First-Order Derivative Transformation in Spectral Data Processing

First-order derivative transformation is a common spectral preprocessing technique aimed at enhancing the trend and characteristic points of the spectral curve, such as absorption and reflection peaks. This method calculates the rate of change of spectral values with respect to wavelength, thereby emphasizing subtle variations in the spectral data.

Koashi (1999) demonstrated that first-order derivative transformation effectively eliminates baseline drifts and slow trends in spectral data, highlighting regions of rapid change. These rapid changes often correspond to key spectral features that are critical for distinguishing between different materials or species. By enhancing these features, first-order derivative transformation improves the resolution of spectral data, making it easier for classification models to detect and differentiate subtle differences between classes.

For the 1D CNN model, the initial classification accuracy was 92.2%. After applying the first-order derivative transformation, the accuracy increased to 97.67%. This significant improvement indicates that the first-order derivative transformation effectively enhanced the spectral features, allowing the 1D CNN model to better distinguish between tree species with minor spectral differences.

Role of Savitzky-Golay Filtering in Spectral Noise Reduction

Savitzky-Golay filtering, a polynomial-based smoothing technique widely employed in spectral data processing, effectively removes noise while preserving essential spectral features. As demonstrated by John, Sadasivan, and Seelamantula (2021), this method performs local polynomial fitting within a sliding window to smooth spectral data—particularly notable for its adaptive capability in non-Gaussian noise environments, which reduces high-frequency noise without distorting underlying spectral characteristics. This adaptive refinement, as outlined in their study, enhances the technique's robustness across diverse spectral datasets, ensuring both noise reduction and feature integrity.

Introduced by Savitzky and Golay (1964), this filtering technique maintains the integrity of important spectral features such as peaks and valleys by fitting a low-degree polynomial to the data within each window. This approach effectively reduces noise interference in the spectral data, enhancing the signal-to-noise ratio and enabling more accurate feature extraction for classification purposes.

In this study, combining Savitzky-Golay filtering with first-order derivative transformation and SMOTE further improved the 1D CNN model's accuracy to 98.89%. This indicates that Savitzky-Golay filtering successfully reduced noise in the spectral data, allowing the model to better capture and utilize the true spectral features for classification, thereby enhancing overall performance.

CONCLUSIONS

1. This study demonstrated that combining data augmentation techniques, particularly synthetic minority oversampling technique (SMOTE), with spectral feature enhancement methods (*e.g.*, first-order derivative transformation and Savitzky-Golay filtering) significantly improved the classification accuracy of hyperspectral images for leguminous tree species.
2. The 1D Convolutional Neural Network (1D CNN), when integrated with these preprocessing and augmentation techniques, achieved an exceptional accuracy of 98.89%, outperforming traditional models like Random Forest, Support Vector Machine (SVM), and Logistic Regression, which also showed improvements after SMOTE application.
3. The findings emphasize the importance of addressing class imbalance and enhancing spectral features in high-dimensional hyperspectral data classification, which is crucial for advancing agricultural information engineering.
4. The framework developed in this study offers valuable strategies for improving classification performance in similar domains, particularly for imbalanced and complex datasets.

5. Limitations include the potential for synthetic samples from SMOTE to reduce generalizability, and the exclusion of spatial information, which may limit classification accuracy. Future research should explore advanced augmentation techniques like Generative Adversarial Networks (GANs) and integrate spatial-spectral hybrid models to improve robustness and applicability.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant number 62265001).

REFERENCES CITED

Aydemir, M. S., and Bilgin, G. (2017). "Semisupervised hyperspectral image classification using small sample sizes," *IEEE Geoscience and Remote Sensing Letters* 14(5), 802-806. DOI: 10.1109/lgrs.2017.2665679

Blagus, R., and Lusa, L. (2013). "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics* 14, article 106. DOI: 10.1186/1471-2105-14-106

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research* 16, 321-357. DOI: 10.1613/jair.953

Chen, Z., Xue, X., Wu, H., Gao, H., Wang, G., Ni, G., and Cao, T. (2024). "Visible/near-infrared hyperspectral imaging combined with machine learning for identification of ten *Dalbergia* species," *Frontiers in Plant Science* 15, article 1413215. DOI: 10.3389/fpls.2024.1413215

Fabijańska, A., Danek, M., and Barniak, J. (2021). "Wood species automatic identification from wood core images with a residual convolutional neural network," *Computers and Electronics in Agriculture* 179, article 105941. DOI: 10.1016/j.compag.2020.105941

Gerasimov, V. A., Gurovich, A. M., Kostrin, D. K., Selivanov, L. M., Simon, V. A., Stuchenkov, A. B., Paltcev, A. V., and Uhov, A. A. (2016). "Raman spectroscopy for identification of wood species," *Journal of Physics: Conference Series* 741, article 012131. DOI: 10.1088/1742-6596/741/1/012131

He, X., and Chen, Y.-S. (2020). "Transferring CNN ensemble for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters* 18.5, 876-880. DOI: 10.1109/LGRS.2020.2988494

John, A., Sadasivan, J., and Seelamantula, C. S. (2021). "Adaptive Savitzky-Golay filtering in non-gaussian noise," *IEEE Transactions on Signal Processing* 69, article 3106450. DOI: 10.1109/tsp.2021.3106450

Koashi, K. (1999). "Spectral data analysis by two-dimensional representation of derivatives," *Applied Spectroscopy* 53.6, 706 - 712. DOI: 10.1366/0003702991947144

Lima, M. D. R., Trugilho, P. F., Bufalino, L., Júnior, A. F. D., Ramalho, F. M. G., Protásio, T. P., and Hein, P. R. G. (2022). "Efficiency of near-infrared spectroscopy in classifying Amazonian wood wastes for bioenergy generation," *Biomass & Bioenergy* 166(11), article 106617. DOI: 10.1016/j.biombioe.2022.106617

Marrs, J., and Ni-Meister, W. (2019). "Machine learning techniques for tree species classification using co-registered LiDAR and hyperspectral data," *Remote Sensing* 11(7), article 819. DOI: 10.3390/rs11070819

Masoumi, A., and Bond, B. H. (2024). "Prediction of equilibrium moisture content and swelling of thermally modified hardwoods by artificial neural networks," *BioResources* 19(4), 6983-6993. DOI: 10.15376/biores.19.4.6983-6993

Pan, X., Yu, Z., and Yang, Z. (2023). "A deep learning multimodal fusion framework for wood species identification using near-infrared spectroscopy GADF and RGB Image," *Holzforschung* 77(11), 001-006. DOI: 10.1515/hf-2023-0062

Ravindran, P., Owens, F. C., Wade, A. C., Vega, P., Montenegro, R., Shmulsky, R., and Wiedenhoeft, A. C. (2021). "Field-deployable computer vision wood identification of Peruvian Timbers," *Frontiers in Plant Science* 12, article 647515. DOI: 10.3389/fpls.2021.647515

Savitzky, A., and Golay, M. J. E. (1964). "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry* 36(8), 1627–1639. DOI: 10.1021/ac60214a047

Zhao, P., Li, Z.-Y., and Wang, C.-K. (2021). "Wood species recognition based on visible and near-infrared spectral analysis using fuzzy reasoning and decision-level fusion," *Journal of Spectroscopy* 2021, 6088435. DOI: 10.1155/2021/6088435

Zhu, S., Zhou, L., Zhang, C., Bao, Y., Wu, B., Chu, H., Yu, Y., He, Y., and Feng, L. (2019). "Identification of soybean varieties using hyperspectral imaging coupled with convolutional neural network," *Sensors* 19(19), article 4065. DOI: 10.3390/s19194065

Article submitted: Jan. 14, 2025; Peer review completed: May 9, 2025; Revised version received: May 29, 2025; Accepted: June 2, 2025; Published: June 19, 2025.
DOI: 10.15376/biores.20.3.6317-6337