Journal of Food Safety and Food Quality
Archiv für Lebensmittelhygiene

J. Food Saf. Food Qual. **2026**; 77(5): 51458
https://doi.org/10.31083/JFSFQ51458

Original Research

Comparison of Different Intelligent Spectral Interval Selection Methods for Near-Infrared Spectroscopy-Based Identification of Simulated Flour Adulteration With Multiple Illegal Additives

Siting Wu[1], Xinyi Dong[1], Chunqi Wang[2,*], Jinming Liu[1,*]

[1]College of Electrical and Information, Heilongjiang Bayi Agricultural University, 163319 Daqing, Heilongjiang, China
[2]College of Food Science, Heilongjiang Bayi Agricultural University, 163319 Daqing, Heilongjiang, China
*Correspondence: wangchunqi1996@foxmail.com (Chunqi Wang); jinmingliu2008@126.com (Jinming Liu)
Academic Editor: Corinna Kehrenberg
Submitted: 3 March 2026 Revised: 30 June 2026 Accepted: 6 July 2026 Published: 9 September 2026

Abstract

Background: Flour is a globally consumed staple food, making its quality and safety crucial for public health. However, the limitations of conventional detection methods and the poor suitability of existing spectral interval selection techniques have constrained the accurate classification of flour adulteration. To address these challenges, this study presents a near-infrared spectroscopy-based approach combined with intelligent spectral interval selection for identifying mixed flour adulteration. **Methods:** Three hybrid intelligent optimization algorithms were developed: genetic simulated annealing (GSA), genetic particle swarm optimization (GPSO), and simulated annealing particle swarm optimization (SAPSO). These algorithms were integrated with interval partial least squares (iPLS) to develop GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS feature interval selection methods. Highly correlated spectral intervals were selected using the cross-validation classification accuracy of partial least squares discriminant analysis (PLSDA) as the objective function. PLSDA and support vector machine (SVM) classification models were constructed using the selected intervals. **Results:** All three methods effectively reduced spectral dimensionality and enhanced model performance. Among them, the GSA-iPLS-based PLSDA model demonstrated the best linear classification performance on the test set, achieving an accuracy and F1 score of 94.61% on the test set and an accuracy of 92.26% with an F1 score of 92.84% on the independent validation set. In contrast, the GPSO-iPLS-based SVM model exhibited the highest nonlinear classification performance, with an accuracy and F1 score of 100% on both the test and independent validation sets, respectively. **Conclusions:** The proposed hybrid intelligent spectral interval selection methods significantly enhance the predictive performance and robustness of flour adulteration detection models. These findings provide a valuable technical framework and theoretical basis for assessing flour quality and safety testing.

Keywords: near-infrared spectroscopy; flour adulteration; chemometrics; spectral analysis; discriminant analysis; support vector machine

1. Introduction

As the primary processed product of wheat, flour is an indispensable global food ingredient rich in carbohydrates, proteins, vitamins, and folic acid. Flour is widely used for producing steamed buns, noodles, dumplings, and pastries [1]. To pursue excessive profits, some manufacturers illegally add unapproved substances such as benzoyl peroxide, talc powder, and azodicarbonamide (ADA) into flour to improve its appearance and processability [2]. However, these additives may pose potential hazards to human health [3]. Relevant regulations specify the types and allowable levels of approved flour additives. However, conventional detection technologies lack sufficient sensitivity for illegal additives and may not distinguish hidden risks associated with the combined use of multiple additives. Therefore, developing accurate detection technologies for qualitative and quantitative analysis of flour adulterants is crucial for ensuring food safety [4].

Near-infrared spectroscopy (NIRS) is an analytical technique that utilizes overtone and combination-band absorption of hydrogen-containing functional groups (–CH, –NH, –OH) for qualitative and quantitative analysis [5,6]. Owing to its rapid, non-destructive, environmentally friendly, and easy-to-use characteristics, NIRS is widely used in agriculture, food science, medicine, and chemical engineering. Therefore, NIRS has emerged as a key tool for modern food analysis, particularly for wheat flour quality assessment [7,8]. Previous studies on flour adulteration detection have mainly focused on NIRS-based methods for identifying individual illegal additives and developing quantitative models for simple binary/ternary adulterated mixtures [9,10,11,12]. However, the rapid identification of adulterants and their types in complex multi-component mixtures remains a significant challenge. Current methods are often inadequate for this task.

The high dimensionality of NIRS data leads to the "curse of dimensionality", characterized by excessive variable dimensionality, information redundancy, and severe collinearity. These issue significantly impairs the prediction accuracy and robustness of chemometric models,

which limits the practical application of NIRS [13]. To address these challenges, feature selection techniques are used to identify characteristic wavelengths or spectral intervals associated with target components. These methods can be classified into individual wavelength variable selection and spectral interval selection [14,15]. Intelligent optimization algorithms have been widely applied for spectral interval selection owing to their strong global search capabilities [16]. However, these algorithms typically select optimal spectral regions based on regression analysis performance indicators, such as root mean square error of cross-validation (RMSECV), which may be suboptimal for classification tasks.

Although intelligent optimization algorithms have powerful search capabilities, each has inherent limitations. For example, the genetic algorithm (GA) converges slowly, the simulated annealing algorithm (SA) is inefficient, and the particle swarm optimization algorithm (PSO) is prone to premature convergence [17,18]. The high dimensionality, strong correlation, and limited sample size of spectral data hinder a single algorithm from achieving both effective global search and accurate convergence. The direct application of these algorithms to whole wavelength selection often results in a divergent solution space due to overly long coding lengths [19]. Therefore, pairwise hybrid optimization strategies are needed to combine the complementary strengths of different algorithms for classification-oriented spectral interval selection. However, the application of these strategies remains underexplored.

From the analysis above, this study aims to identify mixed adulterants in wheat flour using NIRS together with intelligent spectral interval selection. Specifically, the main research objectives are as follows:

(1) Three hybrid intelligent optimization algorithms are developed through the pairwise combination of GA, SA, and PSO to overcome the limitations of conventional single intelligent optimization methods.

(2) These three hybrid algorithms are integrated with interval partial least squares (iPLS) for feature spectral interval selection. This integration enables the acquisition of highly correlated spectral variables for constructing high-accuracy classification models.

(3) A NIRS-based discrimination model is developed for the identification of multiple illegal additives in mixed adulterated wheat flour, which provides reliable technical support for random inspections by quality supervision departments.

2. Materials and Methods

2.1 Preparation of Experimental Samples

Experimental samples were prepared through the blending of three illegal adulterants (ADA, talc powder, and gypsum powder) into unadulterated wheat flour at various proportions. The only additive-free wheat flour was obtained from Nanlai Jiudui Flour Processing Factory, located in Yangcao Town, Anda City, Suihua Prefecture, Heilongjiang Province, China. All three illegal adulterants were purchased from a certified chemical reagent supplier and verified for consistency in chemical composition and particle size. The ADA adulteration level ranged from 0 to 0.06% (w/w). To accurately prepare samples with lower ADA concentrations, a stock flour sample containing 0.1% (w/w) ADA was first prepared and then diluted to the desired concentration. The adulteration levels of talc powder and gypsum powder were each controlled within the range of 0 to 20% (w/w). All adulterated samples were mixed for 3 min using a SCI-VS adjustable vortex mixer (SCILOGEX, Rocky Hill, CT, USA) to ensure uniform dispersion of the adulterants throughout the wheat flour matrix. A total of 840 adulterated wheat flour samples were prepared (**Supplementary Materials**). These samples included 120 samples for each single-adulterant group, 120 samples for each binary adulterant combination, and 120 samples for the ternary mixture containing all three adulterants.

2.2 Spectral Data Acquisition

The NIRS data of the mixed-adulterated wheat flour samples were collected using a TANGO Fourier-transform near-infrared spectrometer (Bruker, Karlsruhe, BW, Germany). Before data acquisition, the instrument was warmed up for 30 min and operated in rotating diffuse reflection mode using an integrated IN311/C sample cup. Spectral acquisition and analysis were performed using OPUS 7.5 software (Bruker, Karlsruhe, BW, Germany). During measurement, the sample height was maintained at least one-third of the cup depth. The IN311/C lid was tightly secured to prevent near-infrared light from penetrating through the sample. Spectral data were recorded over the wavenumber range of 11,542–3946 cm^{-1} with a sampling interval of 8 cm^{-1}. A background spectrum was collected every hour, and each sample was scanned 32 times to obtain a single spectrum. Each sample was independently loaded and measured thrice. The average of the three spectra was used as the raw spectral data for subsequent analysis. After each measurement, the sample cup was thoroughly cleaned to remove residual material and prevent cross-contamination.

2.3 Spectral Data Preprocessing

Spectral data are often affected by instrument noise, background signals, and external environmental factors. Consequently, the direct use of raw spectra for modeling can reduce model accuracy and stability. Therefore, raw spectral preprocessing is crucial before classification modeling to remove noise, correct baseline drift, and eliminate interference from irrelevant variables [20]. In this study, the raw spectra of adulterated wheat flour samples were preprocessed using Savitzky-Golay (SG) smoothing, multiplicative scatter correction (MSC), standard normal variate transformation (SNV), and their combinations. The preprocessing aimed to improve the purity of spectral signals and

their signal-to-noise ratio, thereby providing a reliable data foundation for subsequent feature extraction and model development. Specifically, SG smoothing efficiently eliminates noise interference while maintaining the profiles of signal characteristics. MSC corrects scattering effects that arise from differences in the physical properties of samples, thereby eliminating the interference of background signals on spectral data. SNV can successfully reduce spectral variations among samples and standardize signal intensities, further enhancing the consistency of spectral data and the robustness of the models.

2.4 Sample Set Partitioning

Sample set partitioning is an important step in spectral model development. An appropriate partitioning strategy can effectively improve model generalization and reduce the risk of overfitting or underfitting. This strategy also ensures training samples and test samples are representative and balanced, thereby improving the model's predictive performance [21]. In this study, 20% of the original data was randomly selected as an independent validation set to evaluate the generalization ability of the trained models objectively. The independent validation set was also used to assess model robustness and reduce the risk of overfitting. The remaining 80% of the samples were partitioned into training and test sets using stratified random sampling at a 3:1 ratio [22]. This approach preserved the class distribution across the subsets, thereby preventing the overrepresentation or underrepresentation of specific categories and improving the classification performance and stability of the resulting models.

2.5 Intelligent Spectral Interval Selection Methods

2.5.1 Construction of Hybrid Intelligent Optimization Algorithms

2.5.1.1 Genetic Simulated Annealing Algorithm. The genetic simulated annealing algorithm (GSA) combines the strong search capability of GA with the global optimization advantage of SA. The algorithm integrates a temperature parameter into the fitness functions and incorporates a Metropolis selection-replication strategy. These mechanisms maintain population diversity during the early high-temperature stage, and enhance the fitness of superior chromosomes during the later low-temperature stage. Consequently, GSA effectively alleviates the inherent limitations of GA [23]. Particularly, GSA reduces premature convergence and improves search efficiency during the late stages of evolution. In this study, the GSA was applied to feature spectral wavelength selection using binary encoding. The code length was set equal to the number of spectral variables or intervals to be optimized. To meet the requirements of classification tasks, the accuracy of cross-validation (AccuracyCV) was adopted as the objective function. The fitness function was designed by combining temperature parameters, as defined by the following equation:

$$fit(x) = \exp\left(\frac{f(x) - f_{max}}{t_n}\right) \tag{1}$$

In the equation, $f(x)$ represents the objective function value of the current solution, f_{max} denotes the optimal objective function value in the current population, and t_n indicates the current temperature. The initial temperature is calculated as $t_0 = T(f_o^{max} - f_o^{min})$, and the cooling schedule follows $t_{n+1} = \alpha t_n$, $0 < \alpha < 1$. Here, T is a positive integer, while f_o^{max} and f_o^{min} represent the maximum and minimum objective function values of the initial population, respectively.

GSA enhances its optimization performance by incorporating the Metropolis selection-replication mechanism of SA into the GA framework. GA framework includes roulette-wheel selection with elitist preservation, discrete recombination crossover, and discrete bit mutation. All operations are performed based on fitness function values. The Metropolis selection-replication process is described as follows: For a given solution X_i, a neighborhood solution $X_i^{'}$ is generated, where $f(X_i)$ and $f(X_i^{'})$ denote the objective function values of X_i and $X_i^{'}$, respectively. Δf is defined as $\Delta f = f(X_i) - f(X_i^{'})$. For maximization problems, if $\Delta f < 0$, the neighborhood solution $X^{'}$ is accepted into the next-generation population. If $\Delta f \geq 0$, a random number $r \in [0, 1]$ is generated. If $r < \exp(-\Delta f / t_n)$, the neighborhood solution $X_i^{'}$ is still accepted. Otherwise, the neighborhood solution X_i is retained into the next-generation population.

GSA consists of two key components: the fitness function calculation method and the Metropolis selection-replication strategy. These components enable the algorithm to accept suboptimal solutions with a higher probability at high temperature, thereby maintaining population diversity and mitigating premature convergence. As the temperature decreases, superior chromosomes are more likely to be retained and propagated to the next generation, thereby accelerating algorithm convergence. Fig. 1 shows the execution process of the GSA.

2.5.1.2 Genetic Particle Swarm Optimization Algorithm. To solve discrete optimization problems such as wavelength selection, the standard PSO must use binary discretization. In this study, a binary PSO (BPSO) algorithm was developed in which particle positions were encoded as 0 or 1. BPSO retains the velocity update rule of the standard PSO. Instead, BPSO uses a sigmoid function to map particle velocities to binary probabilities, enabling binary position updates. The corresponding equation is presented below:

$$X_i^n = \begin{cases} 1 & r < \frac{1}{1+\exp(-v_i^n)} \\ 0 & r \geq \frac{1}{1+\exp(-v_i^n)} \end{cases} \tag{2}$$

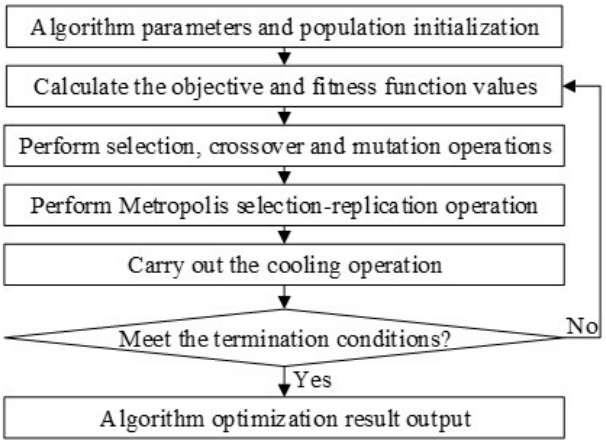

Fig. 1. Flowchart of genetic simulated annealing algorithm.

In the equation, v_i^n and X_i^n denote the velocity and position of the *n*th dimension of particle i, respectively, and r is a random number satisfying $0 \leq r \leq 1$.

The genetic particle swarm optimization algorithm (GPSO) incorporates two strategies (discrete recombination crossover and discrete bit mutation) into the evolutionary process of BPSO. These strategies enhance population diversity, improve the balance between exploration and exploitation, and reduce the risk of stagnation and premature convergence [24]. Consequently, the global convergence and robustness of the algorithm are improved. GPSO uses the same encoding scheme and objective function as GSA. The evolutionary process of GPSO consists of five main steps: particle population initialization, individual and global best solution updates, particle velocity updates, discrete recombination crossover and discrete mutation evolutions, and particle position updates (Fig. 2).

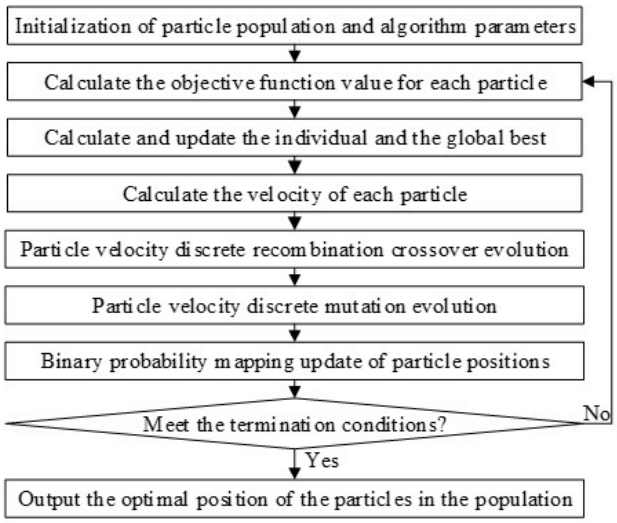

Fig. 2. Flowchart of genetic particle swarm optimization algorithm.

2.5.1.3 Simulated Annealing Particle Swarm Optimization Algorithm. The simulated annealing particle swarm optimization algorithm (SAPSO) incorporates the Metropolis selection-replication evolutionary strategy of SA into the evolutionary process of BPSO. For each particle updated by BPSO, multiple bits are randomly selected for binary bit mutation to generate perturbed solutions. These perturbed solutions are then evaluated using the Metropolis criterion. This strategy effectively overcomes the inherent limitations of both SA and PSO, providing rapid convergence and enhanced global optimization capabilities [25]. SAPSO adopts the same encoding scheme, objective function, initial temperature determination, and cooling schedule as GSA. The evolutionary process of SAPSO consists of five components: the individual best solution updates, the global best solution updates, particle velocity updates, particle position updates, and Metropolis selection-replication (Fig. 3).

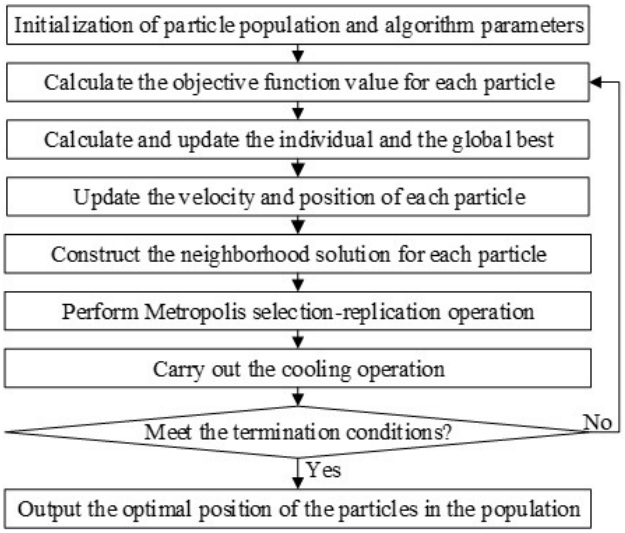

Fig. 3. Flowchart of simulated annealing particle swarm optimization algorithm.

2.5.2 Intelligent Feature Spectral Interval Selection Algorithms

Three intelligent feature spectral interval selection algorithms were developed, including genetic simulated annealing interval partial least squares (GSA-iPLS), genetic particle swarm optimization interval partial least squares (GPSO-iPLS), and simulated annealing particle swarm optimization interval partial least squares (SAPSO-iPLS). Each of them combines the iPLS concept with the strong search capabilities of different intelligent algorithms. First, the spectral data are divided into *n* equal-width intervals. GSA, PSO, and SAPSO are then used to select effective spectral regions for modeling, thereby reducing data dimensionality and improving model performance. All three al-

gorithms use binary encoding. During population initialization, the number of intervals serves as the code length. In this encoding scheme, "1" indicates that the spectral variables included in the corresponding interval are selected for modeling computation. However, "0" indicates that the corresponding interval is excluded from the analysis. For example, if a solution encoding with 12 intervals is encoded as the binary sequence "010000100001", the 2nd, 7th, and 12th intervals (from left to right) are selected for analysis. Consequently, all spectral data corresponding to the spectral variables within these selected intervals are included in the modeling computation. After population initialization and parameter setting (including crossover and mutation probabilities, inertia weight, individual and global learning factors, initial temperature, and cooling schedule), the objective function value of each solution is calculated. The corresponding fitness function value for GSA is calculated. Based on these objective or fitness function values, multiple rounds of each algorithm- specific evolutionary operation are sequentially executed. These operations include roulette-wheel selection with elitist preservation strategy, discrete recombination crossover, discrete mutation, particle velocity and position updates, individual and global optimal solution updates, and Metropolis selection-replication. Through iterative optimization, informative feature spectral regions are selected for model development. Model performance under different interval partition schemes is evaluated to determine the optimal number of spectral intervals and the corresponding spectral feature regions.

2.6 Classification Modeling Methods

Partial least squares discriminant analysis (PLSDA) is a classification method that integrates partial least squares (PLS) with linear discriminant analysis (LDA). PLSDA first applies PLS to extract latent variables that capture the relationship between input features and class labels. LDA is then performed on these latent variables in a low-dimensional space to construct the classification model [26]. PLS establishes linear relationships between the feature space and the categorical space to extract latent variables that characterize class differences. These latent variables are linear combinations of the original features, thereby reducing dimensionality and removing redundant information. LDA is then applied to discriminant analysis on these latent variables to achieve effective separation of samples from different classes. PLSDA combines the dimensionality reduction capability of PLS with the classification and discriminant power of LDA to extract discriminative features from high-dimensional spectral data and maintain strong generalization performance.

Support vector machine (SVM) is a supervised learning approach founded on statistical learning theory, which is widely utilized for both classification and regression tasks. SVM defines an optimal separating hyperplane in the feature space that maximizes the margin between samples and distinct classes, enabling strong generalization performance [27]. Although originally developed for binary classification, SVM can be effectively extended to multiclass classification tasks. SVM utilizes kernel functions to map nonlinearly separable data from a low-dimensional space into a higher-dimensional feature space. In this transformed space, the data become linearly separable. Consequently, SVM can effectively handle complex nonlinear classification tasks and achieves excellent performance on small- to medium-sized, high-dimensional datasets. However, SVM has a high computational cost and is sensitive to parameter selection, making it unsuitable for extremely large datasets.

2.7 Model Evaluation Metrics

Before constructing the classification model, the following three tasks need to be completed: spectral preprocessing, sample set partitioning, and feature spectral interval selection. After these steps, the selected feature wavelengths were used to construct two models: PLSDA and SVM. Model performance was evaluated using four metrics: accuracy, precision, recall, and F1 score [28]. Among these metrics, accuracy is the most widely used indicator of classification performance, defined as the proportion of correctly classified samples among all samples. Precision measures the accuracy of positive predictions and is defined as the proportion of correctly predicted positive samples among all predicted positive samples. Recall is the proportion of actual positive samples correctly identified by the model and reflects its ability to detect positive samples. The F1 score is the harmonic mean of precision and recall and is particularly suitable for evaluating models on imbalanced datasets. The calculation formulas for the above evaluation metrics are presented as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive and false negative samples, respectively.

3. Results

3.1 Spectral Data Analysis and Processing

Spectral data are influenced by factors such as ambient temperature, sample particle size, and instrument performance, which may introduce high-frequency noise, baseline drift, and spectral scattering. To minimize these effects, the collected spectral data were preprocessed using SG smoothing, MSC, SNV, and their combinations. The effectiveness of different preprocessing methods was evaluated based on the AccuracyCV of a 10-fold cross-validation PLSDA model. The preprocessing method with the highest AccuracyCV was selected for subsequent classification modeling (Table 1).

Table 1. Comparison results of preprocessing methods.

Preprocessing methods	AccuracyCV (%)
Null	79.19
SG	79.29
MSC	85.24
SNV	85.83
SG+MSC	83.57
SG+SNV	84.05
MSC+SNV	85.36

AccuracyCV, accuracy of cross-validation; SG, Savitzky-Golay; MSC, multiplicative scatter correction; SNV, standard normal variate transformation.

Table 1 shows that all preprocessing methods improved the classification accuracy compared to the raw spectral data. Notably, MSC and SNV achieved higher accuracies than the other preprocessing methods. Regarding the AccuracyCV values, SNV was ultimately selected as the optimal preprocessing method for identifying mixed adulterated flour samples. The raw and preprocessed spectra are presented in Fig. 4. A comparison of the spectra before and after preprocessing revealed that SNV effectively corrected baseline drift and spectral scattering, thereby enhancing spectral resolution and increasing signal-to-noise ratio [21]. Particularly, the absorption peak at 7000 cm^{-1} (Dashed ellipse marked area in Fig. 4), attributed to ADA adulteration, became more pronounced after SNV preprocessing, providing valuable spectral data for constructing a high-accuracy classification model. This absorption peak corresponds to the first overtone of the N–H stretch vibration of the formamide group ($–CONH_2$) in ADA.

3.2 Feature Spectral Interval Selection

The 840 adulterated flour samples were partitioned into a training set (504 samples), a testing set (168 samples), and an independent validation set (168 samples) using a combination of random and stratified sampling. The optimal spectral feature intervals were identified from the training set using GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS. The number of spectral interval partitions significantly influences the performance of intelligent spectral interval selection methods. Therefore, a two-step optimization strategy was adopted to determine the optimal number of partitions. First, a coarse search was performed with a step size of 10 to identify the preliminary optimal partition number of intervals. Subsequently, a fine search was conducted with a step size of 1 around the preliminary value to determine the final optimal number of partitions.

Taking GSA-iPLS as an example, a coarse search of the optimal partition number of spectral intervals was first conducted with the partition number from 10 to 100 with a step size of 10. When the number of intervals was set to 40, the PLSDA model achieved the highest AccuracyCV value of 92.29%. Therefore, the optimal number of intervals was likely between 31 and 49. A subsequent fine search with a step size of 1 was performed within this range. Ultimately, at the partition number of 38, the PLSDA model achieved the highest AccuracyCV (93.29%) and selected 23 feature intervals containing 1119 spectral variables (Fig. 5A). Similarly, both the preliminary and optimal partition numbers of intervals for GPSO-iPLS were 60, yielding 32 feature intervals containing 986 spectral variables (Fig. 5B). For SAPSO-iPLS, the preliminary partition number of intervals was 50, while the optimal partition number was 57. Under these conditions, 17 feature intervals containing 780 spectral variables were selected (Fig. 5C).

Fig. 5D shows the distribution of optimal feature spectral intervals selected via GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS. The three algorithms exhibited a high degree of consistency in the selected spectral regions. Specifically, 492 spectral variables were common to all three methods, accounting for 43.97%, 49.90%, and 63.08% of the spectral variables selected by GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS, respectively. These results indicate the robust search capabilities of the GSA, GPSO, and SAPSO algorithms. Under the guidance of AccuracyCV, they can effectively realize the selection of feature spectral regions for high-correlation modeling by combining with the iPLS strategy.

3.3 Construction of PLSDA Classification Models

To evaluate the effectiveness of GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS for linear classification modeling, PLSDA models were performed using the highly correlated spectral variables selected by each method. The modeling performances were further compared with that of the full-spectrum and the spectral intervals selected by backward iPLS (BiPLS). Model performance was assessed using accuracy, precision, recall, and F1 score. As shown in Table 2, the classification models built on the spectral regions selected by GSA-iPLS, GPSO-iPLS, SAPSO-iPLS and BiPLS exhibited improved performance compared to the model based on the full-spectrum. Among these, the GSA-iPLS-selected PLSDA model achieved the

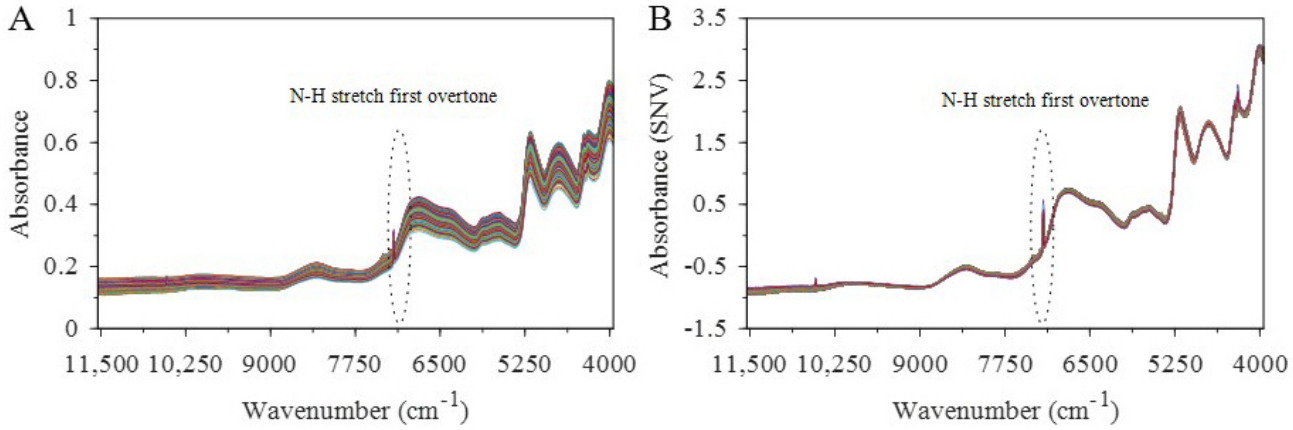

Fig. 4. Spectral data before and after preprocessing. (A) Raw spectral data. (B) Preprocessed spectral data.

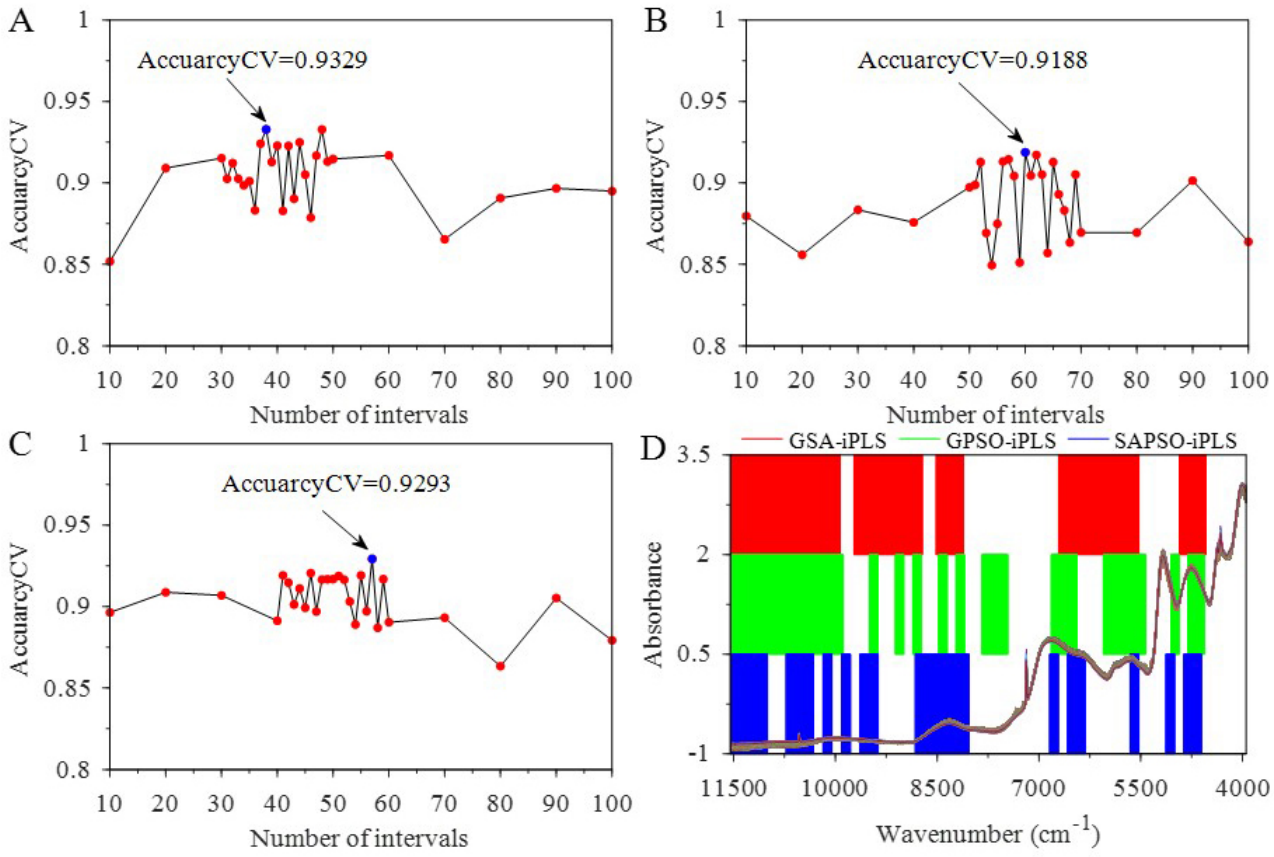

Fig. 5. Results of intelligent feature spectral interval selection. (A) Genetic simulated annealing interval partial least squares (GSA-iPLS). (B) Genetic particle swarm optimization interval partial least squares (GPSO-iPLS). (C) Simulated annealing particle swarm optimization interval partial least squares (SAPSO-iPLS). (D) Spectral intervals selected by three intelligent algorithms.

highest classification performance, with a test set accuracy of 94.61%, precision of 94.67%, recall of 94.55%, and an F1 score of 94.61%. Compared with the full-spectrum model, the GSA-iPLS-based PLSDA model improved accuracy, precision, recall, and F1 score by 9.58%, 9.28%, 9.53%, and 9.41%, respectively. For the independent validation set, the GSA-iPLS-based PLSDA model yielded an accuracy of 92.26%, precision of 92.89%, recall of 92.79%, and an F1 score of 92.84%, representing substantial improvements over the full-spectrum model. Overall, the selection of highly informative spectral variables effectively reduced data dimensionality and enhanced classification performance, enabling the reliable identification of illegal adulterant mixtures in flour.

Table 2. Performance metrics of PLSDA models constructed with different spectral interval selection methods.

Samples	Indicators	Full	BiPLS	GSA-iPLS	GPSO-iPLS	SAPSO-iPLS
	Wavelength	1845	1259	1119	986	780
Training set	Accuracy (%)	88.12	92.46	95.64	95.65	95.85
	Precision (%)	88.45	92.74	95.73	95.65	95.94
	Recall (%)	88.08	92.51	95.57	95.56	95.85
	F1 score (%)	88.27	92.62	95.65	95.60	95.90
Test set	Accuracy (%)	85.03	90.48	94.61	92.17	92.77
	Precision (%)	85.39	90.51	94.67	92.33	92.95
	Recall (%)	85.02	90.38	94.55	92.20	92.77
	F1 score (%)	85.20	90.45	94.61	92.27	92.86
Validation set	Accuracy (%)	86.31	92.86	92.26	92.26	88.10
	Precision (%)	86.76	93.15	92.89	93.19	89.43
	Recall (%)	86.64	92.79	92.79	92.89	88.56
	F1 score (%)	86.70	92.97	92.84	93.04	88.99

PLSDA, partial least squares discriminant analysis; BiPLS, backward iPLS.

3.4 Construction of SVM Classification Models

To evaluate the applicability of GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS to nonlinear modeling, SVM classification models were constructed using the spectral variables selected by each method. The performance of these models was compared with that of full-spectrum and BiPLS-based SVM models. In this study, the radial basis function was adopted as the kernel function, and grid search was used to optimize the kernel parameter and penalty factor. As shown in Table 3, the SVM models based on GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS exhibited excellent classification performance and outperformed the full-spectrum and BiPLS-based model. Among them, the GPSO-iPLS-based SVM model achieved the highest classification performance, with 100% accuracy, precision, recall, and F1 score on the training, test, and independent validation sets. These results demonstrate that all three feature selection algorithms are equally applicable to nonlinear SVM modeling. Therefore, integrating GSA-iPLS, GPSO-iPLS, or SAPSO-iPLS with SVM enables the accurate identification of mixed illegal adulterants in flour.

4. Discussion

4.1 Discussion of Results

A comprehensive analysis of the full-spectrum PLSDA and SVM classification models revealed that the raw spectral data comprised 1845 spectral variables. Most of these variables were redundant and non-discriminative. This high dimensionality increases model complexity, promotes overfitting, and reduces the generalization capability of classification models [29]. Without feature selection, the full-spectrum PLSDA model achieved an 88.12% accuracy and an F1 score of 88.27% on the training set. The corresponding values on the test set were 85.03% and 85.20%. Although the full-spectrum PLSDA model demonstrated acceptable classification performance and met the basic identification requirements, its relatively low identification accuracy and limited stability hinder its practical applicability. It indicates the key role of effective spectral feature selection in improving model efficiency and predictive performance. In contrast, the full-spectrum SVM model exhibited superior performance. It achieved 100% accuracy and a 100% F1 score on the training set, and 98.21% accuracy with an F1 score of 98.30% on the test set. The full-spectrum SVM model significantly outperformed the PLSDA model, reflecting the strong generalization capability of nonlinear SVM algorithms. However, for high-dimensional spectral data, SVM models incur high computational costs and increased training complexity. These limitations highlight the need for effective spectral feature selection to improve modeling efficiency and maintain high classification performance [30].

We compared the classification performance of different PLSDA models constructed using full-spectrum and different feature spectral interval selection algorithms. The results revealed that the feature-selected models consistently outperformed the full-spectrum model. Specifically, the GSA-iPLS-based PLSDA model achieved an accuracy of 94.61% on the test and 92.26% on the independent validation set. Its corresponding F1 scores were 94.61% and 92.84%, respectively. This level of performance is sufficient to meet the requirements for rapid identification of flour adulterated with multiple illegal additives [31]. Compared with the GPSO-iPLS-based model, the GSA-iPLS-based model exhibited slightly lower accuracy and precision on the training and independent validation sets. However, the GSA-iPLS-based model achieved superior overall classification performance on the test set. Similarly, only minor performance differences were observed between GSA-iPLS and SAPSO-iPLS on the training set. The GSA-iPLS-based model exhibited slightly lower training per-

Table 3. Performance metrics of SVM models constructed with different spectral interval selection methods.

Samples	Indicators	Full	BiPLS	GSA-iPLS	GPSO-iPLS	SAPSO-iPLS
	Wavelength	1845	1259	1119	986	780
Training set	Accuracy (%)	100.00	100.00	100.00	100.00	100.00
	Precision (%)	100.00	100.00	100.00	100.00	100.00
	Recall (%)	100.00	100.00	100.00	100.00	100.00
	F1 score (%)	100.00	100.00	100.00	100.00	100.00
Test set	Accuracy (%)	98.21	98.81	100.00	100.00	99.40
	Precision (%)	98.38	98.98	100.00	100.00	99.43
	Recall (%)	98.21	98.86	100.00	100.00	99.41
	F1 score (%)	98.30	98.92	100.00	100.00	99.42
Validation set	Accuracy (%)	98.81	99.40	99.41	100.00	100.00
	Precision (%)	98.49	99.21	99.57	100.00	100.00
	Recall (%)	98.85	99.57	99.38	100.00	100.00
	F1 score (%)	98.67	99.39	99.47	100.00	100.00

SVM, support vector machine.

formance than the SAPSO-iPLS-based model. Nevertheless, GSA-iPLS significantly outperformed SAPSO-iPLS on both the test and independent validation sets. This difference may be partly attributed to overfitting in the SAPSO-iPLS-based PLSDA model. Compared with BiPLS, GSA-iPLS, GPSO-iPLS, and SAPSO-iPLS achieved superior overall classification performance. Notably, GSA-iPLS achieved the highest overall performance among PLSDA models. The advantage is likely attributed to the strong spectral interval search capability of GSA-iPLS. Consequently, GSA-iPLS effectively identifies highly discriminative spectral bands from high-dimensional spectral data, thereby improving both classification accuracy and model generalization [32].

We analyzed the performance of three SVM nonlinear classification models constructed using intelligent feature spectral interval selection methods. All three models achieved 100% accuracy and a 100% F1 score on the training set. This reflects the strong learning capability of SVM and its superior performance compared with the corresponding PLSDA models. The GSA-iPLS-based SVM model achieved an accuracy of 100% and an F1 score of 100% on test sets, and an accuracy of 99.41% and an F1 score of 99.47% on the independent validation set. This result indicates that the selected variables exhibited excellent discriminative power, good adaptability, and generalization capability for nonlinear classification. Similarly, the SAPSO-iPLS-based SVM model exhibited excellent classification performance, achieving accuracies of 99.40% on the test set and 100% on the independent validation set. Its corresponding F1 scores were 99.42% and 100%, respectively. These results further confirm the effectiveness of the hybrid intelligent optimization algorithm. Notably, the GPSO-iPLS-based SVM model achieved the highest overall classification performance. This model attained 100% accuracy and an F1 score of 100% on the training, test, and independent validation sets. Overall, the SVM models constructed using the three intelligent spectral interval selection algorithms achieved performance metrics exceeding 99.40% and consistently outperformed the BiPLS model. This result indicates that SVM models built on optimally selected spectral variables exhibit excellent learning capability and stability, thereby effectively reducing the influence of redundant information in full-spectrum data [33].

Compared with the spectral interval selection methods based on a single intelligent optimization algorithm [34], the three intelligent spectral interval selection algorithms proposed in this study fully exploit the advantages of pairwise combinatorial optimization designs in terms of global convergence and search efficiency, thereby achieving superior performance in selecting optimal characteristic wavelengths for modeling. In contrast to our previous work that applied the GSA combined with iPLS for quantitative regression modeling [35], AccuracyCV was adopted as the objective function of the hybrid intelligent optimization algorithm instead of RMSECV, which makes the algorithms more suitable for solving spectral classification modeling problems. Compared with detection studies of adulteration using multiple types of flour matrices [36], the use of a single flour matrix in this study does indeed limit the model's generalization ability with respect to differences in flour origin and variety. However, this does not undermine the core value of this research in discriminating complex adulteration patterns. The three proposed intelligent spectral interval selection algorithms effectively address the urgent requirement for selecting critical characteristic wavelengths during the construction of classification discriminant models under unary, binary, and ternary adulteration scenarios, where the interference from different adulterants is complex. This study is positioned as a methodological feasibility validation in complex multi-component adulteration scenarios, and future work will introduce multi-origin and

multi-variety flour matrices to evaluate the out-of-domain generalization ability of the model.

4.2 Limitations and Future Perspectives

This study integrated intelligent spectral interval selection with NIRS to identify flour adulterated with mixtures containing three illegal additives. Under the experimental conditions, the developed PLSDA and SVM classification models achieved satisfactory classification performance. However, the study has several limitations resulting from its experimental design and scope. Further research should focus on three key areas: sample representativeness, classification modeling methodologies, and quantitative regression modeling. First, only a single flour variety was used as the base sample for the three-adulterant mixture experiments. Although the use of a single flour type minimized confounding factors and confirmed the feasibility of the proposed NIR-based detection method, the composition of commercial flour products significantly varies. Moreover, real-world flour adulteration is not limited to the direct addition of illegal additives but also includes fraudulent practices such as misrepresenting inferior products as higher-quality products [37]. Second, the performance of intelligent spectral interval selection methods was evaluated using only two conventional modeling approaches (PLSDA and SVM), without investigating deep learning techniques. Additionally, this study mainly focused on the qualitative identification of mixed illegal additives in flour and did not address the quantitative determination of adulterant concentrations. Therefore, future research should focus on constructing a comprehensive flour spectral database containing multiple flour varieties and adulteration scenarios. Furthermore, integrating intelligent spectral interval selection with deep learning models should be investigated to enable simultaneous qualitative and quantitative analysis of flour adulteration. These efforts will facilitate the development of more intelligent, practical, and reliable rapid detection methods for ensuring flour quality and safety.

5. Conclusions

In this study, three hybrid intelligent optimization algorithms (GSA, GPSO, and SAPSO) were integrated with the iPLS strategy to achieve intelligent feature spectral interval selection for flour adulteration classification. This approach effectively addressed limitations of conventional algorithms, including premature convergence, low search efficiency, and poor global exploration. The PLSDA and SVM models based on the optimized spectral intervals consistently outperformed the full-spectrum models. Specifically, the GSA-iPLS-based PLSDA model achieved accuracies of 94.61% on the test set and 92.26% on the independent validation set. The GPSO-iPLS-based SVM model achieved 100% accuracy on both sets. These findings confirm the reliability and strong generalization capability of the proposed intelligent spectral interval selection methods for identifying mixed adulterants in flour. This study provides a valuable methodological framework for developing spectral detection technologies and supporting flour adulteration monitoring.

Availability of Data and Materials

The datasets used and analyzed during the current study are available from the corresponding authors upon reasonable request.

Author Contributions

CW and JL designed the research study. SW and XD performed the research and the data analysis. SW, XD, CW and JL wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research was funded by the National-Level College Student Innovation Training Program of Heilongjiang Province of China, grant number 202410223086.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used DeepSeek-V3.2 in order to check spelling and grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/JFSFQ51458.

References

[1] Kogniwali-Gredibert SB, Feudjio WM, Kongbonga GY, Pale WY, Assongo CK. Front-face fluorescence spectroscopy combined with chemometrics for the discrimination of wheat flour and cassava flour. Journal of Food Composition and Analysis. 2024; 127: 105962. https://doi.org/10.1016/j.jfca.2023.105962

[2] He HJ, Chen Y, Li GL, Wang YL, Ou XQ, Guo JL. Hyperspectral imaging combined with chemometrics for rapid detection of

talcum powder adulterated in wheat flour. Food Control. 2023; 144: 109378. https://doi.org/10.1016/j.foodcont.2022.109378
[3] Dong X, Dong Y, Liu J, Wang C, Bao C, Wang N, et al. Identification and quantitative detection of illegal additives in wheat flour based on near-infrared spectroscopy combined with chemometrics. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy. 2024; 323: 124938. https://doi.org/10.1016/j.saa.2024.124938
[4] Xu L, Chen Z, Deng J, Bai X, Gao Y, Jiang H, et al. Rapid quantitative and qualitative analysis of talcum powder in wheat flour by microwave detection combined with multivariate analysis. Microchemical Journal. 2024; 204: 111125. https://doi.org/10.1016/j.microc.2024.111125
[5] Sitorus A, Arianti ND. Artificial intelligence-assisted NIR spectroscopy for identification of halal mislabeling in commercial meatball products. Microchemical Journal. 2026; 221: 116816. https://doi.org/10.1016/j.microc.2026.116816
[6] Spinelli E, Casamassima R, Marini F. Non-destructive forensic identification and quantification of cocaine through sealed packaging using FT-NIR spectroscopy and hyperspectral imaging. Analytica Chimica Acta. 2026; 1414: 345679. https://doi.org/10.1016/j.aca.2026.345679
[7] Yang H, Xing J, Lei H, Meng M, Luo X, Sang S. Rapid quality evaluation of wheat flour containing moderate level of gluten using near-infrared spectroscopy and chemometrics. Journal of the Science of Food and Agriculture. 2025; 105: 4317–4325. https://doi.org/10.1002/jsfa.14221
[8] Wang Y, Zhang C, Li X, Xing L, Lv M, He H, et al. Machine-Learning-Algorithm-Assisted Portable Miniaturized NIR Spectrometer for Rapid Evaluation of Wheat Flour Processing Applicability. Foods (Basel, Switzerland). 2025; 14: 1799. https://doi.org/10.3390/foods14101799
[9] Duarte ESDA, de Almeida VE, da Costa GB, de Araújo MCU, Véras G, Diniz PHGD, et al. Feasibility study on quantification and authentication of the cassava starch content in wheat flour for bread-making using NIR spectroscopy and digital images. Food Chemistry. 2022; 368: 130843. https://doi.org/10.1016/j.foodchem.2021.130843
[10] Du CS, Sun LJ, Bai HY, Zhao ZD, Li XX, Gai ZD. Quantitative detection of talcum powder in wheat flour based on near-infrared spectroscopy and hybrid feature selection. Infrared Physics & Technology. 2022; 123: 104185. https://doi.org/10.1016/j.infrared.2022.104185
[11] Shi S, Feng J, Ma Y, Cao C, Li L, Jiang Y. Rapid determination of two illegal additives in wheat flour by near-infrared spectroscopy and different key wavelength selection algorithms. LWT. 2023; 189: 115437. https://doi.org/10.1016/j.lwt.2023.115437
[12] Dong X, Dong Y, Liu J, Wu S. Combined use of near infrared spectroscopy and chemometrics for the simultaneous detection of multiple illicit additions in wheat flour. Applied Food Research. 2025; 5: 101263. https://doi.org/10.1016/j.afres.2025.101263
[13] Shi S, Tian N, Liu Q, Yan R, Huang L, Jin Z, et al. Unveiling varietal specificity in non-destructive grape quality monitoring: Explainable AI and feature selection for sugar and organic acid prediction using NIR spectroscopy. Food Research International (Ottawa, Ont.). 2026; 235: 119154. https://doi.org/10.1016/j.foodres.2026.119154
[14] Zhu Y, Wang S, Lv J, Yin J, Kong D. Microplastic content prediction in agricultural soils using a mixed variable selection combined with Transformer-LSTM fusion architecture. Talanta. 2026; 306: 129756. https://doi.org/10.1016/j.talanta.2026.129756
[15] Wang Y, Wang S, Luo H. Development of online monitoring strategy for continuous coating process of transdermal patches based on near-infrared spectroscopy. International Journal of Pharmaceutics. 2026; 694: 126731. https://doi.org/10.1016/j.ijpharm.2026.126731
[16] Wang SY, Zhang XD, Mpango P, Sun H, Bian XH. Firefly Interval Selection Combined With Extreme Learning Machine for Spectral Quantification of Complex Samples. Journal of Chemometrics. 2024; 38: e3578. https://doi.org/10.1002/cem.3578
[17] Xie Y, Liu C, Li W, Du M, Feng K. Optimization of stamping process parameters based on an improved particle swarm optimization-genetic algorithm and sparse auto-encoder-back-propagation neural network model. Engineering Optimization. 2024; 56: 252–273. https://doi.org/10.1080/0305215x.2022.2152018
[18] Huang Y, Ge B, Hipel KW, Fang L, Zhao B, Yang K. Solving the inverse graph model for conflict resolution using a hybrid metaheuristic algorithm. European Journal of Operational Research. 2023; 305: 806–819. https://doi.org/10.1016/j.ejor.2022.06.052
[19] Lin Y, Cai H, Lin S, Ni H. Mid-Infrared-Spectroscopy-Based Method for Identifying Single and Multiple Vegetable Protein Adulterants in Whey Protein. Journal of Applied Spectroscopy. 2025; 91: 1378–1386. https://doi.org/10.1007/s10812-025-01863-8
[20] Yu M, Zeng J, Li Y, Fan W, Wang L, Wang H, et al. Near-infrared prediction of total phosphorus in leaves content in korla fragrant pear with growth period specificity via spectral modeling. Frontiers in Plant Science. 2025; 16: 1666460. https://doi.org/10.3389/fpls.2025.1666460
[21] Yang Z, Zeng F, Gu Y, Huang Y, Fang H, Hassan M, et al. Rapid fat quantification in *Torreya grandis* kernels using portable near-infrared spectroscopy: effects of shelling and kernel processing. Frontiers in Nutrition. 2026; 13: 1830282. https://doi.org/10.3389/fnut.2026.1830282
[22] Zhang GY, Chen KR, Zeng XL, Yin P, Qin MH, Deng SL, et al. Robust and interpretable grading of Baijiu base liquor via NIR spectroscopy: A Bagging-Boosting Hybrid Feature Fusion approach. Food Bioscience. 2026; 79: 108715. https://doi.org/10.1016/j.fbio.2026.108715
[23] Zhang Q, Song L, Zeng Y, Liu Y, Liu S, Wang N. Real-time power optimization strategy for fuel cell ships based on improved genetic simulated annealing algorithm. Electric Power Systems Research. 2025; 245: 111647. https://doi.org/10.1016/j.epsr.2025.111647
[24] Khodabakhshian R, Lavasani HS, Weller P. Optimization of FTIR-PLS models for adulteration detection in sesame oil: a comparative study of genetic algorithm, particle swarm optimization, and a hybrid GA-PSO approach. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy. 2026; 348: 127261. https://doi.org/10.1016/j.saa.2025.127261
[25] He Q, Leng J, Feng Z, Tan L, Dong L. Sustainable optimization of methanol production and distillation processes using an adaptive simulated annealing particle swarm algorithm based on composite sustainability index. Chemical Engineering Science. 2026; 320: 122553. https://doi.org/10.1016/j.ces.2025.122553
[26] Xie LH, Shao GN, Sheng ZH, Hu SK, Wei XJ, Jiao GA, et al. Rapid identification of fragrant rice using starch flavor compound via NIR spectroscopy coupled with GC-MS and Badh2 genotyping. International Journal of Biological Macromolecules. 2024; 281: 136547. https://doi.org/10.1016/j.ijbiomac.2024.136547
[27] Zhang J, Dai L, Zhuang R. Corn Seed Freezing Damage Identification of Different Sides Based on Hyperspectral Imaging and SPA-2DCOS Fusion Algorithm. Molecules (Basel, Switzerland). 2025; 30: 2178. https://doi.org/10.3390/molecules30102178
[28] Kazempour AM, Kiani S, Seyedi SRM, Nadamani AR. AI-Integrated Portable VIS-NIR Spectroscopy for On-Site Authen-

ticity Evaluation and Quality Grading of Mutton-Veal Minced Meat. Journal of Food Science. 2026; 91: e70891. https://doi.org/10.1111/1750-3841.70891

[29] Zou J, Gao H, Wang D, Chen Y, Deng S, Shi N, et al. Enabling rapid and accurate grand discrimination of flue-cured tobacco: a near-infrared hyperspectral and machine learning approach. Frontiers in Plant Science. 2026; 17: 1756218. https://doi.org/10.3389/fpls.2026.1756218

[30] Zuo ZT, Wang YZ, Yao ZY. FT-NIR Spectra of Different Dimensions Combined with Machine Learning and Image Recognition for Origin Identification: An Example of *Panax notoginseng*. ACS Omega. 2025; 10: 7242–7255. https://doi.org/10.1021/acsomega.4c10816

[31] Oliveira MF, Batista DB, Soares TF, Luna AS, Pinto L. Uncovering complex adulteration scenarios on unripe banana flour with hierarchical multiclass modelling of NIR data. Food Control. 2026; 181: 111748. https://doi.org/10.1016/j.foodcont.2025.111748

[32] Wu B, Bai HY, Sun LJ, Fang YR, Han ZB, Zhang SH. Varieties identification of soybean seed using visible/near-infrared spectroscopy based on ensemble wavelength selection method. Infrared Physics & Technology. 2026; 153: 106358. https://doi.org/10.1016/j.infrared.2025.106358

[33] Wu J, Nie J, Hu H, Xu X, Li C, Zhou H, et al. Characterizing Chinese saffron Origin, Age and grade using VNIR hyperspectral imaging and Machine learning. Food Research International (Ottawa, Ont.). 2025; 202: 115585. https://doi.org/10.1016/j.foodres.2024.115585

[34] Wang T, Zheng Y, Xu L, Yun YH. Comprehensive comparison on different wavelength selection methods using several near-infrared spectral datasets with different dimensionalities. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy. 2025; 331: 125767. https://doi.org/10.1016/j.saa.2025.125767

[35] Xu Y, Dong Y, Liu J, Wang C, Li Z. Combination of near infrared spectroscopy with characteristic interval selection for rapid detection of rice protein content. Journal of Food Composition and Analysis. 2025; 137: 106995. https://doi.org/10.1016/j.jfca.2024.106995

[36] Feng GQ, Gu Y, Zhang H, Zhou YN, Wang C, Luo B, et al. Near-infrared spectroscopy unmixing for abundance estimation of multi-source composite adulterated wheat flour. Industrial Crops and Products. 2026; 242: 122881. https://doi.org/10.1016/j.indcrop.2026.122881

[37] Unuvar A, Simsek S, Boyaci IH, Ohm JB, Manthey FA, Koksel H. Detection and quantitation of adulteration in flour and pasta by reversed-phase high-performance liquid chromatography combined with chemometrics. Food Chemistry. 2025; 474: 143159. https://doi.org/10.1016/j.foodchem.2025.143159