Hindawi
Journal of Spectroscopy
Volume 2022, Article ID 2440518, 14 pages
https://doi.org/10.1155/2022/2440518

Research Article

A New Method for Spectral Wavelength Selection Based on Multiple Linear Regression Combined with Ant Colony Optimization and Genetic Algorithm

Qing Huang, Heru Xue, Jiangping Liu, and Xinhua Jiang

College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China

Correspondence should be addressed to Heru Xue; xuehr@126.com

Received 21 August 2022; Accepted 6 October 2022; Published 22 November 2022

Academic Editor: Young Jong Lee

Wavelength selection is one of the key steps in quantitative spectral analysis, which reduces the computation time while also improving the prediction accuracy of the model. In this paper, we propose a wavelength selection algorithm based on the ant colony optimization (ACO), in which the absolute value of the regression coefficient of the multiple linear regression (MLR) model is used as the basis for evaluating the importance of wavelengths, and the absolute value of the regression coefficient after full wavelength MLR modeling is used as the initial pheromone value of the ant colony optimization (MLR-ACO). In each iteration, the absolute value of the regression coefficient corresponding to each wavelength of the individual with the highest fitness value is used as the basis for a pheromone update. The crossover operator is introduced in MLR-ACO (MLR-ACO-GA), and the individuals with the top 100 fitness values in MLR-ACO are used as the initial population of the genetic algorithm (GA). A selected frequency of wavelengths greater than the threshold among MLR-ACO individuals is calculated. A number of coarse interval points are generated according to the selected frequency, and a coarse crossover operation is performed at the coarse interval points. Fine crossover points are randomly generated within the coarse interval, and fine crossover operations are performed within the coarse interval to exploit the potential of combining excellent individuals in MLR-ACO with each other as much as possible. MLR-ACO can well solve the problem of traditional ACO initial pheromone scarcity, and MLR-ACO-GA can avoid MLR-ACO falling into a local optimum to a certain extent and be more flexible in the selection of the number of wavelengths, which can give full play to the advantages of MLR-ACO.

1. Introduction

Spectroscopy is widely used in the fields of agriculture [1, 2], medicine [3, 4], environment [5, 6], and food detection [7, 8] due to its speed, low cost, and nonpollution characteristics. With the advancement of modern spectroscopic instruments, the obtained spectral data contain tens to thousands of wavelengths and can reflect the subtle spectral differences of different constituents in the measured substances. However, the obtained data contain a large number of uncorrelated or redundant features with high collinearity, and these data features usually reduce the prediction accuracy of the model and worsen the experimental results [9, 10]. To solve this problem, many wavelength selection methods have been proposed, and many papers have demonstrated experimentally or theoretically that performing wavelength variable selection can lead to better prediction performance and significant computational time savings. Wavelength selection is a very important and essential key step before performing quantitative analysis [11–14].

In general, wavelength selection methods can be divided into two categories, one is wavelength point selection algorithms, such as successive projection algorithm (SPA) [15], competitive adaptive reweighting sampling (CARS) [16, 17], ant colony algorithm (ACO) [18, 19], genetic algorithm (GA) [20, 21], differential evolution algorithm (DE), sparrow search algorithm (SSA) [22], etc. Other kinds of wavelength interval selection algorithms, such as interval PLS (iPLS) [23], moving window PLS (MWPLS) [24], and some interval partial least squares based on optimization algorithm improvement, such as iPLS-genetic algorithm (GA-IPLS) [25], interval random frog algorithm (IFR) [26], etc. It can be seen that intelligent optimization algorithms are widely used not only in wavelength point selection but also in wavelength interval selection.

In recent years, more and more swarm intelligence optimization methods have been proposed and widely used in wavelength selection. In addition to those mentioned earlier, there are gray wolf optimization (GWO) [27], monarch butterfly optimization (MBO) [28], slime mould algorithm (SMA) [29], hunger games search (HGS) [30], Harris Hawks optimization (HHO) [31], artificial algae algorithm (AAA) [32], etc. All of these methods have achieved good results in the selection of feature wavelengths. Among these optimization algorithms, ACO has been widely studied because of its positive feedback mechanism, fast convergence speed, and high accuracy [18]. The ACO has high efficiency in solving complicated problems, but the traditional ACO also has many defects, such as a lack of initial pheromone and an easy tendency to fall into local optimal solutions.

To solve the problems such as the lack of an initial pheromone, Tong proposed using the importance projection coefficients of variables (VIP) under the full wavelength partial least squares regression (PLSR) model as the initial pheromone of the ACO algorithm and proposed the PLS-VIP-ACO wavelength selection method [33]. Based on Tong's research, Xiaoming et al. proposed an elite ACO based on the validity of variables, while combining forward selection methods to prefer feature wavelengths and using elite ant colony search [34]. Liu et al. proposed an improved adaptive update ACO to improve the convergence and global search capability of the traditional ACO [19]. However, it is worth noting that in their studies, the predictive performance of the PLSR model is used as the criterion for evaluating the selected subset of variables. In the iteration of the ACO, it is necessary to artificially set the values of latent variables in PLSR or set the values of latent variables to a certain range, which on the one hand cannot make the PLSR model optimal, and on the other hand, it takes a lot of time in the process of finding latent variables and the additional calculation of VIP coefficients increases the complexity of the algorithm. In order to improve these problems, this paper combines the multiple linear regression (MLR) method with the ACO and establishes the MLR model for the data at full wavelength, uses the absolute value of the regression coefficient of the MLR model as the criterion for evaluating the importance of wavelength and uses it as the value of the initial pheromone of the ACO to solve the problem of the lack of the initial pheromone of the ACO. To improve the problem that the traditional ACO easily falls into local optimum, the crossover operator in the genetic algorithm is introduced into MLR-ACO. The ten-fold root mean square error of cross-validation (RMSECV) of the MLR model with full wavelength data is calculated and this is used as the threshold value. We count the individuals in the ACO that are larger than the threshold value and calculate their frequency of being selected for each wavelength. Several coarse interval points are generated according to the characteristics of the selected frequencies of wavelengths, coarse crossover operation is performed at the coarse interval points, fine crossover points are randomly generated within the coarse interval, and a fine crossover operation is performed at the fine crossover points. Among them, the coarse crossover is to discover the advantages of coarse intervals combining with each other among different excellent wavelength combinations, and the fine crossover is to explore whether the same intervals combining with each other can produce better subsets, to further exploit the advantages of MLR-ACO.

2. Method and Theory

2.1. Multiple Linear Regression Model. MLR is a common calibration method in quantitative spectroscopy, which focuses on the correlation between an attribute of interest and each wavelength [35–37]. The basic form is

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots w_m x_m + \ldots w_n x_n + e. \tag{1}$$

It is generally written in vector form as

$$Y = W^T X + e, \tag{2}$$

where y denotes the attribute value of interest, x_m denotes the reflectance of the corresponding wavelength, w denotes the corresponding regression coefficient, and e is the error, which follows a normal distribution with the mean value of zero. n denotes the number of wavelengths.

The regression coefficient W (W=$\mathrm{w_1,w_2,...w_n}$) is estimated using the least squares method, and the estimated amount of W is denoted as W^* $(W^* = w^*{}_1,\ w^*{}_2 \ldots, w^*{}_n)$ which can be obtained from the least squares method:

$$W^* = \left(X^T X\right)^{-1} X^T Y. \tag{3}$$

Using the regression vector W^* to predict Y, the predicted value of Y can be calculated by the following equation:

$$Y^* = XW^*. \tag{4}$$

From the appeal equation, we know that X is a fixed value and the value of Y^* is determined by W*. When the absolute value of w* is larger, the greater the influence on y. The absolute value of $w^*{}_i$ reflects the contribution of wavelength i to y. It can be said that the larger the $|w^*{}_i|$ is, the more important the *i-th* wavelength is. Therefore, the absolute value of the regression coefficients obtained by building MLR models for full wavelength data is used as the initial pheromone value of the ACO algorithm.

2.2. MLR-ACO. Inspired by the traditional ACO combining the MLR algorithm regression coefficients with the ACO, the main steps are as follows:

2.2.1. Parameters of the Initialization Algorithm

(1) N: the number of iterations of the ACO needs to be large enough for the algorithm to achieve convergence.

(2) M: the number of ants.

(3) Initial pheromone matrix (*IP*): the data at full wavelength were modeled as MLR, and the absolute value of the regression coefficient (β) corresponding to each wavelength was calculated and used as the pheromone concentration value of the initial pheromone matrix.

(4) HAS: storing the feature wavelengths selected for each ant in the HAS matrix.

(5) HAVE: the feature wavelength of storage ants without selection.

(6) V-MAX: maximum number of feature wavelengths selected per ant.

(7) Q: an important factor.

(8) ρ: pheromone volatile factor.

(9) T: the threshold value.

(10) C: the contribution matrix, the combinations of feature wavelengths selected by ants whose fitness value is greater than the threshold value which is stored in the contribution matrix.

2.2.2. Ant Chooses the Path. Each ant randomly selects a wavelength as the path start point and stores it in the HAS matrix. The HAVE matrix removes the wavelength, and the *IP* matrix removes the pheromone value of the wavelength. The roulette algorithm is used to select the next feature wavelength until the number of selected feature lengths reaches *V*-MAX. The probability of each wavelength being selected is as follows:

$$\frac{P_i(n) = \tau^{\mathrm{i}}_{\mathrm{have}}(n)}{\sum \tau_{\mathrm{have}}}, \tag{5}$$

where P_i is the probability that wavelength i in the HAVE matrix is selected and ${\tau^i}_{\mathrm{have}}$ is the concentration value of the pheromone at wavelength i in the HAVE matrix.

2.2.3. Calculation of the Fitness Value. The combined data of feature wavelengths selected by each ant are built into the MLR model, and the RMSECV of the MLR model is used as the basis for the calculation of the fitness value of the ACO. The fitness value (F) is calculated as shown in equation (6). The ants whose fitness value is greater than the threshold value are selected into the contribution matrix.

$$F = \frac{Q}{RMSECV}. \tag{6}$$

2.2.4. Pheromone Update. The pheromone is updated according to the foraging behavior of ants in the biological world. When all ants finish the iteration, the ant with the highest contemporary adaptation value is selected, and the absolute value of the regression coefficient of the corresponding wavelength after its MLR modeling is used as the basis for the pheromone update, and the pheromone of the corresponding wavelength is strengthened according to the pheromone update formula, and the wavelengths that are not selected will slowly become smaller because of the pheromone volatile concentration. The specific pheromone update equation is as follows:

$$\begin{array}{ll} (1-\rho)\tau^i_{n-1} + |\beta i|, & \text{wavelength } i \text{ is selected,} \\ (1-\rho)\tau^i_{n-1}, & \text{others.} \end{array} \tag{7}$$

We repeat steps 2–4 until the set maximum number of ant colony iterations is reached.

The optimal wavelength combination is selected, and the feature wavelength combination selected by the ant with the highest fitness value among all individuals is the final selection.

2.3. Introduction of Crossover Operator in MLR-ACO. In GA, a crossover operator operation can produce an even better offspring that incorporates the characteristics of both parents. The most common crossover method is the single-point crossover operation, which generates a random crossover point and swaps the feature wavelengths before and after the point between the two parents to generate two new combinations of feature wavelengths. The combination of feature wavelengths with higher fitness values is saved for comparison with the parent. Since its intersection points are randomly generated, the stability of the generated children is not very good. To improve this problem, an improved intersection algorithm is proposed in this paper for spectral features. In the study of the optical parameters of milk by Jun [38], it is known that the absorption coefficients of different components of the same measured sample vary at different wavelengths and are strongly influenced by the content of that component. The absorption coefficients of different components also interact with each other, and it is difficult for us to tell the wavelength interval of the absorption coefficient corresponding to the component of interest directly from the raw spectral image. In calculating the selected frequencies of each wavelength in the MLR-ACO contribution matrix, it is found that the selected frequencies of wavelengths also show corresponding peaks and valleys within a certain wavelength interval. The selected frequency of a wavelength indicates the importance of that wavelength for the property of interest that we need to measure. For this reason, we divide the full wavelength into coarse intervals according to the troughs of the selected frequencies of wavelengths based on the selected frequency map of wavelengths. Here, the valley points of the selected frequencies of wavelengths are used as the coarse crossover points of the crossover operator in the genetic algorithm. The best 100 individuals generated in the MLR-ACO iteration are used as the initial population of the genetic algorithm, and the coarse crossover operation is performed by the coarse crossover point. All individuals are combined as much as possible to discover the best individuals generated by combining different wavelength intervals, and in order to discover the advantages of combining within the same wavelength interval, a fine crossover is randomly generated within the coarse interval, and the two parents are combined with each other in the same interval wavelength. Finally, four offspring were generated cumulatively, and the two wavelength combinations with the highest fitness values were left by comparing them with their two parents. After 30 iterations, the wavelength combination selected by the individual with the highest fitness value is the final selection. The flow chart of MLR-ACO-GA is shown in Figure 1.

3. Data and Software

3.1. Wheat Protein Dataset. The dataset is from the International Diffuse Reflectance Conference (IDRC) and can be downloaded at https://www.cnirs.org/content.aspx?page_id=22&club_id=409746&module_id=239453. It contains spectral data of 248 kinds of wheats measured by 3 spectrometers. The wavelength range is 850–1050 nm, the interval is 2 nm, and there are 100 bands in total. The data also measured protein content values for each wheat sample, which varied from 7.97% to 18.69% with an average of 13.64%. Spectral data and protein content values measured by Instrument C were used in this study.

3.2. Cereal Cheese Protein Dataset. The dataset is downloaded from https://eigenvector.com/resources/data-sets/#grain-sec. In this dataset, the U.S. Department of Energy uses a mixture of three substances to predict the content of casein, glucose, lactic acid, and water in the mixture. Among them, the content of casein varies from 0% to 88.83%, with an average of 29.61%. The value of casein and the measured spectral data are used in this study.

3.3. Corn Protein Dataset. The dataset can be downloaded from the website http://software.eigenvector.com/Data/Corn/index.html. It consists of 80 corn samples measured by three different near-infrared spectrometers. The instruments used are m5, mp5, and mp6. The wavelength range is 1100–2498 nm, and the interval is 2 nm. The data also measures the moisture, oil, protein, and corn content of each corn sample. The spectral data and protein values measured by the M5 near-infrared spectrometer were used in this study. The protein content varied from 7.65% to 9.71%, with an average of 8.66%.

3.4. Equipment and Software. We use a general-purpose computer; the CPU is Intel (R) Core (TM) i5-6500 CPU @ 3.20 GHz 3.19 GHz, the memory is 8 GB, the operating system is Windows 10, and all calculations are implemented on the Python 3.7 platforms.

4. Results and Discussion

Three publicly available datasets of wheat protein, grain casein, and corn protein were used to evaluate MLR-ACO and MLR-ACO-GA, which were eventually compared with five established feature selection algorithms, CARS, SPA, ACO, GA, and DE. SPA is based on vector projection analysis, which compares the magnitude of projection vectors between different wavelengths to find the combination of feature variables with the lowest information redundancy in the spectral data and selects the optimal combination of feature variables by correcting the model. SPA can minimize the collinearity between variables and largely reduce the number of wavelengths needed for modeling. CARS imitates the principle of survival of the fittest in Darwinian evolutionary theory combined with PLS model regression coefficients, and in each iteration, samples are drawn by monte Carlo sampling, and variables with small absolute values of regression coefficients are forced to be removed by the exponential decay function (EDF). The adaptive weighted sampling method is used to further filter the feature wavelengths, and the set with a large value of the regression coefficient weights is retained to create the PLSR model and calculate the RMSECV of this feature wavelength combination. After several iterations, the feature wavelength combination with the lowest RMSECV value is selected as the optimal subset. ACO screens the feature wavelengths by simulating the foraging behavior of ants. It uses the RMSECV of the calibration model to judge the goodness of this combination of feature wavelengths and updates the pheromones of the corresponding wavelengths according to the RMSECV after each iteration. In this paper, only the ants with the highest fitness value in each generation are selected to update the pheromone matrix. From the experiments, it is found that updating the pheromone matrix with the ant with the highest fitness value is much more desirable than updating the pheromone matrix with all ants. By imitating the mechanism of superiority and inferiority in nature, GA iterates repeatedly through the selection operator, crossover operator, variation operator, and three operators to finally select the combination of feature wavelengths with the highest fitness value as the optimal feature wavelength combination. The DE algorithm is very similar to the genetic algorithm in that it also includes the operations of mutation, crossover, and selection, but the specific definitions of these operations are different from those of the genetic algorithm. In this paper, the DE algorithm uses floating-point vector coding to generate the initial population, while the GA uses binary coding. In this paper, SPA, ACO, GA, and DE all use MLR regression models as calibration models, and RMSECV as the evaluation criterion for a subset of variables. However, CARS is used to improve the prediction accuracy for the PLSR model, so the MLR and PLSR models are built for the combination of feature wavelengths screened by CARS and compared with the RMSECV of the MLR and PLSR models for the combination of feature wavelengths screened by both MLR-ACO and MLR-ACO-GA algorithms. The parameter settings of each algorithm were set according to their respective recommendations, 30 tests were performed on each data set, and the RMSECV was recorded.

4.1. Parameter Configuration. In MLR-ACO, there are five parameters that affect the performance of the algorithm. Before the method is used for different data sets, the parameters should first be optimized. The number of iterations N was set to 50, 100, 150, and 200 in order, with N set too small for the algorithm to achieve a fit and too large for N to increase the time complexity of the computation. It was

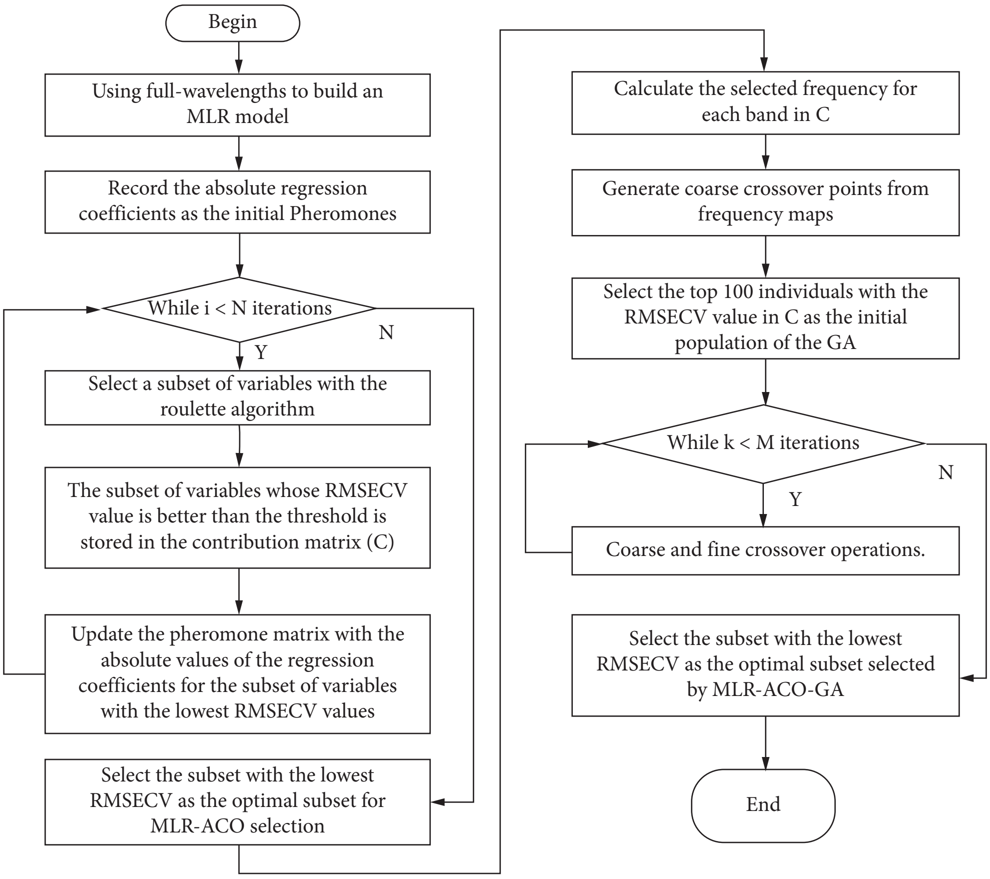
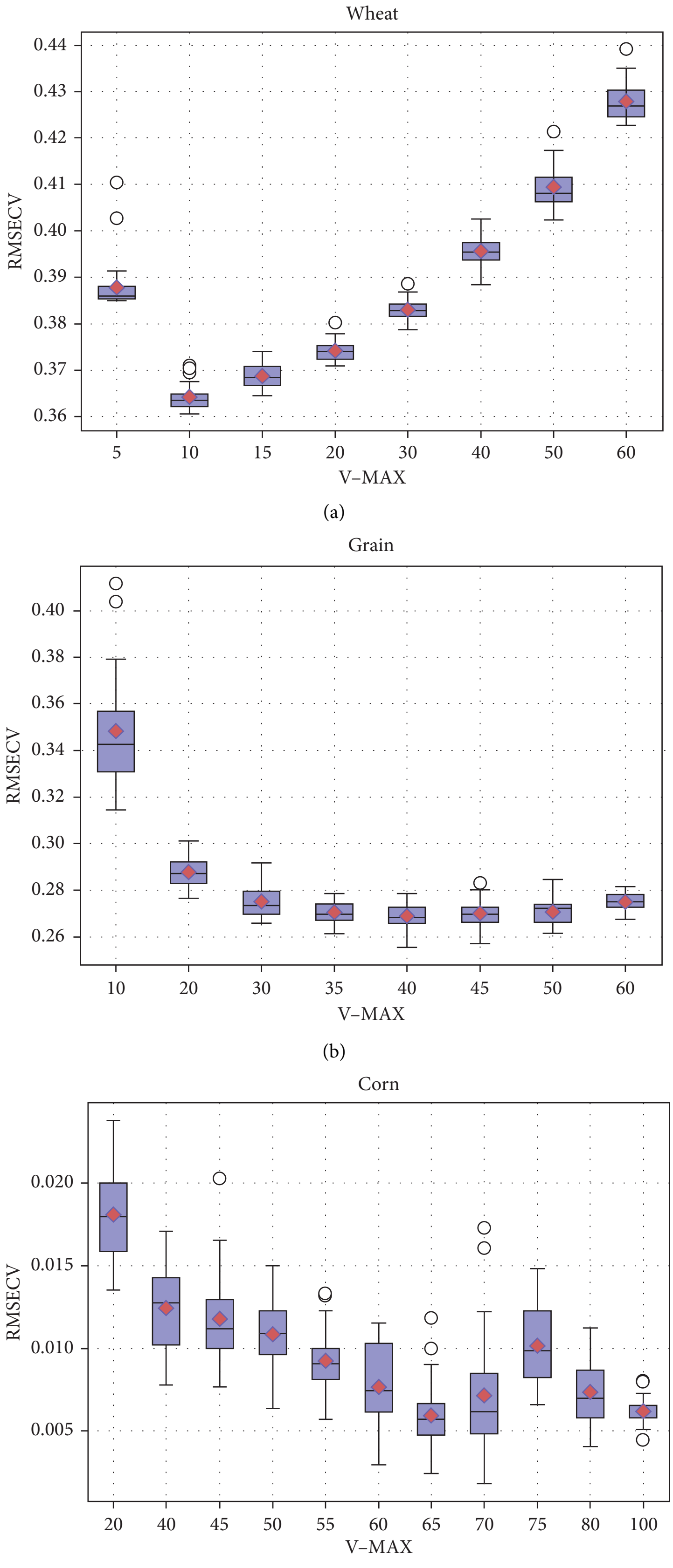

Begin
Using full-wavelengths to build an MLR model
Record the absolute regression coefficients as the initial Pheromones
While i < N iterations
Y
N
Select a subset of variables with the roulette algorithm
The subset of variables whose RMSECV value is better than the threshold is stored in the contribution matrix (C)
Update the pheromone matrix with the absolute values of the regression coefficients for the subset of variables with the lowest RMSECV values
Select the subset with the lowest RMSECV as the optimal subset for MLR-ACO selection
Calculate the selected frequency for each band in C
Generate coarse crossover points from frequency maps
Select the top 100 individuals with the RMSECV value in C as the initial population of the GA
While k < M iterations
Y
N
Coarse and fine crossover operations.
Select the subset with the lowest RMSECV as the optimal subset selected by MLR-ACO-GA
End

Figure 1: Flow chart of MLR-ACO-GA algorithm.

found through experiments that the MLR-ACO algorithm reached its fit when N was set to 50 in the wheat protein and grain casein datasets. In the corn protein dataset, the MLR-ACO algorithm reached the fit only when N was set at 100. This is because the corn dataset has a higher number of wavelengths compared to the other two datasets and requires a longer iteration time. The larger the number of ants M, the higher the accuracy of the algorithm, and also the higher the time complexity. The number of ants in all three datasets was finally set at 80. The pheromone volatility factor ρ was set to 0.3, 0.5, and 0.7, respectively. ρ was too small, the ants might lose the global search ability, and ρ was too large, which would affect the convergence speed. After experiments, it was found that satisfactory results could be achieved when ρ was taken as 0.3 and 0.5. In the wheat protein and grain casein datasets, the results were slightly better when ρ was taken as 0.3, and in the corn protein dataset, the results were better when ρ was taken as 0.5. Q is the pheromone significance factor, and Q is set to 1 for all three data sets. V_MAX is one of the most important parameters in the MLR-ACO algorithm. If V-MAX is set too large, some irrelevant information variables cannot be eliminated, which will reduce the computational efficiency. If V-VAX is set too small, some important variables may be excluded, and the accuracy of the prediction model will be reduced. In the wheat protein and grain casein datasets, V-MAX was first set to 10, 20, 30, 40, 50, and 60 in that order, and after determining the optimal value in this interval, the final *V-MAX* value was determined in intervals of 5 within the value range. In the corn protein data set, the *V-MAX* values were first set to 20, 40, 60, 80, and 100 in that order, and after determining the optimal value in this interval, the final *V-MAX* value was determined in intervals of 5 within this value range. Figure 2 shows the box line plots of the RMSECV for 30 experiments with different *V-MAX* values for the three data sets, respectively. We can know that the optimal values of *V-MAX* for the three data sets are 15, 40, and 65, respectively. When *V-MAX* is set too small, some important wavelengths cannot be selected, which reduces the prediction performance. When *V-MAX* is set too large, the model accuracy is reduced instead, which can be seen as proof of Occam's razor theory that better prediction performance can be achieved by using fewer wavelengths [39].

The population size is set to 100 from the ants with the top 100 fitness values after the MLR-ACO iterations are completed. The number of iterations is set to 30 because the initial population is already excellent and the number of iterations does not need to be set very widely. The coarse crossover was calculated based on the individuals with fitness values greater than T produced by the 30 MLR-ACO iterations, with a threshold T equal to the reciprocal of the full wavelength MLR modeling RMSECV. In addition, the coarse crossover probability and fine crossover probability were set to 0.5 and 1, respectively. The frequency of



(a)

(b)

(c)

Figure 2: Boxplots of RMSECV under 30 different *V-MAX* values for MLR-ACO testing on three datasets: (a) wheat protein dataset, (b) grain casein dataset, and (c) corn protein dataset.

individuals being selected in the contribution matrix of the three datasets and the coarse crossover point settings are shown in Figure 3. The wheat protein dataset generated 20 coarse crossover points, the cereal casein dataset generated 17 coarse crossover points, and the corn protein dataset generated 42 coarse crossover points.

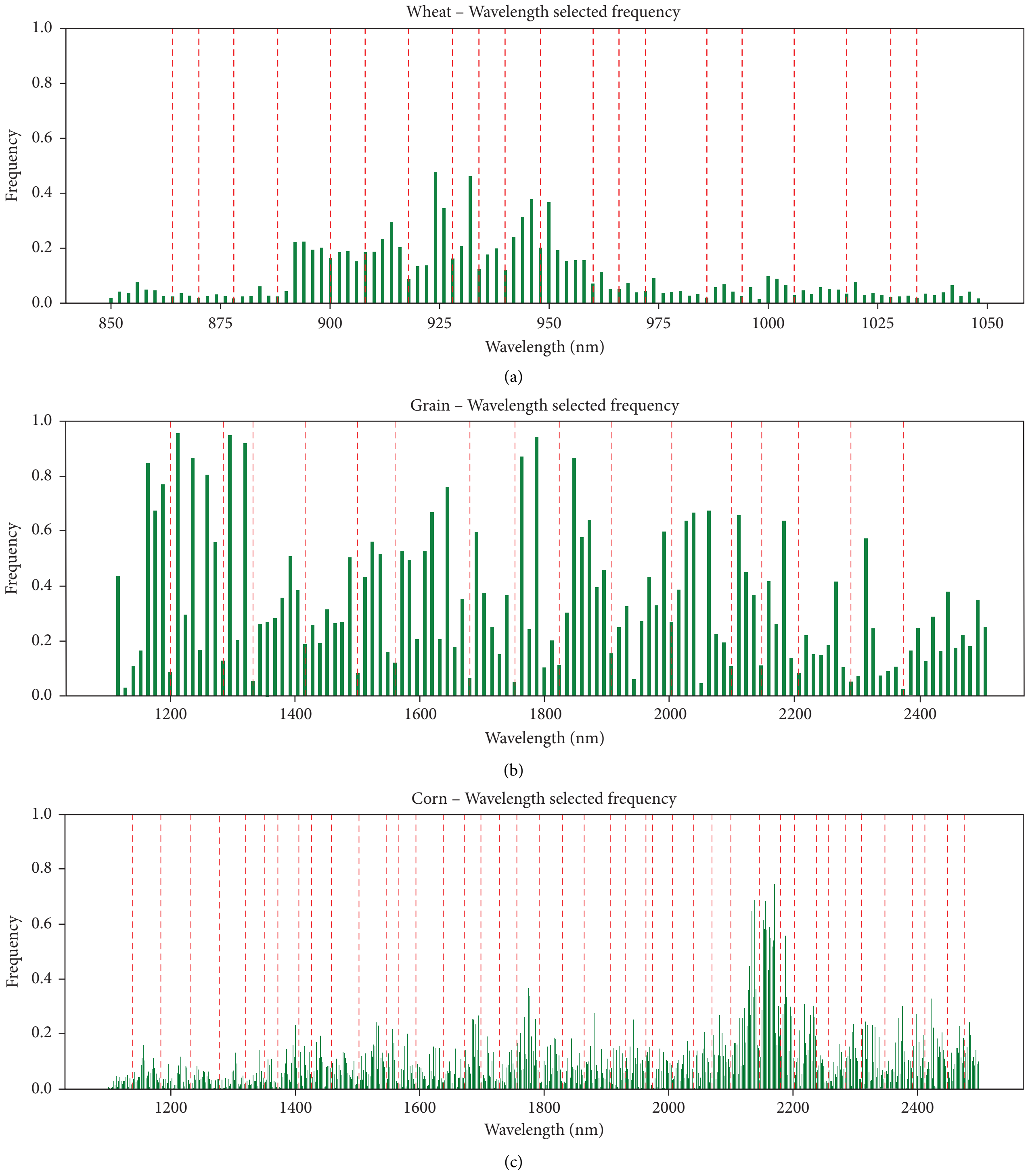
4.2. Wheat Protein Dataset. For wheat protein data, SPA, CARS, ACO, GA, and DE were used for comparison with MLR-ACO and MLR-ACO-GA, respectively. Each algorithm, except SPA was run 30 times and its RMSECV was recorded, and the results of the full wavelength model are listed together as shown in Table 1. From Table 1, it can be seen that MLR-ACO, MLR-ACO-GA, SPA, and ACO can be selected for a combination of variables with a smaller number of wavelengths, but the performance of SPA prediction is not very good. Except for SPA, the prediction performance of all algorithms is better than the full wavelength, where CARS, ACO, GA, DE, MLR-ACO, and MLR-ACO-GA have 20.56%, 38.38%, 31.04%, 35.89%, 40.23%, and 40.32% lower mean RMSECV compared to the full wavelength, respectively. MLR-ACO-GA performs CARS, GA, and DE which not only require a larger number of bands compared to MLR-ACO, MLR-ACO-GA, and ACO algorithms but also the model accuracy cannot be further improved. The MLR-ACO algorithm can jump out of the local optimum after adding the crossover operator. The standard deviation is reduced from 0.0026 to 0.0023, and the algorithm becomes more stable.

The spectrograms of wheat protein data and the frequencies of the seven different methods selected for the wavelengths on the wheat protein data set for the test experiments are shown in Figure 4. As can be seen from the observation of Figure 4, it is not directly evident from the spectral images that the spectral absorption bands are related to the frequencies of the variables selected, which once again should confirm the conclusion in 2.3. From the figure, the absorption spectra selected by MLR-ACO and MLR-ACO-GA are roughly the same, mainly including the complex regions of protein molecular characteristic absorption such as the stretching vibration or bending vibration of C-H, N-H, and O-H bonds, their interaction, and the influence of the external environment. The selected absorption bands of MLR-ACO and MLR-ACO-GA are mainly concentrated near 900 nm, 925 nm, and 950 nm. A few wavelengths are also selected near 860 nm, 1000 nm, and 1025 nm. Among them, 900 nm corresponds to the quadruple frequency absorption band of C-H and 950 nm corresponds to the triple frequency absorption band of the O-H bond. And the other selected wavelengths are difficult to match accurately with a certain chemical bond. However, the experimental results show that these wavelengths play an important role in the modeling. It is worth noting that the high-frequency wavelengths selected by the two algorithms, MLR-ACO and MLR-ACO-GA, basically match those selected by the other five algorithms, but MLR and MLR-ACO-GA discard more irrelevant information variables.

4.3. Grain Protein Dataset. For the grain protein data, SPA, CARS, ACO, GA, and DE were used for comparison with MLR-ACO and MLR-ACO-GA, respectively. Each algorithm except SPA was run 30 times and its RMSECV was recorded, and the results of the full wavelength model are listed together as shown in Table 2. From Table 2, it can be seen that the final modeling results of all seven wavelength selection algorithms outperformed the full wavelength model, and the mean values of SPA, CARS, ACO, GA, DE, MLR-ACO, and MLR-ACO-GA were reduced by 37.28%, 48.61%, 54.04%, 49.75%, 52.51%, 57.22%, and 57.46%. This shows that feature wavelength selection is very important before performing quantitative correction models. Among these seven algorithms, MLR-ACO-GA has the best prediction performance, and MLR-ACO is the second, but the number of feature wavelengths required is not the least, which is because the MLR-ACO algorithm believes that when the number of wavelengths is taken to be about 30, not all effective information variables can be selected to make the model effect optimal. From Figure 2(b), we can see that when the *V-MAX* parameter is set to 20, the prediction effect of CARS, SPA, and ACO can already be achieved, but the MLR-ACO algorithm believes that a *V-MAX* of 20 is not the optimal parameter value. Of course, *V-MAX* can be set to 20 if considered from the perspective of time complexity.

The spectrograms of the cereal casein data and the frequencies of the variables selected by the seven different methods for the experimental tests on the cereal casein dataset are shown in Figure 5. It can be seen from the figure that the high-frequency wavelengths selected by the six algorithms are the same, mainly around 1152 nm, 1248 nm, around 1500 nm, 1752 nm, and 2028 nm. Among the three algorithms with better prediction performance, CARS, MLR-ACO, and MLR-ACO-GA, the selected frequencies of the eight bands of 1152 nm, 1164 nm, 1200 nm, 1224 nm, 1248 nm, 1752 nm, 1776 nm, and 2028 nm are more than 70%. Compared with the CARS algorithm, MLR-ACO, MLR-ACO-GA, ACO, and GA additionally select wavelengths in the range of 1344–1392 nm and some other wavelengths. Among them, the spectral regions near 1152 nm, 1500 nm, 1752 nm, and 2028 nm correspond to the C-H bond triple frequency absorption band, N-H bond double frequency absorption band, C-H double frequency absorption band, and O-H combined frequency absorption band, respectively. The other chosen wavelengths are difficult to match exactly to a particular chemical bond, but experimental results show that these wavelengths play an important role in modeling. The selected wavelengths of the swarm intelligence class algorithm are more dispersed because the total number of wavelengths is only 117. The swarm intelligence algorithm has a good global search capability and can exploit the advantages of the combination of different wavelengths as much as possible.

4.4. Corn Protein Dataset. For the corn protein data, SPA, CARS, ACO, GA, and DE were compared with MLR-ACO, MLR-ACO-GA, and each algorithm was run 30 times, except SPA and its RMSECV was recorded, and the results of



(a)

(b)

(c)

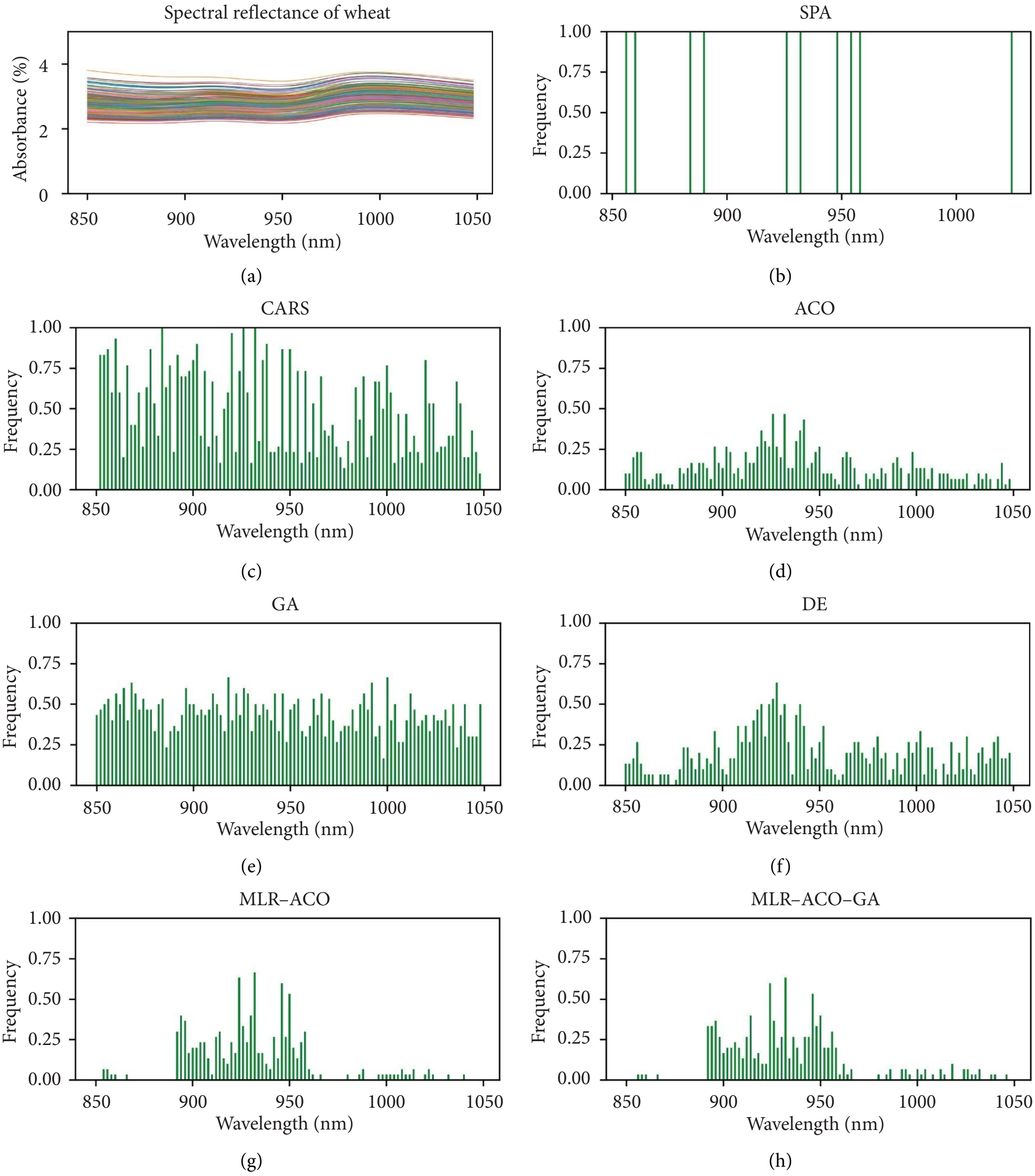
Figure 3: Frequency of selected wavelengths for individuals with MLR-ACO greater than the threshold for different datasets: (a) wheat protein dataset, (b) grain casein dataset, and (c) corn protein dataset.

the full wavelength model are presented together as shown in Table 3. All the algorithms except SPA outperformed the full wavelength prediction. The mean values of CARS, ACO, GA, DE, MLR-ACO, and MLR-ACO-GA were reduced by 40.15%, 62.45%, 41.80%, 55.08%, 91.91%, and 92.52%, respectively, compared to the full wavelength RMSECV. There are 700 wavelengths in the corn protein dataset, which is about 7 times the number of wavelengths of cereal casein and wheat protein, and the advantages of MLR-ACO and MLR-ACO-GA over other algorithms are more prominent as the number of wavelengths increases. SPA has a high RMSECV although only two bands were collected, and it is clear that SPA is not applicable to the corn protein dataset. ACO, GA, and DE have good prediction results, but the number of

Table 1: The results of different methods on the wheat protein dataset.

Method	RMSECV[a]	RMSECV[b]	nVar[a]	nVar[b]
Full	0.6093	—	100	—
SPA	0.6377	—	10	—
CARS	0.4839 ± 0.0607	0.4076–0.6070	49.03 ± 26.25	11–99
ACO	0.3754 ± 0.0036	0.3667–0.3810	13.76 ± 3.14	8–21
GA	0.4201 ± 0.0071	0.4044–0.4350	43.96 ± 7.44	30–59
DE	0.3906 ± 0.0029	0.3845–0.3964	20.16 ± 3.15	14–25
MLR-ACO	0.3641 ± 0.0026	0.3605–0.3708	10	—
MLR-ACO-GA	0.3636 ± 0.0023	0.3603–0.3694	10.5 ± 1.6	8–14

[a]Statistical results of mean ± standard deviation of 30 replicate simulations of different methods. [b]Statistical results of the variation range of 30 repeated simulations with different methods.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

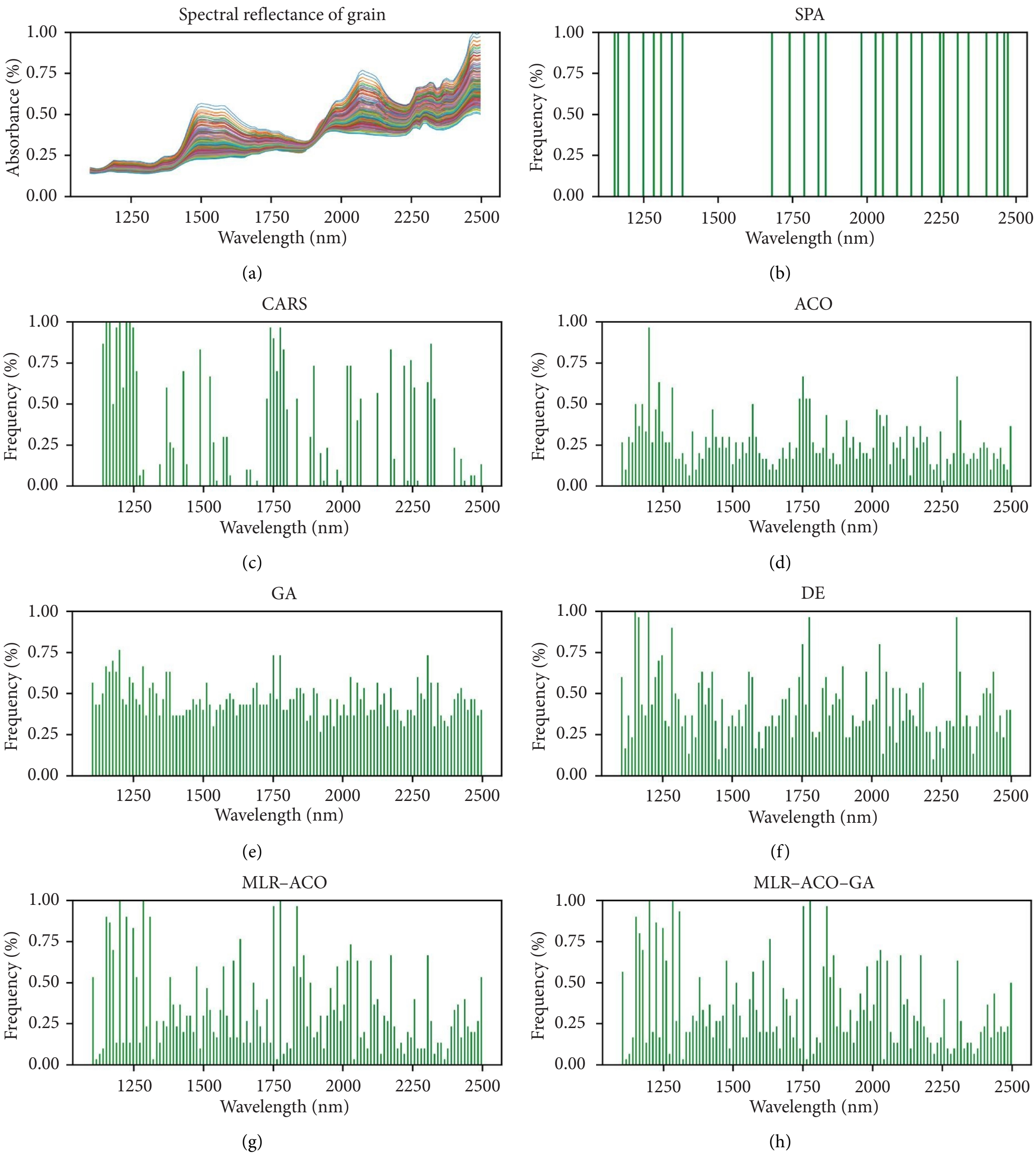
(h)

Figure 4: (a) Spectrograms of the wheat dataset; (b–h) frequencies of variables selected by different wavelength selection methods on the grain dataset. (b) SPA, (c) CARS, (d) ACO, (e) GA, (f) DE, (g) MLR-ACO, and (h) MLR-ACO-GA.

Table 2: The results of different methods on the grain protein dataset.

Method	RMSECV[a]	RMSEC[b]	nVAR[a]	nVAR[b]
Full	0.6288	—	117	—
SPA	0.3943	—	27	—
CARS	0.3231 ± 0.0231	0.2870–0.3900	30.93 ± 6.86	22–40
ACO	0.2889 ± 0.0060	0.2805–0.3088	31.66 ± 4.28	21–43
GA	0.3159 ± 0.0102	0.2791–0.3332	54.2 ± 45.37	43–66
DE	0.2985 ± 0.0035	0.2876–0.3047	49.43 ± 4.26	42–60
MLR-ACO	0.2689 ± 0.0049	0.2552–0.2759	40	—
MLR-ACO-GA	0.2674 ± 0.0043	0.2552–0.2759	40.46 ± 1.33	37–43

[a] Statistical results of mean ± standard deviation of 30 replicate simulations of different methods. [b] Statistical results of the variation range of 30 repeated simulations with different methods.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 5: (a) Spectrograms of the grain dataset; (b–h) frequencies of variables selected by different wavelength selection methods on the grain dataset. (b) SPA, (c) CARS, (d) ACO, (e) GA, (f) DE. (g) MLR-ACO, and (h) MLR-ACO-GA.

Table 3: The results of different methods on the corn protein dataset.

Method	RMSECV[a]	RMSECV[b]	nVAR[a]	nVAR[b]
Full	0.0735	—	700	—
SPA	0.4250	—	2	—
CARS	0.0440 ± 0.0314	0.0202–0.1483	50.16 ± 14.73	29–83
ACO	0.0276 ± 0.0037	0.0201–0.0338	150.9 ± 12.03	119–172
GA	0.0428 ± 0.0028	0.0386–0.0503	341.3 ± 13.46	317–365
DE	0.0330 ± 0.0013	0.0359–0.0306	329.06 ± 12.45	308–357
MLR-ACO	0.0059 ± 0.0020	0.0024–0.0118	65	—
MLR-ACO-GA	0.0055 ± 0.0018	0.0024–0.0113	67.2 ± 1.6	62–69

[a]Statistical results of mean ± standard deviation of 30 replicate simulations of different methods. [b]Statistical results of the variation range of 30 repeated simulations with different methods.

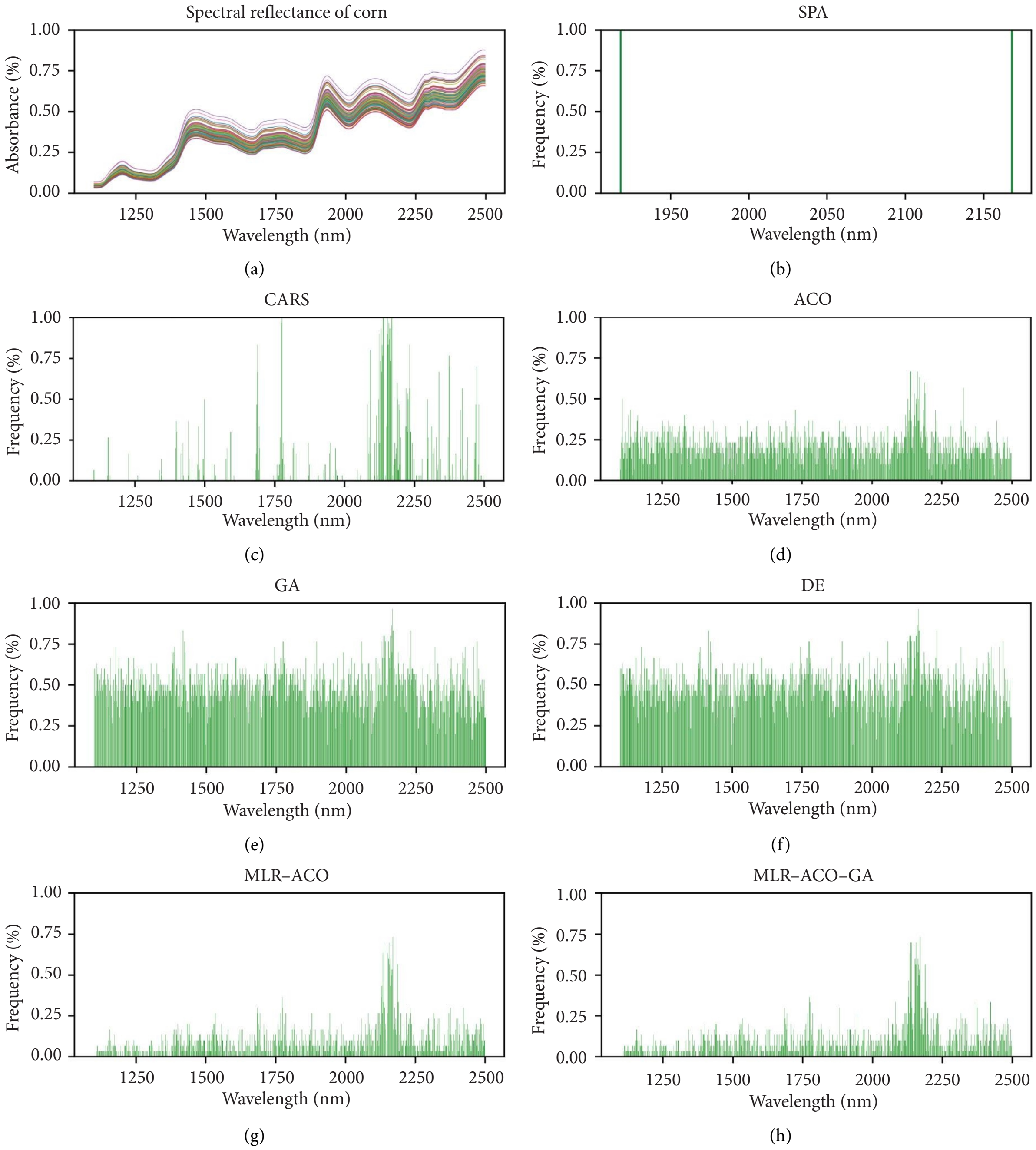


Figure 6: (a): Spectrograms of the corn protein dataset; (b–h) frequencies of variables selected on the grain dataset by different wavelength selection methods. (b) SPA, (c) CARS, (d) ACO, (e) GA, (f) DE, (g) MLR-ACO, and (h) MLR-ACO-GA.

selected wavelengths is much higher compared to the other methods. The MLR-ACO and MLR-ACO-GA perform best in terms of prediction accuracy. The number of selected wavelengths is slightly more than CARS, but the model accuracy is higher and the stability is also the best. It is worth noting that, observing Figure 2(c), it can be seen that when the *V-MAX* parameter in MLR-ACO is set to 20, the average value of RMSECV is 0.0181, which is lower than SPA, CARS, ACO, GA, and DE. In practical applications, the value of the *V-MAX* parameter can be set according to both time demand and accuracy demand.

The spectrograms of the corn protein data and the frequencies of the wavelengths selected by the seven different methods for the experiments on the corn protein data set are shown in Figure 6. The wavelengths selected by the six algorithms are mainly around 1750 nm, 1776 nm, 2168 nm, and 2374 nm. The wavelengths selected by the six algorithms are mainly around 1750 nm, 1776 nm, 2168 nm, and 2374 nm. Among them, 1750 nm corresponds to the C–H bond double frequency absorption band, 2168 nm corresponds to the N–H bond frequency absorption band, and 2374 nm corresponds to the C–H bond frequency absorption band. As can be seen from the figure, the wavelengths selected by the CARS algorithm are relatively concentrated, while the ACO, GA, and DE algorithms are relatively divergent, which for ACO is due to the lack of an initial pheromone. For GA and DE, this is because the random population is randomly generated and the final result is directly related to the initial population. The improved MLR-ACO and MLR-ACO-GA algorithms just make up for the defects of GA and ACO. On the other hand, the bionic algorithm has better global search capability and can exploit the advantages of combination between different bands as much as possible. It can be observed in the frequency diagram that each algorithm produces certain peaks and valleys in frequency for a certain wavelength interval, and the positions of the peaks and valleys of these six algorithms are the same.

4.5. Comparison of the Results of the PLS Correction Model Established by CARS, MLR-ACO, and MLR-ACO-GA. The results of the PLS correction models for the combinations of the selected feature variables for 30 tests of the MLR-ACO and MLR-ACO-GA and CARS algorithms are shown in Table 4. In the wheat protein dataset, the accuracy of PLS correction models for the selected combinations of variables in CARS, MLR-ACO, and MLR-ACO-GA was better compared to the full wavelength, and the mean value of RMSECV was reduced by 3% in CARS, 15% in MLR-ACO, and 10% in MLR-ACO-GA, respectively, compared to the full wavelength. MLR-ACO predicted the best results and was more stable. In the cereal casein dataset, the accuracy of the RLSR correction models for the selected combinations of variables in CARS, MLR-ACO and MLR-ACO-GA were better compared to the full wavelength, with a 25.56% reduction in the mean value of RMSECV for CARS compared to the full wavelength, and a 37.32% and 35.17% reduction in RMSECV for MLR-ACO and MLR-ACO-GA compared to the full wavelength. Both MLR-ACO and MLR-ACO-GA performed better due to CARS. in the maize protein dataset, the accuracy of RLS correction modeling was also better for the combination of variables selected for CARS, MLR-ACO and MLR-ACO-GA compared to the full wavelength, with a 79.20% reduction in CARS compared to the full wavelength RMSECV mean, and a 79.20% reduction in MLR-ACO and MLR-ACO-GA compared to the full wavelength RMSECV mean. ACO and MLR-ACO-GA reduce the full wavelength RMSECV values by 78.27% and 78.42% compared to the full wavelength, respectively. CARS performs the best and is the most stable among the three algorithms, but the minimum RMSECV values of MLR-ACO and MLR-ACO-GA are better than CARS. Apparently, the combination of feature wavelengths selected by MLR-ACO and MLR-ACO-GA can also achieve good results in the PLSR model.

Table 4: Results of PLS modeling with different methods on different datasets.

Dataset	Method	RMSECV[a]	RMSECV[b]
Wheat	Full	0.4305	—
	CARS	0.4188 ± 0.0142	0.3893–0.4356
	MLR-ACO	0.3654 ± 0.0034	0.3607–0.3732
	MLR-ACO-GA	0.3850 ± 0.0243	0.3655–0.4517
Grain	Full	0.4433	—
	CARS	0.3321 ± 0.0298	0.2889–0.4293
	MLR-ACO	0.2779 ± 0.0119	0.2576–0.3151
	MLR-ACO-GA	0.2874 ± 0.0186	0.2652–0.2889
Corn	Full	0.1108	—
	CARS	0.0230 ± 0.0042	0.0168–0.0372
	MLR-ACO	0.0245 ± 0.0078	0.0115–0.0424
	MLR-ACO-GA	0.0239 ± 0.0070	0.0099–0.0392

[a]Statistical results of mean ± standard deviation of 30 replicate simulations of different methods. [b]Statistical results of the variation range of 30 repeated simulations with different methods.

5. Conclusion

In this paper, we propose an improved algorithm based on the ant colony algorithm, combining ACO with MLR and adding the crossover operator of the genetic algorithm to MLR-ACO to combine them into MLR-ACO-GA. The MLR-ACO makes up for the defects of the original ant colony algorithm well, and the MLR-ACO-GA further exploits the advantages of the MLR-ACO-GA algorithm. Compared with other methods, these two algorithms are highly accurate and require fewer wavelengths, but require more time to complete the iterations. Our future work will try to solve this problem and apply the algorithm to practical applications. It is worth noting that these two algorithms can be applied to the selection of feature wavelengths for NIR spectral data and can also be extended to other data requiring quantitative analysis for the selection of feature wavelengths.

Data Availability

The data used in this study are a public dataset; the source is detailed in the text.

Conflicts of Interest

The authors declare that they have no conflicts of interests.

Authors' Contributions

Qing Huang conceptualized the study, developed methodology, developed software, and wrote the original draft. Heru Xue supervised the study and reviewed and edited the manuscript. Jiangping Liu and Xinhua Jiang reviewed and edited the manuscript.

Acknowledgments

The authors gratefully acknowledge the support of the National Natural Science Foundation of China (61461041 and 31960494), Inner Mongolia Autonomous Region Major Science and Technology Project (2021ZD0003), and Natural Science Foundation of Inner Mongolia Autonomous Region (2022MS06026).

References

[1] W. H. Maes and K. Steppe, "Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture," *Trends in Plant Science*, vol. 24, no. 2, pp. 152–164, 2019.

[2] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sensing*, vol. 12, no. 16, p. 2659, 2020.

[3] B. Li, R. Fu, H. Tan et al., "Characteristics of the interaction mechanisms of procyanidin B1 and procyanidin B2 with protein tyrosine phosphatase-1B: analysis by kinetics, spectroscopy methods and molecular docking," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 259, Article ID 119910, 2021.

[4] C. C. Huang, "Applications of raman spectroscopy in herbal medicine," *Applied Spectroscopy Reviews*, vol. 51, no. 1, pp. 1–11, 2016.

[5] P. J. Dev, A. Sukenik, D. R. Mishra, and I. Ostrovsky, "Cyanobacterial pigment concentrations in inland waters: novel semi-analytical algorithms for multi- and hyperspectral remote sensing data," *Science of the Total Environment*, vol. 805, 2022.

[6] S. Chen, Y. Gao, K. Fan et al., "Prediction of drought-induced components and evaluation of drought damage of tea plants based on hyperspectral imaging," *Frontiers of Plant Science*, vol. 12, Article ID 695102, 2021.

[7] G. Ren, J. Ning, and Z. Zhang, "Multi-variable selection strategy based on near-infrared spectra for the rapid description of dianhong black tea quality," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 245, Article ID 118918, 2021.

[8] J. Xinhua, X. Heru, Z. Lina, G. Xiaojing, W. Guodong, and B. Jie, "Nondestructive detection of chilled mutton freshness based on multi-label information fusion and adaptive BP neural network," *Computers and Electronics in Agriculture*, vol. 155, pp. 371–377, 2018.

[9] M. Li, Y. Feng, Y. Yu et al., "Quantitative analysis of polycyclic aromatic hydrocarbons in soil by infrared spectroscopy combined with hybrid variable selection strategy and partial least squares," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 257, Article ID 119771, 2021.

[10] M. Martínez Galera, D. Picón Zamora, J. L. Martínez Vidal, and A. Garrido Frenich, "Selection of the simpler calibration model for multivariate analysis by partial least squares," *Analytical Letters*, vol. 35, no. 5, pp. 921–941, 2002.

[11] M. Zhang, S. Zhang, and J. Iqbal, "Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares," *Chemometrics and Intelligent Laboratory Systems*, vol. 128, pp. 17–24, 2013.

[12] P. Mishra and E. J. Woltering, "Identifying key wavenumbers that improve prediction of amylose in rice samples utilizing advanced wavenumber selection techniques," *Talanta*, vol. 224, Article ID 121908, 2021.

[13] R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data," *Analytica Chimica Acta*, vol. 692, pp. 63–72, 2011.

[14] J. Geng, C. Yang, Q. Luo, L. Lan, and Y. Li, "iPCPA: interval permutation combination population analysis for spectral wavelength selection," *Analytica Chimica Acta*, vol. 1171, Article ID 338635, 2021.

[15] R. K. H. Galvao, M. C. U. Araújo, W. D. Fragoso et al., "A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm," *Chemometrics and Intelligent Laboratory Systems*, vol. 92, no. 1, pp. 83–91, 2008.

[16] H.-Y. Zhen, R.-J. Ma, Y. Chen, X. P. Sun, and C. L. Ma, "Study on prediction model of malathion pesticide concentration absorption spectra based on CARS and K-S," *Spectroscopy and Spectral Analysis*, vol. 40, no. 5, pp. 1601–1606, 2020.

[17] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Analytica Chimica Acta*, vol. 648, no. 1, pp. 77–84, 2009.

[18] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, and M. Akhond, "Ant colony optimisation: a powerful tool for wavelength selection," *Journal of Chemometrics*, vol. 20, pp. 146–157, 2006.

[19] T. Liu, T. Xu, F. Yu, Q. Yuan, Z. Guo, and B. Xu, "A method combining ELM and PLSR (ELM-P) for estimating chlorophyll content in rice with feature bands extracted by an improved ant colony optimization algorithm," *Computers and Electronics in Agriculture*, vol. 186, Article ID 106177, 2021.

[20] S. Cateni, V. Colla, and M. Vannucci, "General purpose input variables extraction: a genetic algorithm based procedure give A gap," in *Proceedings of the International Conference on Intelligent Systems Design and Applications*, IEEE, Pisa, Italy, December 2009.

[21] B. K. Lavine and C. G. White, "Boosting the Performance of Genetic Algorithms for Variable Selection in Partial Least Squares Spectral Calibrations," *Applied Spectroscopy*, vol. 71, 2017.

[22] A. G. Gad, K. M. Sallam, R. K. Chakrabortty, M. J. Ryan, and A. A. Abohany, "An improved binary sparrow search algorithm for feature selection in data classification," *Neural Computing & Applications*, vol. 34, no. 18, pp. 15705–15752, 2022.

[23] F. Allegrini and A. C. Olivieri, "A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis," *Analytica Chimica Acta*, vol. 699, no. 1, pp. 18–25, 2011.

[24] J. H. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression

with applications to mid-infrared and near-infrared spectroscopic data," *Analytical Chemistry*, vol. 74, no. 14, pp. 3555–3565, 2002.

[25] R. Leardi and L. Norgaard, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions," *Journal of Chemometrics*, vol. 18, no. 11, pp. 486–497, 2004.

[26] Y.-H. Yun, H.-D. Li, L. R. E Wood et al., "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 111, pp. 31–36, 2013.

[27] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," *Advances in Intelligent Systems and Computing*, Addis Ababa University, Addis Ababa, Ethiopia, pp. 1–13, 2015.

[28] L. Sun, S. Si, J. Zhao, J. Xu, Y. Lin, and Z. Lv, "Feature selection using binary monarch butterfly optimization," *Applied Intelligence*, vol. 208, 2022.

[29] H. Jia, W. Zhang, R. Zheng, S. Wang, X. Leng, and N. Cao, "Ensemble mutation slime mould algorithm with restart mechanism for feature selection," *International Journal of Intelligent Systems*, vol. 37, no. 3, pp. 2335–2370, 2021.

[30] B. J. Ma, S. Liu, and A. A. Heidari, "Multi-strategy Ensemble Binary Hunger Games Search for Feature Selection," *Knowledge-Based Systems*, vol. 248, 2022.

[31] Y. Zhang, R. Liu, X. Wang, H. Chen, and C. Li, "Boosted binary harris hawks optimizer and feature selection," *Engineering with Computers*, vol. 37, 2020.

[32] T. Bahaeddin, U. Sait Ali, and K. Ersin, "Binary artificial algae algorithm for feature selection," *Applied Soft Computing*, vol. 120, 2022.

[33] L. Tong, "A new method and application of molecular spectrum wavelength selection based on ant colony algorithm," Master Thesis, Zhejiang University, Hangzhou, China, 2017.

[34] Z. Xiaoming, M. Zhikang, and L. Shaowen, "Application of ant colony algorithm in wavelength selection of soil available phosphorus near infrared spectrum%," *Jiangsu Agricultural Science*, vol. 47, no. 19, pp. 227–231, 2019.

[35] J. Xu, H. Zhang, D. Yang, J. Zhang, J. Qian, and L. Liu, "The determination of a diesel solidifying point by near infrared spectroscopy," *Petroleum Science and Technology*, vol. 31, no. 19, pp. 1974–1979, 2013.

[36] J. Wang, M. Shi, P. Zheng, and S. Xue, "Quantitative analysis of lead in tea samples by laser-induced breakdown spectroscopy," *Journal of Applied Spectroscopy*, vol. 84, no. 1, pp. 188–193, 2017.

[37] H. Jin and P. Luo, "Study on the accuracy of photoacoustic spectroscopy system based on multiple linear regression correction algorithm," *AIP Advances*, vol. 11, no. 9, Article ID 095314, 2021.

[38] W. Jun, "Feasibility study on discrimination of adulterated milk based on optical parameters," Master Thesis, Tianjin University, Tianjin, China, 2017.

[39] D. Li, J. Svensson, H. Thomsen, F. Medina, A. Werner, and R. Wolf, "Bayesian soft X-ray tomography using non-stationary Gaussian Processes," *Review of Scientific Instruments*, vol. 84, no. 8, Article ID 083506, 2013.