frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/fpls.2025.1597673

Check for updates

OPEN ACCESS

EDITED BY
Ning Yang,
Jiangsu University, China

REVIEWED BY
Wenzheng Bao,
Xuzhou University of Technology, China
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Sotiris Kotsiantis,
University of Patras, Greece

*CORRESPONDENCE
Song Yang
ztriyangs@163.com
Cong Nie
niec@ztri.com.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 16 April 2025
ACCEPTED 28 July 2025
PUBLISHED 20 August 2025

CITATION
Wang C, Fu Y, Wan R, Zhao L, Wang H, Guo J, Liu Q, Li S, Ma S, Wang Z, Huang W, Liu H, Yang S and Nie C (2025) Using preprocessed datasets to construct and interpret multiclass identification models. *Front. Plant Sci.* 16:1597673. doi: 10.3389/fpls.2025.1597673

COPYRIGHT
© 2025 Wang, Fu, Wan, Zhao, Wang, Guo, Liu, Li, Ma, Wang, Huang, Liu, Yang and Nie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Using preprocessed datasets to construct and interpret multiclass identification models

Cong Wang[1†], Yufeng Fu[2†], Ran Wan[1], Le Zhao[1], Hongbo Wang[1], Junwei Guo[1], Qiang Liu[2], Shan Li[3], Shengtao Ma[2], Zhicai Wang[3], Wei Huang[3], Huimin Liu[1], Song Yang[1*] and Cong Nie[1*]

[1]Key Laboratory of Tobacco Chemistry, Zhengzhou Tobacco Research Institute of China National Tobacco Corporation (CNTC), Zhengzhou, China, [2]Technology Center, China Tobacco Henan Industrial Co., Ltd., Zhengzhou, China, [3]Technology Center, China Tobacco Gansu Industrial Co., Ltd., Lanzhou, China

Introduction: Image and near-infrared (NIR) spectroscopic data are widely used for constructing analytical models in precision agriculture. While model interpretation can provide valuable insights for quality control and improvement, the inherent ambiguity of individual image pixels or spectral data points often hinders practical interpretability when using raw data directly. Furthermore, the presence of imbalanced datasets can lead to model overfitting and consequently, poor robustness. Therefore, developing alternative approaches for constructing interpretable and robust models using these data types is crucial.

Methods: This study proposes using preprocessed data—specifically, morphological features extracted from images and chemical component concentrations predicted from NIR spectra—to build multiclass identification models. Combined kernel SVM based models were proposed to identify the rice variety and cultivation region of tobacco. The determination of kernel parameters and percentage of different types of kernel functions were accomplished by PSO, which make the approach self-adaptive. Feature importance and contribution analyses were conducted using Shapley additive explanations (SHAP).

Results: The resulting models demonstrated high robustness and accuracy, achieving classification success rates of 97.9 and 97.4% via n-fold cross validation on rice and tobacco datasets, respectively, and 97.7% on an independent test set (tobacco dataset 2). This analysis identified key variables and elucidated their specific contributions to the model predictions.

Discussion: This study expands the applicability of image and NIR spectroscopic data, offering researchers an effective methodology for investigating factors crucial to the quality control and improvement of agricultural products.

KEYWORDS

multiclass identification, preprocessed data, kernel support vector machine, model interpretation, SHAP, image analysis, near-infrared spectroscopy

1 Introduction

Objective data analysis and machine-learning techniques have been widely employed to construct pattern recognition and regression models in agriculture. Applications include yield prediction (Fita et al., 2025), chemical composition analysis (Rawal et al., 2024), disease and pest diagnosis (Joshi et al., 2024), and soil and land management (Naeimi et al., 2024). Additionally, various chemometric methods have been investigated and applied to achieve high accuracy in agricultural analyses (Stefanov et al., 2010; Jamwal et al., 2021; Xu et al., 2023). These research efforts have significantly enhanced the accuracy and efficiency of relevant tasks while reducing associated costs.

Among the various data-acquisition methods, image analysis and near-infrared (NIR) spectroscopy are commonly used owing to their non-destructive and efficient nature (Antolínez García and Cáceres Campana, 2023; Tian et al., 2023; Zhang et al., 2024). However, most previous studies focusing on these data types have not included model interpretation. A significant challenge is that the direct meaning of individual image pixels, raw spectral data points, or features derived from dimensionality reduction techniques can be ambiguous, thereby hindering practical model interpretation. For the data pre-treatment, instead of giving data straightforward meaning, many research forced on the images recombination and dimensionality reduction for the purpose of increasing the accuracy and robustness of model. Previous studies have demonstrated that morphological features extracted from images can be used to establish identification models for various subjects, including rice and dolphins (Cinar and Koklu, 2019; Sheng et al., 2023). In the field of modeling for medical purpose, Chen et al. (2025) proposed a feature reconstruction method to reconstruct raw features from Conical Beam CT images to eventually detect cleft lip and palate. Furthermore, NIR spectra provide rich structural information of samples, resulting from multiplicative and ensemble absorption of X-H vibrations within hydrogen-containing functional groups (Xiao et al., 2023). Multiple chemical components in crops or fruits can be quantitatively predicted from NIR spectra (Zushi et al., 2025; Wang et al., 2023; Rawal et al., 2024). The straightforward meaning was signed to the data by feature extraction technique and chemical composition prediction, which could provide fundamental of conducting model interpretation method on those processed data.

Model interpretation provides valuable information crucial for controlling or improving the quality of agricultural products. The development of interpretable machine learning has led to the emergence of several methods, including permutation feature importance (Fisher et al., 2019), local interpretable model-agnostic explanations (Ribeiro et al., 2016), and Shapley additive explanations (SHAP) (Lundberg and Lee, 2017). Among these, SHAP has a solid theoretical foundation based on cooperative game theory, offering unique advantages such as fairness guarantees and the ability to provide contrastive explanations (Molnar, 2024). Consequently, it has garnered significant research attention and has been applied in diverse fields, including revealing causes of citrus fruit cracking (Abekasis et al., 2024), visually explaining liver microsomal stability models (Long et al., 2024), and facilitating feature selection in rolling-bearing fault diagnosis (Santos et al., 2024). It has also been used in feature selection in modeling for diagnosis and clinical decisions (Kha et al., 2021).

In machine learning, deep learning approaches, particularly convolutional neural networks (CNNs), have undergone rapid development and achieved impressive performance on various tasks (Luo et al., 2024). However, it is generally accepted that effectively training CNNs requires a substantial amount of data (Xu et al., 2019; Thanapol et al., 2020). However, in many practical research scenarios, available datasets are limited in size and often imbalanced. Compared with CNNs, support vector machines (SVMs) offer distinct advantages in handling smaller datasets (Han et al., 2021; Guan et al., 2021). SVM, originally proposed by Vapnik (1995), is based on the principles of Vapnik–Chervonenkis (VC) dimension theory and Structural Risk Minimization. It achieves good generalizability by striking an optimal balance between model complexity and learning capability, even with limited samples (Li et al., 2024). SVM is also suitable for high-dimensional feature space. Khanh Le et al. (2023) created a SVM based pipeline, including $\chi 2$ and recursive feature elimination as feature selection method, LDA as dimensionality reduction method, to predict protein crystallization propensity. Another key factor contributing to the popularity of SVM is its ability to model complex non-linear relationships through the use of appropriate kernel functions (Zhang and Wang, 2011). Consequently, the selection, construction, and optimization of suitable kernel functions and associated strategies have become active research areas (Song et al., 2008; Taqvi et al., 2022). Kernel functions can be broadly categorized into global kernels, known for their strong generalizability (e.g., linear (LKF), polynomial (PKF), and sigmoid (SKF) kernels), and local kernels, recognized for strong learning ability (e.g., radial basis function (RBF) kernel). Different kernel functions possess unique characteristics suitable for different data structures (Wang and Fang, 2019). However, because the underlying features of a dataset are often unknown beforehand, selecting the most suitable kernel function can be challenging. One approach to constructing potentially superior kernel functions involves using optimization algorithms to linearly combine multiple kernel types. This strategy aims to leverage the respective advantages of different kernels, potentially leading to enhanced model performance. It has been successfully employed for hyperspectral imagery classification (Lin & Yan, 2015), asset price prediction (Zhu et al., 2022), face recognition (Hu et al., 2022), and wind speed prediction (Tian, 2020).

This study aims to construct multiclass identification models and provide practical model interpretation using preprocessed data derived from commonly used agricultural sensing techniques. To achieve this, a morphological feature dataset for rice, originally extracted from images, was obtained from previous studies (Koklu et al., 2021; Cinar and Koklu, 2022). Additionally, two imbalanced tobacco datasets were collected, and their chemical compositions were predicted from corresponding NIR spectra using previously established chemometric models (Liang et al., 2022; Guo et al., 2023). This study employs a combined kernel function in SVM

incorporating both linear and non-linear, as well as global and local kernels. Four kernel parameters and three contribution percentages within the combined kernel were optimized via particle swarm optimization (PSO) at the same time, which made this approach a self-adaptive kernel method. The resulting multiclass identification models demonstrated both strong learning and generalization capabilities. Crucially, compared to raw image or spectral data, these preprocessed features are inherently more interpretable. We employed SHAP analysis, summary and dependence plots to illustrate how different features influence class identification.

Based on the rapidly developing of the feature extraction technique and the NIR based researches of chemical composition prediction, we posit that our study—utilizing interpretable, preprocessed data derived from these techniques—is transferable and extends the applicability of image analysis and NIR spectroscopy. Moreover, models based on these interpretable preprocessed data hold significant potential for identifying key factors influencing the quality control and improvement of agricultural products.

2 Materials and methods

A study workflow is illustrated in Figure 1. In summary, attempts were made to combine images/NIR data, combined kernel SVM, PSO and SHAP. Firstly, data with straight forward meaning was extracted from original images/NIR data. Training set and test set was randomly separated. Then the Cross Validation (CV) was conducted on training set to determine the parameters and percentages of combined kernel. SVM model was trained with the optimized kernel. Multicollinearity between variables was checked before conducting SHAP. Group permutation was applied in SHAP to give final model interpretation. The details of each step are described in following sections.

2.1 Samples and data collection

The rice dataset employed in this study was obtained from previous research (Koklu et al., 2021; Cinar and Koklu, 2022) and is publicly available (https://www.muratkoklu.com/datasets/). It comprises 75,000 images of five distinct rice varieties (Arborio, Basmati, Ipsala, Jasmine, and Karacadag), with 15,000 images acquired for each variety. In this study, only the 12 morphological features, out of 106 features in original research, were used. Compared to the 90 color features, the meaning of morphological features is straightforward. The 4 shape features are a combination of certain features from the 12 morphological features.

Two distinct tobacco datasets were collected from two tobacco companies without prior sample screening based on quality attributes. In both datasets, comprising Chinese domestic tobacco samples, the class labels corresponded to eight main cultivation regions (Luo et al., 2019). Non-domestic samples were labeled according to their country of origin. The number of classes (regions or countries) exceeded ten in both datasets. Furthermore, both exhibited significant class imbalance, with disparate numbers of samples per class (details provided in Table 1).

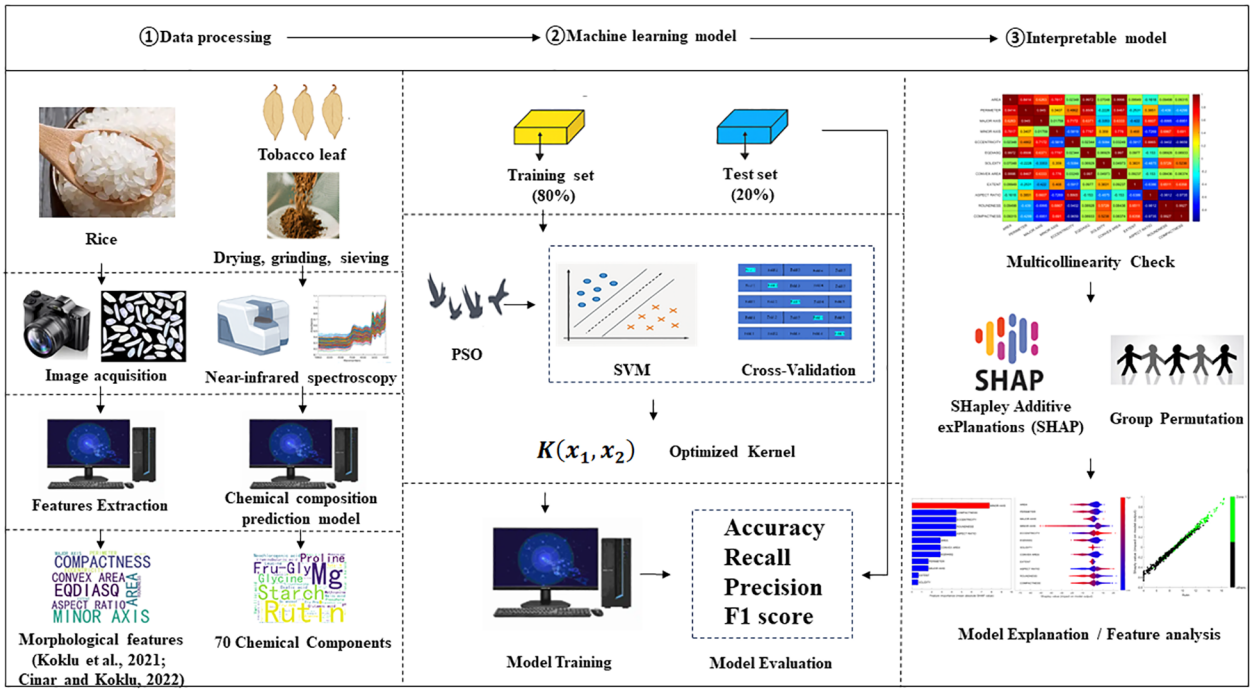

FIGURE 1
Flowchart illustrating the overall research workflow.

TABLE 1 Number of samples per cultivation region in the tobacco datasets.

Zone	Dataset 1	Dataset 2
Domestic Zone 1	363	175
Domestic Zone 2	74	40
Domestic Zone 3	8	25
Domestic Zone 4	56	164
Domestic Zone 5	139	17
Domestic Zone 6	65	31
Domestic Zone 7	7	—
Domestic Zone 8	7	22
Brazil	48	71
Zimbabwe	30	61
America	—	20
Zambia	—	12
Total	797	638

Tobacco samples were handled following the procedures described previously (Liang et al., 2022; Guo et al., 2023). All tobacco samples were dried in a drying room at 40°C for 1–3 days, ground to a certain granularity using a whirlwind grinding mill, and sieved through a 60-mesh sieve. The moisture content of the samples ranged between 6 and 8% and was analyzed by the oven-drying method. NIR spectra were recorded for all tobacco samples using an Antaris II NIR spectrophotometer (Thermo Electron Co., USA). Measurements were performed in triplicate, and each measurement comprised 64 co-added scans recorded at a resolution of 8 cm^{-1} in the wavenumber range of 4000–10000 cm^{-1}. Chemical composition data for these samples were obtained using pre-established chemometric models that predict compound concentrations from NIR spectra (Liang et al., 2022; Guo et al., 2023; Li et al., 2025). According to Liang's report, the average R^2 of routine chemicals, polyphenolic compounds, organic acids, amino acids, Amadori compounds, and other compounds for the EDM-PLS models were 0.949, 0.88, 0.862, 0.867, 0.945, and 0.891, respectively. The specific chemical compounds included in the analysis are listed in Table 2.

TABLE 2 Chemical components (n = 70) measured in the tobacco datasets.

No.	Type	Compound name	Amount
1	Routine chemicals	Total sugar, Reducing sugar, Total alkaloid, Total N, Potassium ion(K), Chloridion (Cl), Starch	7
2	Ion	Sulfate, Phosphate, Calcium(Ca), Magnesium (Mg),	4
3	Polyphenolic compounds	Neochlorogenic acid, Chlorogenic acid, Cryptochlorogenic acid, Scopoletin, Rutin,	5
4	Organic acids	Oxalic acid, Propanedioic acid, Succinic acid, Malic acid, Citric acid, Vanillic acid, Myristic acid, Palmitic acid, Oleic acid and Linolenic acid, Linoleic acid, Stearic acid, arachidic acid	12
5	Amino acid	Aspartic acid, L-Threonine, Serine, L-Asparagine, Glutamic acid, Glutamine, Glycine, Alanine, Valine, Cystine, Methionine, L-isoleucine, Leucine, Tyrosine, Phenylalanine, γ-aminobutyric acid, Lysine, Histidine, Tryptophan, Arginine, Proline	21
6	Amadori compounds	N-(1-Deoxy-d-glucose-1-yl) Ammonia (Glu-An), N-(1-deoxy-D-fructos-1-yl) aminobutyric(Fru-Amb), N-(1-deoxy-D-fructos-1-yl) Histidine(Fru-His), N-(1-deoxy-D-fructos-1-yl) Proline(Fru-Pro), N-(1-deoxy-D-fructos-1-yl) Valine(Fru-Val), N-(1-deoxy-D-fructos-1-yl) Threonine(Fru-Thr), N-(1-deoxy-D-fructos-1-yl) Glycine(Fru-Gly), N-(1-deoxy-D-fructos-1-yl) Alanine(Fru-Ala), N-(1-deoxy-D-fructos-1-yl) Asparagine (Fru-Asn), N-(1-deoxy-D-fructos-1-yl) Asparticacid(Fru-Asp), N-(1-deoxy-D-fructos-1-yl) Glutarnine(Fru-Gln), N-(1-deoxy-D-fructos-1-yl) Glutamicacid (Fru-Glu), N-(1-deoxy-D-fructos-1-yl) Isoleucine(Fru-Ile), N-(1-deoxy-D-fructos-1-yl) Leucine(Fru-Leu), N-(1-deoxy-D-fructos-1-yl) Tyrosine(Fru-Tyr), N-(1-deoxy-D-fructos-1-yl) Phenylalanine(Fru-Phe), N-(1-deoxy-D-fructos-1-yl) Tryptophan(Fru-Trp)	17
7	Others	Dichloromethane extraction, pH value, Solanesol, Neophytadiene	4
Total			70

2.2 Combined kernel and optimization

In the non-linearly separable data, SVM utilizes the kernel trick: the input data are mapped into a higher-dimensional feature space via a kernel function, where linear separation becomes feasible. The one-vs-all (OVA) approach was adopted, where M individual binary SVM classifiers are trained (M is the total number of classes), each separating one class from all the others. This choice was guided by reports suggesting potentially higher accuracy than the OVO strategy in certain contexts (Taqvi et al., 2022). A key aspect of this study is the use of a combined kernel function. Specifically, LKF, PKF, and RBF kernels (detailed in Table 3)—encompassing linear/non-linear and global/local types—are linearly combined as follows (Equation 1):

$$K(x_1, x_2) = p_1 \star x_1^T x_2 + p_2 \star (a \cdot x_1^T x_2 + b)^q + p_3 \star exp\left(\frac{-|||x_1 - x_2|||^2}{\sigma^2}\right) \quad (1)$$

The optimal kernel percentages (p_1, p_2, p_3) and kernel-specific parameters (a, b, q, σ^2) were determined via PSO. PSO is a swarm intelligence algorithm developed by Eberhart and Kennedy (1995),

TABLE 3 Kernel functions used for SVM modeling in this study.

Kernel	Formula
LKF	$K(x_1, x_2) = x_1^T x_2$
PKF	$K(x_1, x_2) = (a \cdot x_1^T x_2 + b)^q$
RBF	$K(x_1, x_2) = \exp(\frac{-\|\|x_1 - x_2\|\|^2}{\sigma^2})$

inspired by the social foraging behavior of bird flocks. Owing to its simplicity, robustness, and efficiency, PSO has undergone considerable development and found widespread application (Zhang and Teng, 2009). In PSO, a population of particles (representing potential solutions) is initialized randomly within the search space, and each is evaluated by the fitness function at its current location. During each iteration, each particle adjusts its position based on its current velocity (inertia), its own best-known positions, and the best-known position found by the entire swarm, with some random perturbations as follow (Equations 2 and 3):

$$v_i^{k+1} = w*v_i^k + c_1*r_1*(pbest_i^k - x_i^k) + c_2*r_2*(gbest_i^k - x_i^k) \quad (2)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (3)$$

where v_i^k and x_i^k represent the velocity and position of the i-th particle at iteration k. The parameters c_1 and c_2 are acceleration factors, both set to 0.8 in this study. Additionally, r_1 and r_2 are random numbers uniformly distributed in the range [0, 1]. A time-decreasing inertia weight (w) strategy was employed to balance global and local search capabilities during optimization. CV is a standard technique used to assess the generalizability of classification models (Ong et al., 2005; Koklu et al., 2021). In this study, the average classification accuracy obtained from k-fold CV (with k typically set to 5 or 10) was used as the fitness function for the PSO algorithm. This directs the optimization process towards finding kernel parameters and weightings that yield models with strong generalizability. The kernel parameters and their percentages were optimized via PSO at the same time, which made this approach a self-adaptive kernel method.

2.3 Model evaluation

In this study, the model evaluation approach varied across different datasets. For the rice dataset, ten-fold CV was applied to ensure that the results were comparable with previous research findings. For Tobacco Dataset 1, a five-fold CV was used to evaluate the models as the sample sizes for a few cultivation regions were smaller than ten. For Tobacco Dataset 2, evaluation involved randomly selecting approximately 20% of the samples as a test set, with the remaining samples constituting the training set. A five-fold CV was conducted on this training set to determine the optimal kernel parameters. Subsequently, the model was trained using the determined parameters, and the predictive accuracy on the test set was used to evaluate its performance.

2.4 Shapley value and SHAP

The Shapley value, coined by Shapley (1997), assigns payouts to features depending on their contribution to the model's prediction (total payout). It represents the average marginal contribution of a feature's value across all possible feature combinations. The Shapley value was estimated using the approximation method detailed in Algorithm 1, employing Monte-Carlo sampling as proposed by Štrumbelj and Kononenko (2014). SHAP (Lundberg and Lee, 2017), by using Shapley values, provides global interpretation methods derived from aggregations of these individual Shapley values.

For m = 1, ..., M:

- Randomly select an instance z from the data matrix X
- Generate a random permutation o of the feature indices
- Order instance x: $x_0 = (x_{(1)}, ..., x_{(j)}, ..., x_{(p)})$.
- Order instance z: $z_0 = (z_{(1)}, ..., z_{(j)}, ..., z_{(p)})$.
- Construct two new instances.
- With j: $x_{+j} = (x_{(1)}, ..., x_{(j-1)}, x_{(j)}, z_{(j+1)}, ..., z_{(p)})$.
- Without j: $x_{-j} = (x_{(1)}, ..., x_{(j-1)}, z_{(j)}, z_{(j+1)}, ..., z_{(p)})$.
- Compute marginal contribution: $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$

End for

- Compute the average Shapley value: $\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M}\phi_j^m$

Algorithm 1. Approximating the contribution of the j-th feature for model f.

Here, x is the selected instance being explained, j is the index of the feature whose contribution is being estimated, and M is the number of iterations, which was set to 300 in this study. Meanwhile, group permutation was conducted for the variables if they were highly correlated (|R|>0.8) with j.

3 Results and discussion

3.1 Rice variety identification: model construction and evaluation

Rice is one of the most widely produced and consumed cereal crops globally. The quality attributes of rice, such as cooking properties, aroma, and taste, are closely related to its variety.

TABLE 4 Classification accuracy results (%) of ten-fold CV on the rice dataset.

Variety	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Average
Basmati	97.1	97.4	96.9	98.2	97.8	98.1	97.5	97.8	97.2	97.6	97.5
Arborio	97.1	96.9	96.9	96.5	96.0	96.8	96.9	97.2	96.5	96.6	96.7
Jasmine	98.0	98.3	98.6	98.3	97.7	98.0	98.4	97.9	98.1	98.5	98.2
Ipsala	99.4	99.6	99.2	99.7	99.3	99.4	99.5	99.7	99.3	99.6	99.5
Karacadag	98.4	97.4	97.4	96.7	98.5	98.1	97.4	98.1	97.7	97.8	97.8
Total	98.0	97.9	97.8	97.9	97.9	98.1	97.9	98.1	97.8	98.0	97.9

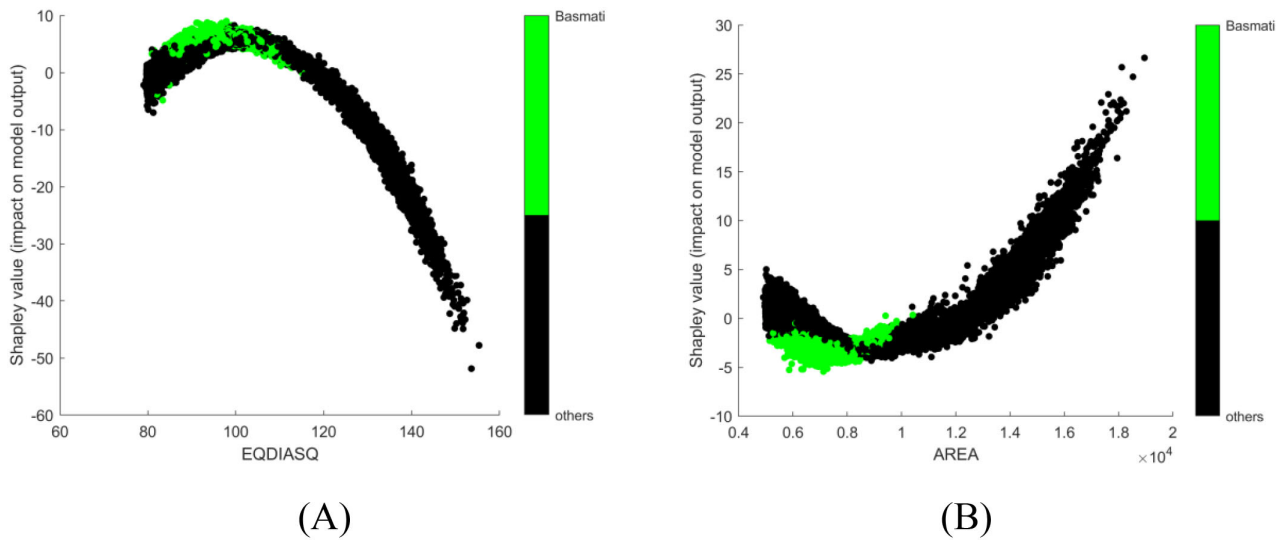

FIGURE 2
SHAP dependence plots, no corrective measures are implemented for multicollinearity: **(A)** EQDIASQ; **(B)** AREA.

TABLE 5 Mean values of morphological variables for the five rice varieties.

Variable	Arborio	Basmati	Ipsala	Jasmine	Karacadag
AREA	7531.717	7563.938	14048.645	6267.308	6484.379
PERIMETER	339.852	426.906	476.498	347.781	299.810
MAJOR AXIS	137.585	202.336	197.071	157.076	114.959
MINOR AXIS	70.459	48.494	91.817	50.951	72.426
ECCENTRICITY	0.857	0.970	0.884	0.945	0.774
EQDIASQ	97.790	97.975	133.549	88.545	90.797
SOLIDITY	0.977	0.970	0.977	0.972	0.983
CONVEX AREA	7712.890	7797.524	14373.349	6442.758	6597.790
EXTENT	0.683	0.504	0.663	0.590	0.726
ASPECT RATIO	1.958	4.194	2.153	3.089	1.591
ROUNDNESS	0.818	0.522	0.776	0.642	0.906
COMPACTNESS	0.711	0.485	0.678	0.565	0.791

No units were provided in the original data extracted from images.

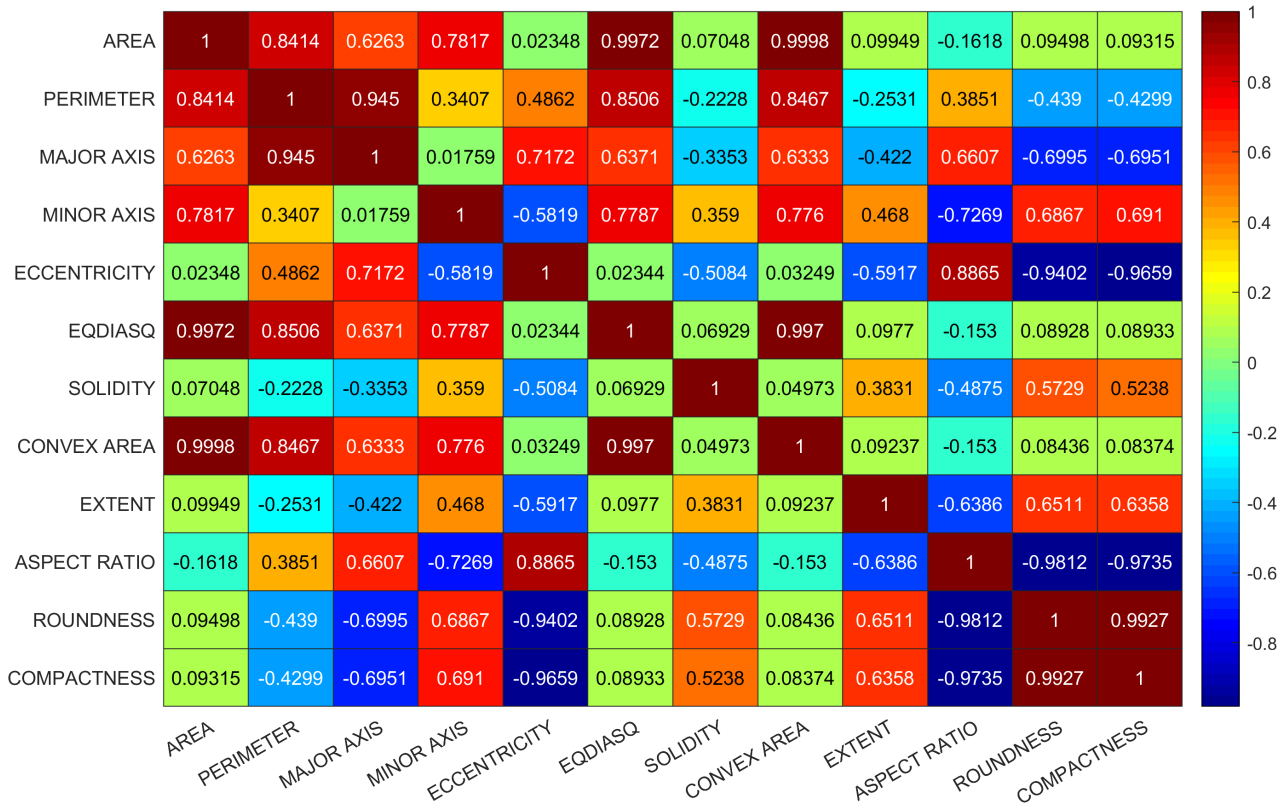

FIGURE 3
Heatmap of correlation with correlation coefficient between variables.

First, a linear SVM was used to construct an identification model, achieving an average total accuracy of 97.3% in ten-fold CV (Supporting Information Table S1). Subsequently, the PKF was evaluated, with the "PolynomialOrder" parameter set to 2 within the "fitcsvm" function. This approach yielded an improved average total accuracy of 97.9% (Table 4), and the accuracy for each variety also improved. Considering that only 12 morphological features were used in this study, as opposed to the 106 features

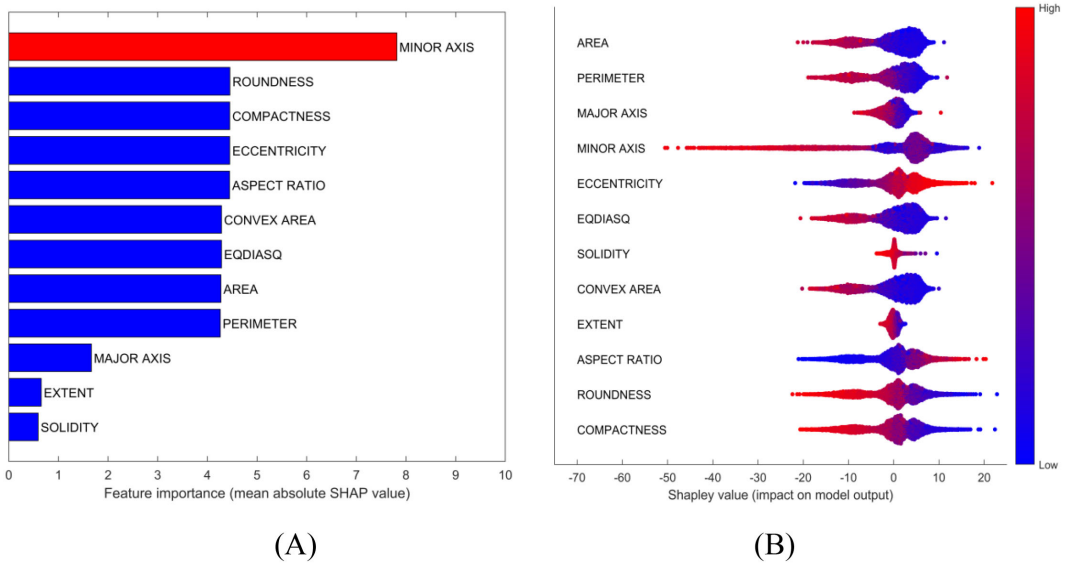

FIGURE 4
Model interpretation plots for the rice dataset: **(A)** Feature importance ranking based on mean absolute SHAP values; **(B)** SHAP summary plot illustrating feature impacts, where red and blue points correspond to high and low feature values, respectively.

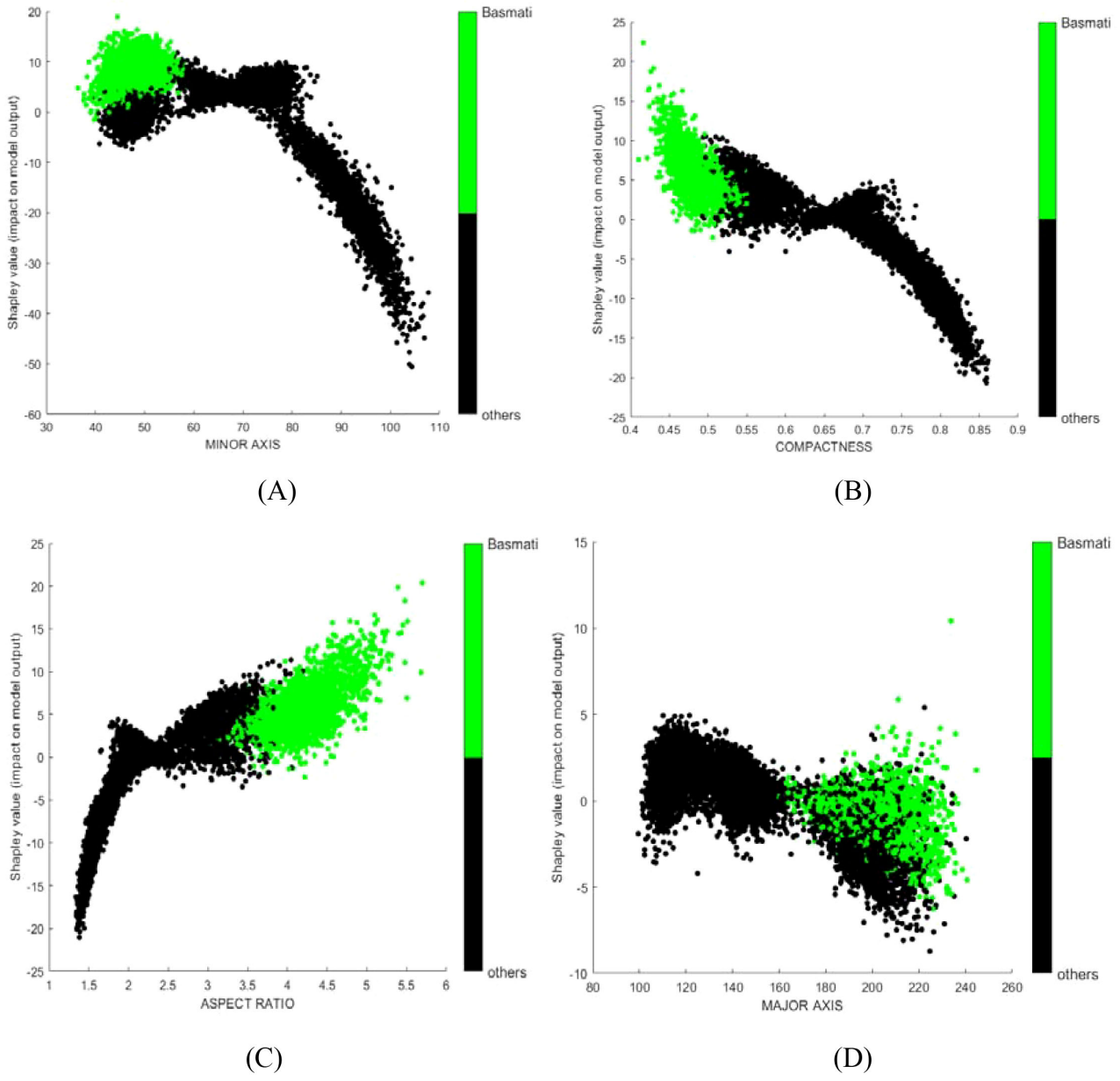

FIGURE 5
SHAP dependence plots for selected features from the rice dataset model: **(A)** MINOR AXIS; **(B)** ROUNDNESS; **(C)** ASPECT RATIO; **(D)** MAJOR AXIS. Green points represent samples of the target class; black points represent samples from other classes.

employed in the original research, this performance level is considered acceptable.

3.2 Rice variety identification: model interpretation

Cinar and Koklu (2021) previously conducted variable analysis on this dataset using analysis of variance, chi-squared test, and gain ratio, providing an importance order for effective features. Although such statistical information is valuable, it does not directly interpret the classification model itself. This study applied the SHAP approach, using one OVA model (Basmati vs. others) as an illustrative example.

In the first trial, no corrective measures are implemented in SHAP to deal with multicollinearity issue and obvious conflict was observed in interpretation. For instance, EQDIASQ exhibited a quadratic relationship between the variable value and the corresponding Shapley value (Figure 2A). The mean EQDIASQ values of Ipsala, Jasmine, and Karacadag were either higher or lower than those of Basmati and Arborio (Table 5). It is easy to accept the interpretation that the model would tent to output lower Shapley values (indicating less support for Basmati) if the EQDIASQ value is excessively high or low relative to the typical Basmati range.

TABLE 6 Average classification accuracies (%) of five-fold CV on tobacco dataset 1.

Zone label	LKF	PKF	RBF	SKF	Combined kernel SVM
Domestic Zone 1	98.1	99.2	99.5	99.2	99.7
Domestic Zone 2	84.0	86.7	73.1	25.8	88.0
Domestic Zone 3	40.0	80.0	50.0	0.0	80.0
Domestic Zone 4	94.8	98.3	96.7	96.5	98.3
Domestic Zone 5	98.6	98.6	97.9	95.0	98.6
Domestic Zone 6	95.4	93.8	93.8	90.8	95.4
Domestic Zone 7	100.0	100.0	100.0	80.0	100.0
Domestic Zone 8	90.0	90.0	70.0	70.0	90.0
Brazil	94.0	96.0	92.0	96.0	96.0
Zimbabwe	96.7	100.0	100.0	100.0	100.0
Total Accuracy	95.4	96.9	94.9	89.2	97.4
Note: Kernel parameter	—	a = 1/65, b = 1.5, q = 3	σ^2 = 32	a = 1/65, b = -1.2	In Table 7

However, Area, which is highly correlated to EQDIASQ (Figure 3, heatmap of correlation with correlation coefficient between variables), showed an opposite quadratic relationship (Figure 2B). A similar pattern was observed in the dependence plot for CONVEX AREA. This counterintuitive interpretation could be caused by high correlations between those variables.

Potential solutions to address interpretation issues arising from multicollinearity include permuting correlated features together to obtain a mutual Shapley value or employing conditional sampling (Molnar, 2024). Therefore, group permutation was implemented in this study, which means that the highly correlated variables will be permuted together. SHAP uses mean absolute Shapley values to evaluate variable importance (Figure 4A). For each variables, its absolute Shepley value from all the samples will averaged to give the length of bar in figure. The SHAP summary plot gives global view of contributions of the variables (Figure 4B). In the summary plot, each variable is represented by a dotted line along the horizontal axis. Red dots indicate high values of the variable in a given sample, whereas blue dots represent low values. A higher Shapley value signifies that the variable makes a positive contribution toward classifying the sample into the target class (Basmati in this example), whereas a lower value indicates a contribution towards classifying it as one of the other varieties. From the perspective of the constructed model, MINOR AXIS was identified as the most important variables for identifying Basmati rice, and it showed a negative contribution towards the Basmati classification. The basically same importance and contribution were signed to the correlated variables. ROUNDNESS, COMPACTNESS, ECCENTRICITY and ASPECT RATIO were signed as the second important variables as a group.

SHAP dependence plots were used to further investigate how individual variables affect the identification outcome. For instance, MINOR AXIS consistently made a negative contribution in the model (Figure 5A), which aligns with the observation that the mean value of MINOR AXIS for Basmati is the lowest among all rice varieties considered (Table 5, mean value of each variables of each class). COMPACTNESS and ASPECT RATIO showed negative and positive contribution (Figures 5B, C), respectively, which also aligns with their mean value. Some variables, like MAJOR AXIS showed very limited contribution (Figure 5D).

TABLE 7 Optimized parameters for the combined kernel function obtained via PSO.

Kernel percentage (%)			Kernel parameters			
			PKF			RBF
LKF	PKF	RBF	a	b	q	σ^2
0.0	44.5	55.5	65	1.5	3	32

3.3 Cultivation region identification: model construction and evaluation

The quality of many agricultural products, including fruits, mushrooms, tobacco, and traditional Chinese medicines, is significantly affected by their cultivation regions (Li et al., 2018; Wei et al., 2018; Kim et al., 2020; Jiang et al., 2020; Li et al., 2022).

Given that the tobacco datasets were limited in size, imbalanced, and included multiple region classes, linear methods were evaluated first. Stepwise Fisher linear discriminant analysis (LDA) resulted in serious overfitting for Domestic Zones 3, 7, and 8. By contrast, linear SVM showed higher or similar CV accuracy, implying better generalizability in this context (Supporting Information Table S2, complete CVs in Supplementary Tables S3, S4). LDA relies heavily on variance and scatter matrices (Ghojogh et al., 2022), and small sample sizes may not provide stable estimations for these parameters.

Because potential non-linear relationships between the chemical variables and region classes might exist, kernel SVM models were

TABLE 8 Standard deviation and confidence intervals of recall and total accuracy.

Cultivation region	Repeat 1	Repeat 2	Repeat 3	Repeat 4	Repeat 5	Repeat 6	Repeat 7	Repeat 8	Repeat 9	Repeat 10	Average	Standard deviation	Confidence intervals (95%)
Domestic Zone 1	100.0%	100.0%	97.4%	100.0%	100.0%	94.9%	94.6%	100.0%	100.0%	97.6%	98.5%	2.2%	98.5% ± 1.4%
Domestic Zone 2	87.5%	85.7%	100.0%	90.0%	44.4%	83.3%	100.0%	66.7%	60.0%	100.0%	81.8%	18.9%	91.8% ± 11.7%
Domestic Zone 3	80.0%	100.0%	100.0%	100.0%	66.7%	100.0%	100.0%	80.0%	100.0%	75.0%	90.2%	13.2%	90.2% ± 8.2%
Domestic Zone 4	97.0%	100.0%	97.1%	96.4%	97.1%	100.0%	94.9%	100.0%	97.1%	100.0%	98.0%	1.9%	98.0% ± 1.2%
Domestic Zone 5	100.0%	100.0%	100.0%	80.0%	100.0%	71.4%	100.0%	100.0%	50.0%	100.0%	90.1%	17.5%	90.1% ± 10.8%
Domestic Zone 6	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	100.0% ± 0%
Domestic Zone 8	100.0%	80.0%	100.0%	100.0%	100.0%	66.7%	100.0%	100.0%	100.0%	83.3%	93.0%	12.0%	93.0% ± 7.4%
Brazil	100.0%	94.1%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.4%	1.9%	99.4% ± 1.2%
Zimbabwe	100.0%	100.0%	100.0%	100.0%	100.0%	90.0%	92.9%	100.0%	100.0%	100.0%	98.3%	3.7%	98.3% ± 2.3%
America	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	100.0% ± 0%
Zambia	100.0%	66.7%	NaN	50.0%	75.0%	66.7%	50.0%	100.0%	100.0%	100.0%	78.7%	21.7%	78.7% ± 14.2%
Total Accuracy	97.7%	96.9%	98.4%	96.3%	93.7%	93.7%	95.4%	96.3%	96.6%	97.8%	96.3%	1.6%	96.3% ± 1.0%

NaN means no sample was selected as test set of this zone due to the random sampling.

10.3389/fpls.2025.1597673

10.3389/fpls.2025.1597673

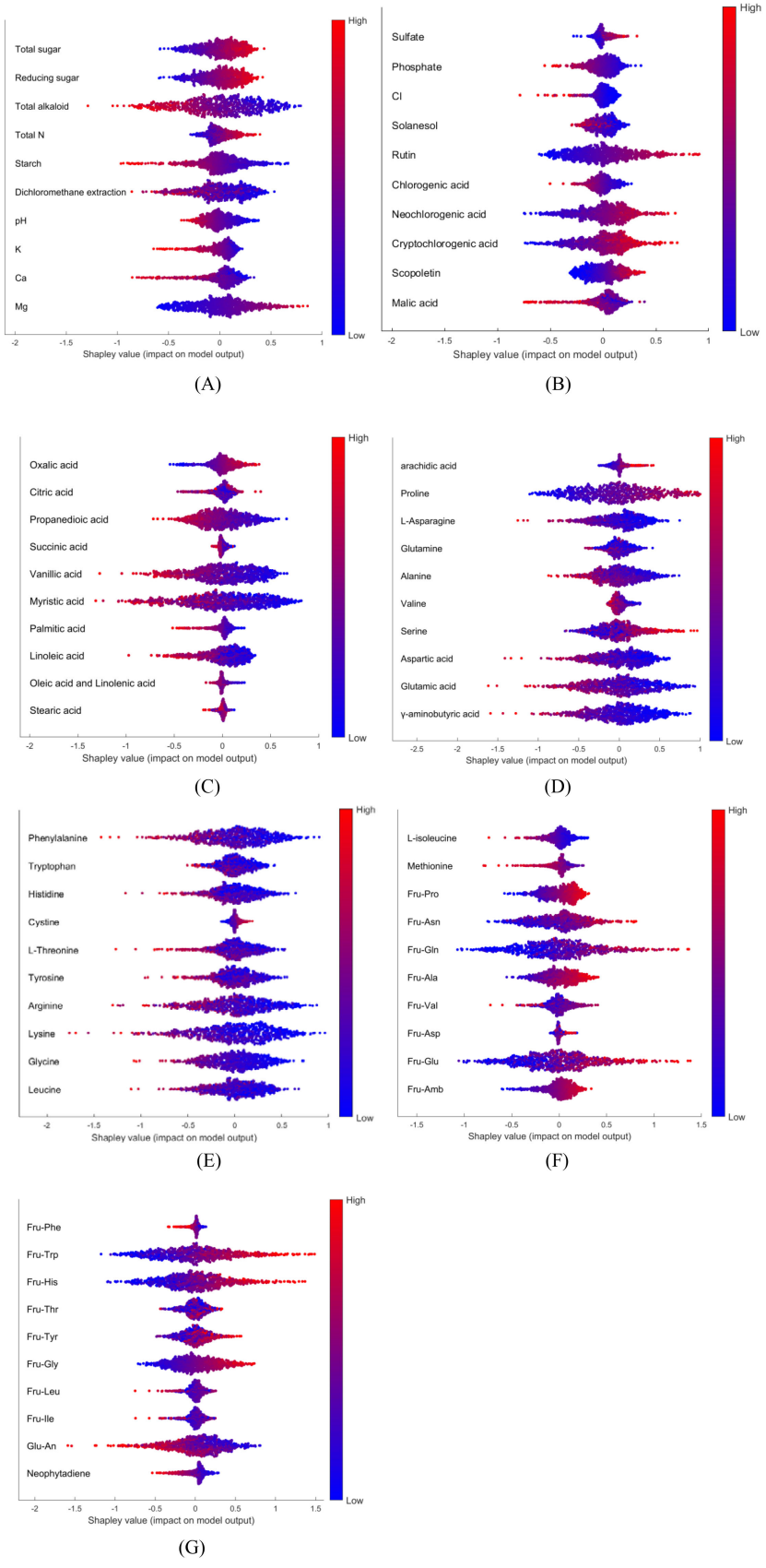

FIGURE 6
SHAP summary plot for all 70 compounds, Red and blue points correspond to high and low chemical levels, respectively.: **(A)** part1; **(B)** part2; **(C)** part3; **(D)** part4; **(E)** part5; **(F)** part6; **(G)** part7.

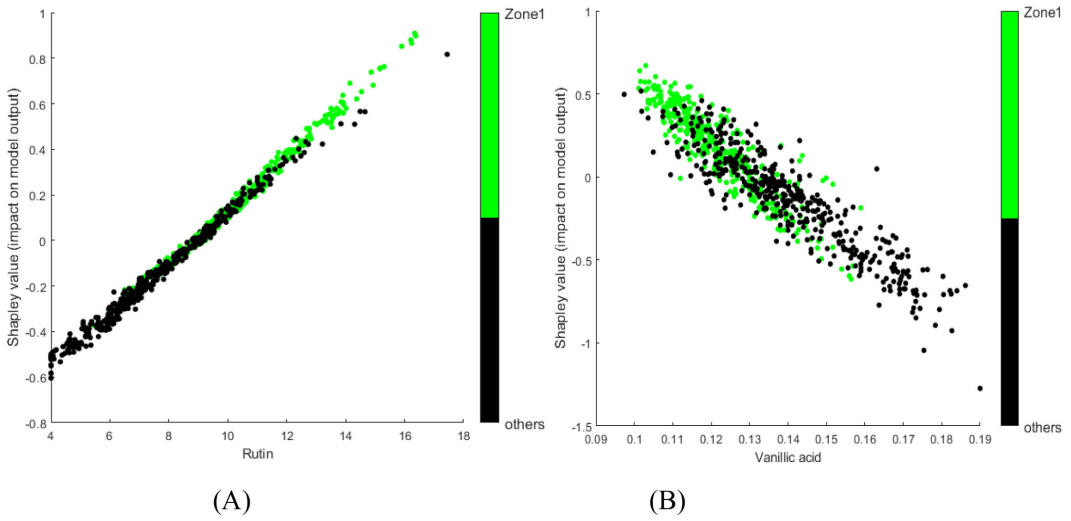

FIGURE 7
SHAP dependence plots for selected chemical components from the tobacco dataset model: **(A)** rutin; **(B)** vanillic acid. Green points represent samples from the target class; black points represent samples from other classes.

subsequently established. The kernel function parameters for each kernel type were set manually through multiple trials, and the best results are listed in Supplementary Tables S5–S7. The PKF generally demonstrated better generalizability than the other kernel functions tested. Notably, satisfactory average accuracies were achieved even for groups with small sample sizes. Following these initial explorations, the combined kernel approach was implemented. A performance comparison of combined kernel SVM against models using single, manually tuned kernel functions is presented in Table 6. The kernel parameters optimized by PSO are listed in Table 7. Combined kernel SVM achieved the highest, or equally highest, average CV accuracies for all individual regions and the highest overall accuracy. The confusion matrix, including F1-score, precision, recall, and five-fold CV results for combined kernel SVM are provided in Supporting Information (Supplementary Tables S8, S9). The over-sampling was also conducted to fix the imbalance issue. However, limited improvement was achieved in LDA and even lower accuracy was achieved in SVM (details and results in Supporting Information, Table S11). The paired samples t-test of different methods can be found in Supporting Information (Supplementary Table S12).When comparing with other methods, the combined kernel SVM without over-sampling showed significant difference in most case.

To further evaluate the combined kernel SVM method, an independent test set was randomly selected from Tobacco Dataset 2. The combined kernel SVM approach, including parameter optimization via CV on the training set, was applied to the training set. The resulting model was then used to predict the classes of the test set samples. The results are presented in Supporting Information (Supplementary Tables S13–S15). The model demonstrated a high accuracy of 97.7% on the independent test set and F1-score range for each class is from 0.923 to 1. To further evaluate the robustness of model, 10 times repeating experiments was conducted to give Standard Deviation (SD) and Confidence Intervals (CI). For each trial, training set and independent test set was randomly separated. The approach mentioned above was re-conducted. The SD and CI of recall and total accuracy were listed in the Table 8. Generally, the model showed good accuracy and rubustness. 5 zones showed standard deviations more the 10%. 3 zones showed confidence intervals more than ±10%. The possible reason is the limited amount of samples. Therefore, it is necessary to keep collecting samples in future work. An extra dataset was collected as external test data. The Total accuracy is 95.2% and the confusion matrix, including F1-score, precision, recall, was presented in Supporting Information (Supplementary Table S16).

3.4 Cultivation region identification: model interpretation

One OVA model (Domestic Zone 1 vs. others), constructed using Tobacco Dataset 1, was analyzed using the SHAP approach. Group permutation was also applied in this case (Full Heatmap of 70 compounds in Supporting Information, Supplementary Figure S1). Figures 6A–G presents the global SHAP analysis for selected compounds relevant to identifying samples from Domestic Zone 1 versus other regions. As depicted in Figure 6, certain compounds exhibit a positive contribution towards identifying Domestic Zone 1 samples, such as magnesium (Mg) and Oxalic acid. Conversely, some compounds show a negative contribution, such as starch and vanillic acid. The overlap of red and blue dots for some variables, such as succinic acid and stearic acid (Figure 6C), indicates that these features did not consistently contribute either positively or negatively to the model's output across all samples.

To further interpret the model, SHAP dependence plots were generated to illustrate the influence of chemical levels on the

10.3389/fpls.2025.1597673

TABLE 9 Chemical components identified as most important for distinguishing domestic zone 1 samples based on SHAP analysis.

No.	Compound	Slope	R^2	Absolute slope value	Positive (+) or negative (-) contribution
1	Proline	2.452	0.885	2.452	+
2	γ-aminobutyric acid	-1.939	0.825	1.939	–
3	Mg	1.779	0.980	1.779	+
4	Fru-Trp	1.765	0.861	1.765	+
5	Glycine	-1.633	0.828	1.633	–
6	Fru-His	1.572	0.826	1.572	+
7	Total alkaloid	-1.566	0.911	1.566	–
8	Rutin	1.529	0.992	1.529	+
9	Vanillic acid	-1.529	0.864	1.529	–
10	Starch	-1.396	0.963	1.396	–
11	Fru-Gly	1.364	0.973	1.364	+
12	Neochlorogenic acid	1.343	0.933	1.343	+
13	Propanedioic acid	-1.277	0.898	1.277	–
14	Cryptochlorogenic acid	1.108	0.891	1.108	+
15	Total sugar	0.939	0.965	0.939	+
16	Reducing sugar	0.928	0.978	0.928	+
17	Fru-Amb	0.911	0.973	0.911	+
18	Fru-Ala	0.897	0.881	0.897	+
19	Oxalic acid	0.824	0.936	0.824	+
20	Phosphate	-0.808	0.978	0.808	–
21	pH	-0.739	0.968	0.739	–
22	L-isoleucine	-0.736	0.896	0.736	–
23	Fru-Pro	0.706	0.958	0.706	+
24	K	-0.694	0.912	0.694	–
25	Scopoletin	0.640	0.942	0.640	+
26	Chlorogenic acid	-0.595	0.934	0.595	–
27	Cl	-0.563	0.905	0.563	–
28	Total N	0.562	0.885	0.562	+
29	Palmitic acid	-0.547	0.878	0.547	–
30	Ca	-1.076	0.791	1.076	N/A
31	Neophytadiene	-0.566	0.771	0.566	N/A
32	arachidic acid	0.451	0.753	0.451	N/A
33	Linoleic acid	-0.828	0.748	0.828	N/A
34	Myristic acid	-1.655	0.736	1.655	N/A
35	L-Asparagine	-1.526	0.724	1.526	N/A
36	Cystine	0.194	0.723	0.194	N/A
37	Fru-Gln	1.595	0.722	1.595	N/A
38	Sulfate	0.282	0.706	0.282	N/A

(Continued)

TABLE 9 Continued

No.	Compound	Slope	R^2	Absolute slope value	Positive (+) or negative (-) contribution
39	Alanine	-1.215	0.682	1.215	N/A
40	Aspartic acid	-1.397	0.673	1.397	N/A
41	Glutamic acid	-1.680	0.671	1.680	N/A
42	Methionine	-0.748	0.669	0.748	N/A
43	Glu-An	-1.312	0.666	1.312	N/A
44	Fru-Glu	1.388	0.659	1.388	N/A
45	Fru-Asn	1.116	0.653	1.116	N/A
46	Lysine	-1.812	0.652	1.812	N/A
47	Phenylalanine	-1.424	0.644	1.424	N/A
48	Fru-Phe	-0.216	0.619	0.216	N/A
49	Arginine	-1.362	0.599	1.362	N/A
50	Dichloromethane extraction	-1.017	0.590	1.017	N/A
51	Malic acid	-0.682	0.576	0.682	N/A
52	Solanesol	-0.366	0.562	0.366	N/A
53	Valine	-0.350	0.539	0.350	N/A
54	Serine	0.870	0.530	0.870	N/A
55	Succinic acid	-0.106	0.443	0.106	N/A
56	Leucine	-1.051	0.440	1.051	N/A
57	L-Threonine	-0.787	0.427	0.787	N/A
58	Histidine	-0.806	0.423	0.806	N/A
59	Tyrosine	-0.666	0.351	0.666	N/A
60	Fru-Tyr	0.397	0.339	0.397	N/A
61	Oleic acid and Linolenic acid	-0.127	0.237	0.127	N/A
62	Tryptophan	-0.324	0.180	0.324	N/A
63	Glutamine	-0.258	0.137	0.258	N/A
64	Citric acid	-0.184	0.104	0.184	N/A
65	Fru-Asp	0.049	0.065	0.049	N/A
66	Fru-Val	0.175	0.042	0.175	N/A
67	Fru-Thr	0.074	0.022	0.074	N/A
68	Stearic acid	-0.008	0.001	0.008	N/A
69	Fru-Leu	0.012	0.000	0.012	N/A
70	Fru-Ile	-0.003	0.000	0.003	N/A

identification. Figures 7A, B show the SHAP dependence plots for rutin and vanillic acid, respectively, overlaid with sample classification information. Green dots represent samples from Domestic Zone 1, whereas black dots represent those from other regions. For a single variable, overlap between samples from different classes can often be observed in the middle range of chemical values, regardless of whether the overall contribution is positive or negative. However, clear trends can be observed. With increasing levels of rutin or decreasing levels of vanillic acid, a greater proportion of samples were identified as belonging to

Domestic Zone 1 rather than other regions. Some previous reports provide corroborating evidence from a different perspective. The content levels of rutin, total sugar, and total N are fairly high, while the K levels are relatively low, in samples from Domestic Zone 1 (Luo et al., 2019). These chemical components showed correspondingly positive (rutin, total sugar, total N) or negative (K) contributions in our SHAP analysis (Figures 6A, B). Similarly, Wang et al. (2022) found that the chemical components Fru-Pro, Fru-Gln, and Fru-His are typically high in samples from Domestic Zone 1, and these components also showed clear positive contributions in our analysis (Figures 6F, G). Compared to traditional methods of analyzing variable differences between groups, this model interpretation approach is arguably more efficient, yielding detailed and straightforward insights into feature contributions.

Remarkably, many variables displayed a strong linear relationship between their value and their corresponding Shapley value (contribution). The dependence plots of variables exhibiting a coefficient of determination (R^2) lower than 0.8 between the variable and Shapley values were manually verified. None of these showed clear relationship of other kind such as quadratic relationship. This observation suggests that the absolute value of the slope formed by the points in a SHAP dependence plot could serve as an alternative measure of variable importance, particularly when a clear linear trend exists. Based on this assumption, we identified the important compounds for distinguishing between samples from Domestic Zone 1 and other samples using this slope-based metric (Table 9). The compounds with R^2 higher than 0.8 was listed in row 1 to 29, ordered by absolute slope value from high to low. The compounds with R^2 lower than 0.8 was listed in row 30 to 70 and their positive/negative contribution was labeled as N/A. Considering the potentially imbalanced distribution of Shapley values across a variable's range, this slope-based method might offer a fairer assessment of importance than the mean absolute Shapley value for certain cases.

4 Conclusions

This study employed preprocessed data—specifically, morphological features extracted from images and chemical component data predicted from NIR spectra—as inputs to construct multiclass identification models. The proposed combined kernel SVM model demonstrated high accuracy and robustness. In contrast to some previous studies, practical model interpretation was achieved by applying SHAP to the models constructed with these preprocessed data types. The detailed contributions of individual variables were clarified using SHAP summary and dependence plots. Furthermore, the analysis suggested that the absolute value of the slope observed in SHAP dependence plots shows potential as an alternative metric for evaluating variable importance. These results indicate that accurate and robust models can be constructed from imbalanced, preprocessed data using the PSO optimized combined kernel SVM, while simultaneously allowing for practical model interpretation to provide detailed variable analysis. This approach broadens the application scope of image and NIR spectrum data. Looking forward, this methodology is considered transferable and applicable for exploring key variables related to the quality and characteristics of various agricultural products.

Data availability statement

Data and Matlab code are available at github (https://github.com/cvn77andf1/Tobacco-data-and-cod).

Author contributions

CW: Methodology, Visualization, Formal analysis, Investigation, Writing – original draft. YF: Methodology, Visualization, Investigation, Writing – review & editing, Writing – original draft. RW: Conceptualization, Writing – review & editing, Software, Visualization. LZ: Resources, Project administration, Methodology, Supervision, Writing – review & editing. HW: Resources, Project administration, Methodology, Writing – review & editing, Supervision. JG: Software, Resources, Project administration, Writing – review & editing. QL: Writing – review & editing, Resources, Validation. SL: Writing – review & editing, Resources, Formal Analysis. SM: Formal Analysis, Writing – review & editing, Resources. ZW: Validation, Writing – review & editing, Funding acquisition. WH: Validation, Writing – review & editing, Resources. HL: Writing – review & editing, Investigation, Supervision. SY: Supervision, Funding acquisition, Writing – review & editing, Project administration. CN: Supervision, Writing – review & editing, Project administration, Funding acquisition.

Funding

The author(s) declare that no financial support was received for the research, and/or publication of this article.

Conflict of interest

Authors YF, QL, SM were employed by the company Technology Center, China Tobacco Henan Industrial Co., Ltd. and authors SL, ZW, WH were employed by the company Technology Center, China Tobacco Gansu Industrial Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1597673/full#supplementary-material

References

Abekasis, D., Sadka, A., Rokach, L., Shiff, S., Morozov, M., Kamara, I., et al. (2024). Explainable machine learning for revealing causes of citrus fruit cracking on a regional scale. *Precis. Agric.* 25, 589–613. doi: 10.1007/s11119-023-10084-y

Antolínez García, A., and Cáceres Campana, J. W. (2023). Identification of pathogens in corn using near-infrared UAV imagery and deep learning. *Precis. Agric.* 24, 783–806. doi: 10.1007/s11119-022-09951-x

Chen, B., Li, N., and Bao, W. (2025). CLPr_in_ML: cleft lip and palate reconstructed features with machine learning. *Curr. Bioinf.* 20, 179–193. doi: 10.2174/0115748936330499240909082529

Cinar, I., and Koklu, M. (2019). Classification of rice varieties using artificial intelligence methods. *Int. J. Intell. Syst. Appl. Eng.* 7, 188–194. doi: 10.18201/ijisae.2019355381

Cinar, I., and Koklu, M. (2021). Determination of effective and specific physical features of rice varieties by computer vision in exterior quality inspection. *Selcuk. J. Agric. Food Sci.* 35, 229–243. doi: 10.15316/SJAFS.2021.252

Cinar, I., and Koklu, M. (2022). Identification of rice varieties using machine learning algorithms. *J. Agric. Sci.* Tarim Bilimleri Dergisi, 28 (2), 307–325. doi: 10.15832/ankutbd.862482

Eberhart, R., and Kennedy, J. (1995). "A new optimizer using particle swarm theory," in *Presented at Proceedings of the Sixth International Symposium on Micro Machine and Human Science* (Nagora, Japan). doi: 10.1109/MHS.1995.494215

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 177. Available online at: https://www.jmlr.org/papers/volume20/18-760/18-760.pdf (Accessed December 10, 2024).

Fita, D., Rubio, C., Franch, B., Castiñeira-Ibáñez, S., Tarrazó-Serrano, D., and San Bautista, A. (2025). Improving harvester yield maps postprocessing leveraging remote sensing data in rice crop. *Precis. Agric.* 26, 1573–1618. doi: 10.1007/s11119-025-10219-3

Ghojogh, B., Karray, F., and Crowley, M. (2022). *Fisher and Kernel Fisher Discriminant Analysis: Tutorial.* ArXiv, doi: 10.48550/arXiv.1906.09436

Guan, S., Wang, X., Hua, L., and Li, L. (2021). Quantitative ultrasonic testing for near-surface defects of large ring forgings using feature extraction and GA-SVM. *Appl. Acoust.* 173, 107714. doi: 10.1016/j.apacoust.2020.107714

Guo, J., Zhao, L., Liang, Y., Wang, D., Shang, P., Li, H., et al. (2023). Moisture-adaptive corrections of NIR for the rapid simultaneous analysis of 70 chemicals in tobacco: A case study on tobacco. *Microchem. J.* 189, 108522. doi: 10.1016/j.microc.2023.108522

Han, T., Zhang, L., Yin, Z., and Tan, A. C. C. (2021). Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine. *Measurement* 177, 109022. doi: 10.1016/j.measurement.2021.109022

Hu, L., Chen, X., Guo, G., and Chen, L. (2022). Nonnegative matrix factorization with combined kernels for small data representation. *Expert Syst. Appl.* 208, 118155. doi: 10.1016/j.eswa.2022.118155

Jamwal, R., Amit,, Kumari, S., Sharma, S., Kelly, S., Cannavan, A., et al. (2021). Recent trends in the use of FTIR spectroscopy integrated with chemometrics for the detection of edible oil adulteration. *Vibrational. Spectrosc.* 113, 103222. doi: 10.1016/j.vibspec.2021.103222

Jiang, D., Qi, G., Hu, G., Mazur, N., Zhu, Z., and Wang, D. (2020). A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors. *Infrared. Phys. Technol.* 111, 103494. doi: 10.1016/j.infrared.2020.103494

Joshi, P., Sandhu, K. S., Singh Dhillon, G. S., Chen, J., and Bohara, K. (2024). Detection and monitoring wheat diseases using unmanned aerial vehicles (UAVs). *Comput. Electron. Agric.* 224, 109158. doi: 10.1016/j.compag.2024.109158

Kha, Q.-H., Le, V.-H., Hung, T. N. K., and Le, N. Q. K. (2021). Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q co-deletion in lower-grade gliomas. *Cancers* 13, 5398. doi: 10.3390/cancers13215398

Khanh Le, N. Q., Li, W., and Cao, Y. (2023). Sequence-based prediction model of protein crystallization propensity using machine learning and two-level feature selection. *Briefings Bioinf.* 24, 1–8. doi: 10.1093/bib/bbad319

Kim, D. S., Choi, M. H., and Shin, H. J. (2020). Extracts of *Moringa oleifera* leaves from different cultivation regions show both antioxidant and antiobesity activities. *J. Food Biochem.* 44, e13282. doi: 10.1111/jfbc.13282

Koklu, M., Cinar, I., and Taspinar, Y. S. (2021). Classification of rice varieties with deep learning methods. *Comput. Electron. Agric.* 187, 106285. doi: 10.1016/j.compag.2021.106285

Li, B., Li, W., Guo, J., Wang., H., Wan, R., Liu, Y., et al. (2025). Outlier removal with weight penalization and aggregation: A robust variable selection method for enhancing near-infrared spectral analysis performance. *Anal. Chem* 97, 7325–7332. doi: 10.1021/acs.analchem.4c07007

Li, P., Jiang, Z., Sun, T., Wang, C., Chen, Y., Yang, Z., et al. (2018). Comparison of structural, antioxidant and immuno-stimulating activities of polysaccharides from *Tremella fuciformis* in two different regions of China. *Int. J. Food Sci. Technol.* 53, 1942–1953. doi: 10.1111/ijfs.13782

Li, Q., Yang, S., Li, B., Zhang, C., Li, Y., and Li, J. (2022). Exploring critical metabolites of honey peach (*Prunus persica* (L.) Batsch) from five main cultivation regions in the north of China by UPLC-Q-TOF/MS combined with chemometrics and modeling. *Food Res. Int.* 157, 111213. doi: 10.1016/j.foodres.2022.111213

Li, Z., Deng, S., Hong, Y., Wei, Z., and Cai, L. (2024). A novel hybrid CNN–SVM method for lithology identification in shale reservoirs based on logging measurements. *J. Appl. Geophys.* 223, 105346. doi: 10.1016/j.jappgeo.2024.105346

Liang, Y., Zhao, L., Guo, J., Wang, H., Liu, S., Wang, L., et al. (2022). Just-in-time learning-integrated partial least-squares strategy for accurately predicting 71 chemical constituents in Chinese tobacco by near-infrared spectroscopy. *ACS Omega.* 7, 38650–38659. doi: 10.1021/acsomega.2c04139

Lin, Z., and Yan, L. (2015). A support vector machine classifier based on a new kernel function model for hyperspectral data. *GISci. Remote Sens.* 53, 85–101. doi: 10.1080/15481603.2015.1114199

Long, T. Z., Jiang, D. J., Shi, S. H., Deng, Y. C., Wang, W. X., and Cao, D. S. (2024). Enhancing multi-species liver microsomal stability prediction through artificial intelligence. *J. Chem. Inf. Model.* 64, 3222–3236. doi: 10.1021/acs.jcim.4c00159

Lundberg, S., and Lee, S. I. (2017). "A unified approach to interpreting model predictions," in *31st conference on Neural Information Processing Systems (NIPS).* doi: 10.5555/3295222.3295230

Luo, D., Wang, B., and Qiao, X. (2019). Explanation of national regionalization of leaves style of flue-cured tobacco. *Acta Tabacaria. Sin.* 25, 4. doi: 10.16472/j.Chinatobacco.2019.218

Luo, N., Xu, D., Xing, B., Yang, X., and Sun, C. (2024). Principles and applications of convolutional neural network for spectral analysis in food quality evaluation: A review. *J. Food Compos. Anal.* 128, 105996. doi: 10.1016/j.jfca.2024.105996

Molnar, C. (2024). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Available online at: https://christophm.github.io/interpretable-ml-book/.

Naeimi, M., Daggupati, P., and Biswas, A. (2024). Image-based soil characterization: A review on smartphone applications. *Comput. Electron. Agric.* 227, 109502. doi: 10.1016/j.compag.2024.109502

Ong, C., Smola, A., and Williamson, R. C. (2005). Learning the kernel with hyperkernels. *J. Mach. Learn. Res.* 6, 1043–1071. Available online at: https://www.jmlr.org/papers/volume6/ong05a/ong05a.pdf.

Rawal, A., Hartemink, A., Zhang, Y., Wang, Y., Lankau, R. A., and Ruark, M. D. (2024). Visible and near-infrared spectroscopy predicted leaf nitrogen contents of

potato varieties under different growth and management conditions. *Precis. Agric.* 25, 751–770. doi: 10.1007/s11119-023-10091-z

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "'Why should I trust you?' Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Acad. Med.* doi: 10.1145/2939672.2939778

Santos, M. R., Guedes, A., and Sanchez-Gendriz, I. (2024). SHapley additive exPlanations (SHAP) for efficient feature selection in rolling bearing fault diagnosis. *Learn. Knowl. Extr.* 6, 316–341. doi: 10.3390/make6010016

Shapley, L. (1997). "A value for n-person games," in *Contributions to the Theory of Games, II*. Ed. H. W. Kuhn (Princeton: Princeton University Press), 69–79. doi: 10.1515/9781400829156-012

Sheng, M., He, Q., Yu, D., and Su, B. (2023). Age-groups classification of Irrawaddy dolphins based on dorsal fin geometric morphological features. *Ecol. Indic.* 154, 110506. doi: 10.1016/j.ecolind.2023.110506

Song, H., Ding, Z., Guo, C., Li, Z., and Xia, H. (2008). *Research on combination kernel function of support vector machine International Conference on Computer Science and Software Engineering* (Wuhan, China: IEEE), 838–841. doi: 10.1109/CSSE.2008.1231

Stefanov, I., Baeten, V., Abbas, O., Colman, E., Vlaeminck, B., De Baets, B., et al. (2010). Analysis of milk odd- and branched-chain fatty acids using fourier transform (FT)-Raman spectroscopy. *J. Agric. Food Chem.* 58, 10804–10811. doi: 10.1021/JF102037G

Štrumbelj, E., and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* 41, 647–665. doi: 10.1007/s10115-013-0679-x

Taqvi, S. A. A., Zabiri, H., Uddin, F., Naqvi, M., Tufa, L. D., Kazmi, M., et al. (2022). Simultaneous fault diagnosis based on multiple kernel support vector machine in nonlinear dynamic distillation column. *Energy Sci. Eng.* 10, 814–839. doi: 10.1002/ese3.1058

Thanapol, P., Lavangnananda, K., Bouvry, P., Pinel, F., and Lepŕevost, F. (2020). "Reducing overfitting and improving generalization in training convolutional neural network (CNN) under limited sample sizes in image recognition," in *th International Conference on Information Technology (InCIT)* (IEEE, New York), 300–305. doi: 10.1109/InCIT50588.2020.9310787

Tian, Z. (2020). Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM. *Eng. Appl. Artif. Intell.* 91, 103573. doi: 10.1016/j.engappai.2020.103573

Tian, W., Li, Y., Guzman, C., Ibba, M. I., Tilley, M., Wang, D., et al. (2023). Quantification of food bioactives by NIR spectroscopy: Current insights, long-lasting challenges, and future trends. *J. Food Compos. Anal.* 124, 105708. doi: 10.1016/j.jfca.2023.105708

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory* (USA: SpringerVerlag (New York). doi: 10.1007/978-1-4757-2440-0

Wang, Z., and Fang, B. (2019). RETRACTED ARTICLE: Application of combined kernel function artificial intelligence algorithm in mobile communication network security authentication mechanism. *J. Supercomput.* 75, 5946–5964. doi: 10.1007/s11227-019-02896-5

Wang, B., Jia, G., Zheng, W., Chen, Y., Liu, C., Ai, D., et al. (2022). Correlations between Amadori compound contents and quality of flue-cured tobacco leaves from different ecoregions. *Tob. Sci. Technol.* 55, 33–40. doi: 10.16135/j.issn1002-0861.2021.0609

Wang, D., Zhao, F., Wang, R., Guo, J., Zhang, C., Liu, H., et al. (2023). A Lightweight convolutional neural network for nicotine prediction in tobacco by near-infrared spectroscopy. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1138693

Wei, B., Wang, R., Hou, K., Wang, X., and Wu, W. (2018). Predicting the current and future cultivation regions of Carthamus tinctorius L. using MaxEnt model under climate change in China. *Glob. Ecol. Conserv.* 16, e00477. doi: 10.1016/j.gecco.2018.e00477

Xiao, H., Chen, Z., Yi, S., and Liu, J. (2023). Rapid detection of maize seed germination rate based on Gaussian process regression with selection kernel function. *Vib. Spectrosc.* 129, 103595. doi: 10.1016/j.vibspec.2023.103595

Xu, X., Gao, B., Li, Y., Liu, J., Han, L., and Liu, X. (2023). The effect of temperature on the identification of NIR animal fats and oils species and its mechanism. *Vib. Spectrosc.* 124, 103498. doi: 10.1016/j.vibspec.2023.103498

Xu, Q., Zhang, M., Gu, Z., and Pan, G. (2019). Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs. *Neurocomputing* 328, 69–74. doi: 10.1016/j.neucom.2018.03.080

Zhang, Y. N., and Teng, H. F. (2009). Detecting particle swarm optimization. *Concurrency. Computat. Pract. Exp.* 21, 449–473. doi: 10.1002/cpe.1347

Zhang, R., and Wang, W. (2011). Facilitating the applications of support vector machine by using a new kernel. *Expert Syst. Appl.* 38, 14225–14230. doi: 10.1016/j.eswa.2011.04.235

Zhang, K., Yan, F., and Liu, P. (2024). The application of hyperspectral imaging for wheat biotic and abiotic stress analysis: A review. *Comput. Electron. Agric.* 221, 109008. doi: 10.1016/j.compag.2024.109008

Zhu, B., Ye, S., Wang, P., Chevallier, J., and Wei, Y. M. (2022). Forecasting carbon price using a multi-objective least squares support vector machine with mixture kernels. *J. Forecasting.* 41, 100–117. doi: 10.1002/for.2784

Zushi, K., Yamamoto, M., Matsuura, M., Tsutsuki, K., Yonehana, A., Imamura, R., et al. (2025). Machine learning and multiple linear regression models can predict ascorbic acid and polyphenol contents, and antioxidant activity in strawberries. *J. Sci. Food Agric.* 105, 1159–1169. doi: 10.1002/jsfa.13906