OE Optical Engineering

RESEARCH PAPER

Non-destructive detection of specific enzyme activities in tomato plants using visible-near infrared spectroscopy

Ao Li[a,b] and Jin-Shi Cui[a,*]

[a]Changchun University of Science and Technology, College of Optoelectronic Engineering, Changchun, China
[b]Changchun University of Science and Technology, CUST-ITMO Joint Institute of Optics and Fine Mechanics, Changchun, China

ABSTRACT. Tomato is an important vegetable that is rich in antioxidants, vitamins, and minerals and has significant economic and health value. In this study, hyperspectral images in the wavelength range of 370 to 1715 nm were first preprocessed to improve the data quality and comparability. Subsequently, the tomatoes were chemically destroyed, and the average activities of peroxidase enzyme, phenylalanine ammonia-lyase enzyme, and α-amylase enzyme were 5.108 mU/g, 7.347 U/g, and 35.856 U/g, respectively. Then, two spectral selection algorithms, the genetic algorithm (GA) and the successive projection algorithm, were used to extract effective wavelength bands from high dimensional spectral data. And the extracted effective wavelength variables were combined with partial least squares (PLS) regression to build the optimal spectral selection model GA-PLS. Finally, three additional spectral prediction models were created by combining the GA-selected spectra with three other algorithms: support vector machine, particle swarm optimization-backpropagation neural network, and random forest. After comparing the predictive performance of four models, it was found that the GA-PLS model had the highest prediction accuracy and stability. Furthermore, compared with tomato stems, the near infrared (NIR) bands of tomato leaves were more accurate in predicting the enzyme content of tomato plants. It was found that the GA-PLS model had a better prediction performance for the three enzymes in the NIR band of leaves with $\overline{R}_p$ (average coefficient of determination for the three enzymes) and $\overline{\text{RMSEP}}$ (average root means square error of the three enzymes) of 0.815 and 1.659, respectively. This provides an effective method for phytochemical composition analysis using hyperspectral imaging.

Keywords: feature selection; genetic algorithm; successive projection algorithm; hyperspectral images; hyperspectral prediction model

Paper 20240169G received Feb. 17, 2024; revised Jun. 5, 2024; accepted Jun. 5, 2024; published Jul. 2, 2024.

1 Introduction

The tomato (*Solanum Lycopersicon L.*) is an important crop widely used for dietary supplementation. The quality and yield of the tomato plant is influenced by a number of factors, including the content of specific enzymes in the plants. The content of some specific plant enzymes reflects the physiological and metabolic status of the plant. It is closely related to plant growth and development, stress tolerance, disease resistance, and antioxidant ability.[1] Therefore, accurate, rapid, and non-destructive determination of tomato plant enzyme content is very important for

*Address all correspondence to Jin-Shi Cui, cuijinshi0804@hotmail.com

assessing the quality and yield of the tomato plant. The three enzymes that are most responsive to tomato growth are α-amylase (α-AMS), peroxidase (POD), and phenylalanine ammonia-lyase (PAL).[2,3] α-AMS is a key enzyme involved in starch metabolism. It breaks down starch into soluble sugars, which affects the sweetness and taste of tomato fruits.[4] POD is an important enzyme involved in the antioxidant defense system of plants. This enzyme catalyzes the breakdown of hydrogen peroxide into water and oxygen, thereby eliminating reactive oxygen radicals and protecting cell membranes and organelles from oxidative damage.[5] PAL is a key enzyme involved in the phenylalanine metabolic pathway. This enzyme catalyzes the conversion of phenylalanine to cinnamic acid, which promotes the synthesis of secondary metabolites, such as anthocyanins.[6]

Currently, the mainstream enzyme activity assays include spectrophotometry, chemical assays, enzyme kinetics, immunological methods, colourimetric methods, and reflectance spectrometry, etc. They all have some advantages and disadvantages. Spectrophotometry determines the enzyme activity by measuring the change in absorbance of light at a specific wavelength. The advantage is that it is suitable for the determination of the reaction involving coenzyme NAD+ or NADP+, with higher sensitivity, and the disadvantages are that it is only suitable for dilute solutions with absorbance values between 0.05 and 1.0 and is sensitive to temperature changes. Chemical assays usually refer to the determination of enzyme activity using the color change produced by a chemical reaction. The advantages are that it is highly accurate and provides quantitative information, and the disadvantage is that the steps may be more complicated as it requires specific reagents and equipment. Enzyme kinetic methods determine enzyme activity by measuring the rate of an enzymatic reaction. The advantage is that it provides detailed information about the kinetics of the enzymatic reaction. The disadvantage is that the experimental conditions are stringent, e.g., temperature and pH control. Immunological methods determine the enzyme activity using the principle of specific binding of antibodies to the enzyme. The advantages are that it is specific and can be used for the detection of enzyme activity in complex samples, and the disadvantages are that it requires specific antibodies and has a high cost. The colorimetric method quantifies enzyme activity by color change. The advantages are its simple operation and ease of observing, and the disadvantages are that the color of the sample may interfere with it and its accuracy is limited. Reflectance spectroscopy estimate enzyme activity by determining the reflectance characteristics of the sample to light. The advantages are that it is non-invasive, fast, and convenient, and the disadvantage is that it may not be as accurate as the chemical detection method. Currently, reflectance spectroscopy is mostly used in non-destructive testing, but the innovation of this paper is that the spectral range of the analysis is from 370 to 1700 nm, which avoids errors caused by insufficient analytical spectral data. This paper also uses a spectral selection algorithm to extract the characteristic bands for the spectral analysis, which improves the accuracy and efficiency of the analysis. The object of the study is the target enzyme rather than the product of the enzyme activity or a specific ingredient related to the quality of the tomato. In this paper, we also use several prediction algorithms for comparative analysis to get the best model for tomato enzyme activity prediction.

Enzyme activity, also known as enzyme viability, is the ability of an enzyme to catalyze a specific chemical reaction. The indirect inference of changes in enzyme activity within tomatoes through reflectance spectral analysis is based primarily on the relationship between spectral properties and changes in the chemical composition within tomatoes. The key point here is that enzymes, as catalysts, affect the biochemical reactions within tomatoes, which in turn changes the content and ratio of its chemical constituents. These chemical components include sugars, organic acids, pigments, etc., which have specific spectral properties for light absorption and reflection.

When enzyme activity in tomatoes changes, such as during ripening, it affects the concentration of these chemicals. Because these chemicals have different absorption and reflectance of light at specific wavelengths, changes in these chemicals can be inferred indirectly by measuring the spectra reflected from the tomato surface. From these variations, changes in enzyme activity can thus be inferred.

This paper focuses on three plant enzymes related to tomato quality and yield: α-AMS, POD, and PAL. It establishes a nondestructive method for the determination of the three enzyme activities. The reflectance spectra of leaves and stems of tomatoes is analyzed, and the

relationship between the reflectance spectra of different enzyme activities and the enzyme content is used to establish a model to predict the enzyme activities of tomatoes.[7] Finally, the partial least squares (PLS) algorithm was mainly utilized to predict the enzymatic activity of tomatoes.

To control the accuracy of the spectral assay, the chemical assay was used as the gold standard. The activities of these three enzymes are determined by double antibody sandwich assay, i.e., using a specific antibody-coated microtiter plate, which binds to the enzyme in the specimen and horseradish peroxidase (HRP)-labeled antibody to the enzyme, forming an antibody-antigen-enzyme-labelled antibody complex. Then the TMB (3,3',5,5'-Tetramethylbenzidine) chromogen, which produces a yellow product catalyzed by HRP enzyme, is added, and the shade of its color is positively correlated with the enzyme level in the specimen. Absorbance (optical density) was measured at 450 nm using a spectrophotometer, and the concentration of enzyme activity in the samples was calculated from a standard curve according to the Beer–Lambert law. The activities of all three enzymes are based on the production of detectable products from the specific reactions that they catalyze. Plant POD catalyzes a reaction in which hydrogen peroxide oxidizes phenolics to produce quinones. These compounds are further condensed or condensed with other molecules to produce darker colored compounds. In this reagent kit, guaiacol (o-methoxyphenol) was used as a phenolic substance, which is oxidized by hydrogen peroxide to reddish brown 4-o-methoxyphenol catalyzed by POD. This substance has maximum light absorption at 470 nm; plant α-AMS catalyzes the reaction of the starch decomposition to soluble sugars, producing reducing sugars.. These reducing sugars can undergo a redox reaction with the color rendering agent TMB, which produces a blue product catalyzed by the enzyme HRP, which is converted to a yellow product under acidic conditions and which has maximum light absorption at 450 nm. PAL catalyzes the decarboxylation of phenylalanine to trans-cinnamic acid. In turn, trans-cinnamic acid can undergo a redox reaction with the color developer TMB to produce a blue product catalyzed by the HRP enzyme, which is converted to a yellow product under acidic conditions with a maximum light absorption wavelength of 450 nm.[8,9]

Spectral analysis technology has received wide attention and application as a non-destructive testing technology. Using visible (VIS)-near infrared (NIR) reflectance spectroscopy, a method for detecting enzyme activity in tomatoes was proposed; it improved the accuracy and stability of the model by preprocessing the spectral data and selecting variables, and provided a new method for the rapid detection of enzyme activity in tomatoes.[10] They can be categorized into NIR spectra, mid-infrared spectra, VIS spectra, ultraviolet spectra, etc., according to the wavelength range of electromagnetic radiation; they can also be categorized into hyperspectral, multispectral, monospectral, etc., according to the magnitude of spatial resolving power;[11] in fact, by this technique, there are Raman, ellipsometric, Brillouin, Fourier, and so on. In agricultural non-destructive testing, spectral analysis technology can be used to quickly, accurately, and non-destructively determine the quality, composition, nutrition, safety, and other indicators of agricultural products, providing a scientific basis for agricultural production and processing.[12] Using visible (VIS)-NIR reflectance spectroscopy, a non-destructive testing model for tomato hardness was established, and the noise and redundancy of the data were reduced by wavelet transform and principal component analysis of spectral data, which provided a basis for tomato quality assessment.[13] In this paper, more enzymes are detected in a non-destructive manner to control the yield and quality of tomato plants.

In terms of spectral selection, many spectral selection methods have been used for the analysis of plant enzyme activities, such as the successive projection algorithm (SPA), the uninformative variable removal algorithm, the genetic algorithm (GA), and the interval PLS (iPLS). Among them, the most used and the most effective ones are the GA and SPA algorithms. In this paper, by comparative analysis, GA is better than SPA in the spectral selection of tomato reflectance spectra, and the selected bands are from 1460 to 1574 nm. In addition, some machine learning based wavelength selection algorithms, such as support vector machine (SVM), random forest (RF), and particle swarm optimization-backpropagation (PSO-BP) are also used to compare and analyze the optimal prediction algorithms.[14] GAs can be used for spectral wavelength selection, i.e., screening the wavelengths with the highest correlation with the target variables from high dimensional spectral data, thus reducing the complexity and noise of the data and improving the accuracy and efficiency of modeling. The advantage of GAs is that it can handle nonlinear, multi-peak, discrete optimization problems with strong global search capability and

robustness. PLS can be used for regression analysis of spectral data, i.e., predicting the value of a target variable based on spectral information, e.g., the activity of a specific enzyme in tomato plants.[15]

In recent years, various technologies, such as NIR spectroscopy, VIS/NIR spectroscopy, hyperspectral imaging technology, and Raman spectroscopy, have been employed in agriculture to analyze the water content, proteins, fats, starches, fibers, and other components of agricultural products. These technologies are also used for the classification of varieties, the detection of fungal infections and pesticide residues, and other quality indicators. To enhance the accuracy and stability of detecting quality indicators, such as color, sugar, and acidity, in agricultural products, the comprehensive detection of both internal quality and external characteristics is necessary. This will enable qualitative and quantitative analysis of agricultural products, which has numerous applications.[16,17]

Therefore, the aim of this paper is to explore the effect of changes in enzyme activity within tomato plants on spectra to predict enzyme activity. This experiment requires a large amount of accurate data support in the preliminary data collection for modeling; however, the portable spectrometer is not enough to perfectly collect the spectral data that we need. Therefore, the spectral data acquisition requires digging out the tomato plant, but the tomato will have changes in its enzyme activity after leaving the original growth environment. Thus, in this paper, the tomato plant is placed in a vacuum bag for liquid nitrogen freezing to ensure that the enzyme activity of the tomato body and the enzyme activity of the tomato grown in the land is the same. Digging out the plant and liquid nitrogen freezing are only used in the pre-spectral data acquisition, so for the proposed model, a non-destructive testing method is used.

2 Materials and Methods

2.1 Hyperspectral Imaging System

The hyperspectral imaging system used in the proposed method consists of four main components: (1) a bottom horizontal support plate, linked to a stepper motor (Leashine, 20CM003, 100 r/min), which can be moved horizontally; (2) a camera mount for fixing and changing the vertical position of the camera (the vertical distance between the camera and the platform is 50 cm); (3) hyperspectral cameras (Specim FX10-CHINA, Specim FX17-CHINA); and (4) a dual illumination system. The hyper-spectral imaging system is shown in Fig. 1.

The obtained hyperspectral images comprise hundreds of sequential bands of tomato leaves and stems. White and black references were used to correct the hyperspectral images.[18] R represents a ratio comparing the images with a white background; 100 represents having the same

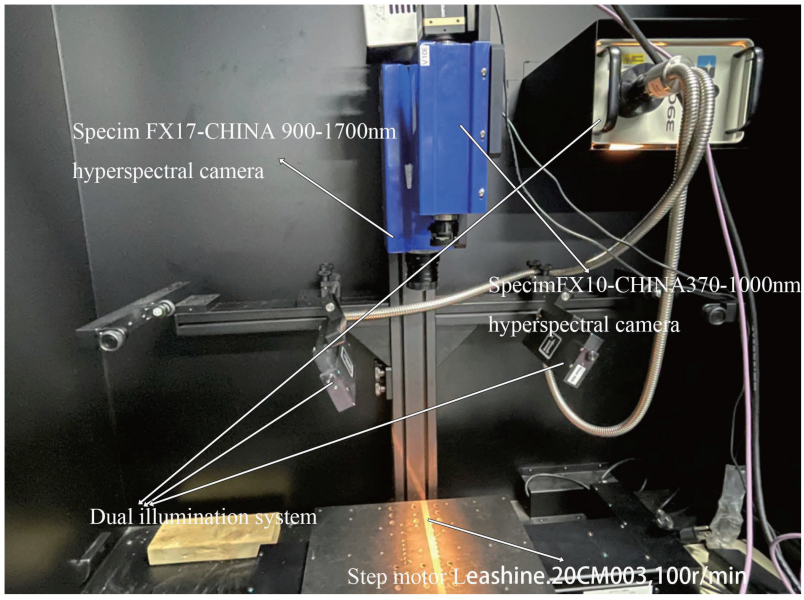

Fig. 1 Schematic diagram of the hyperspectral imaging system.

reflectivity as a white background, and 0 represents having the same reflectivity as a black background. The value is calculated as:

$$R = \frac{i_{\text{raw}} - i_{\text{dark}}}{i_{\text{white}} - i_{\text{dark}}},$$

where i_{raw} represents the reflected light intensity of the hyperspectral tomato image, i_{dark} represents the reflected light intensity on the standard black background, and i_{white} represents the reflected light intensity on the standard white background.

2.2 Spectrophotometer

The absorbance at 450 nm is measured using a spectrophotometer (UV3600_UV-Vis-NIR dual-beam spectrophotometer-OPTOSKY-CHINA).

Reflectance spectral data of tomato plants were collected using a hyperspectral system with a wavelength range of 370 to 1715 nm and a spectral resolution of 1 nm. Tomato leaf and stem samples were collected at different growth stages (1 to 8 weeks) and under the same treatments. The contents of α-AMS, POD, and PAL were determined by traditional chemical methods.[19] Hyperspectral-based prediction models for α-AMS, POD, and PAL content in the tomato plants were then developed using various modeling techniques. The study compared and analyzed the prediction accuracies to obtain relatively good detection models. Figure 2 shows the flowchart of this study.

2.3 Gold-Standard Detection of Enzyme Concentration

In this paper, the chemically detected enzyme activity was used as the gold standard for comparative analysis.

The three enzyme activities were calculated by determining the enzyme concentration in the sample through a linear regression equation and then multiplying by the dilution factor to obtain the actual concentration of the sample. The specific equations are as follows:

1. POD active unit:

$$(U) = \Delta OD470 \text{ min}^{-1} \text{ g}^{-1} \text{ FW},$$

2. α-AMS active unit:

$$(U) = \Delta OD450 \text{ min}^{-1} \text{ g}^{-1} \text{ FW},$$

3. PAL active unit:

$$(U) = \Delta OD450 \text{ min}^{-1} \text{ g}^{-1} \text{ FW}.$$

All three equations indicate that the change in absorbance per minute per gram of fresh weight (FW) sample at either 470 or 450 nm is its POD activity unit (U). Statistical analyses of the chemical assays for the three enzyme activities are presented in Table 1.

2.4 Spectral Selection

Spectral selection is a technique used to extract effective wavelength variables from high-dimensional spectral data. This method reduces redundancy and noise in the data, leading to improved accuracy and stability of the model.[20] This study employed two spectral selection algorithms, namely GA and SPA, which were combined with PLS. A spectral prediction model was developed for predicting enzyme activities in tomato plants.

GA and SPA are two very effective methods in the field of spectral selection. The advantage of GA lies in its global search capability, which can simulate the natural selection and genetic mechanism in the process of biological evolution and find the optimal wavelength combinations through crossover, mutation, and selection operations of populations. This method not only is capable of handling large-scale datasets but also can effectively avoid falling into local optimal solutions, thus improving the generalization ability of the model.[14]

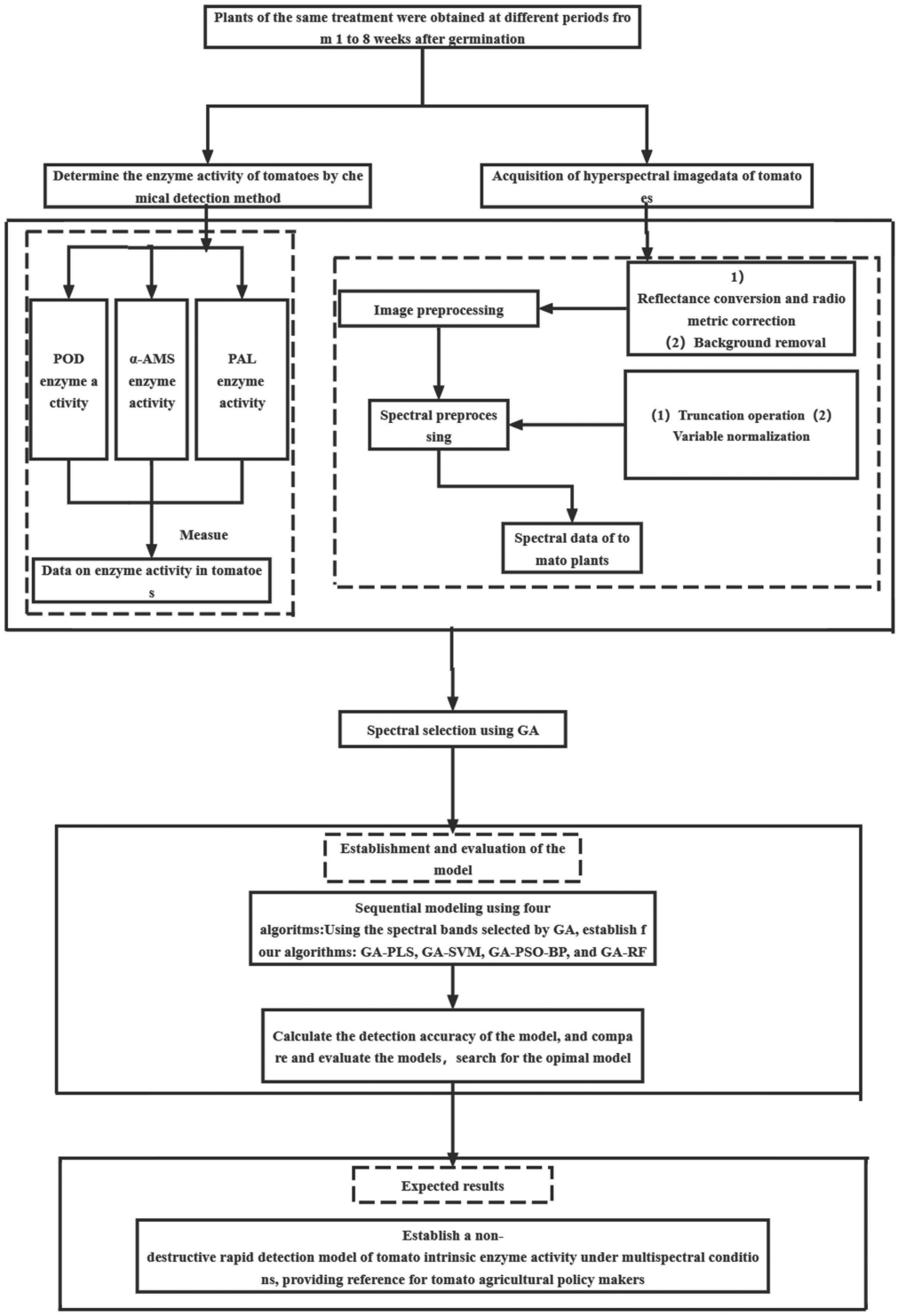

Fig. 2 Research flowchart.

On the other hand, SPA, as a forward variable selection method, selects wavelengths by minimizing the covariance between variables. It is computationally efficient, easy to implement, and particularly suitable for dealing with spectral data with a high degree of covariance. SPA selects the variables step by step by successive projections, which reduces the complexity of the model while maintaining its stability and prediction accuracy.[21]

Spectral selection allows the analyst to select specific spectral regions to observe, allowing for more accurate identification and quantification of components in a sample. It allows for

Table 1 Descriptive statistical analysis of enzyme concentration detected by chemistry.

	POD mU/g	PAL U/g	AMS U/g
Average	5.108	7.347	35.845
Standard error	0.073	0.060	0.440
Standard deviation	0.514	0.721	5.215
Variance	0.265	0.520	27.20
Minimum value	3.81	6.12	25.8
Maximum value	6.21	9.01	44.37

increasing the sensitivity; by selecting specific spectral lines, the sensitivity to detect specific elements or compounds can be increased. It reduces interference; good selectivity enables the determination of elements and compounds with similar chemical properties, reducing spectral interference from other elements. Finally, it improves the analysis speed; spectral selection can quickly locate the spectral region of interest, speeding up the analysis process.

The following equation was used for modeling using a chemical detection algorithm as the gold standard assay to predict enzyme activity in tomatoes:

$$y = X \bullet b + e,$$

where y is the target variable, i.e., the predicted value of enzyme activity. X is the independent variable, i.e., the data obtained by chemical assay. b is the regression coefficient calculated by the PLS algorithm. e is the error term, which represents the difference between the predicted and actual values.

GA is a global optimization algorithm that simulates the processes of natural selection, crossover, and mutation to gradually improve the fitness of the population. A GA was used to select the best subset of wavelengths as input variables for the PLS regression. Each wavelength variable is regarded as a gene, each wavelength subset is regarded as a chromosome, and the prediction error is taken as the fitness function.[22]

SPA is a spectral selection algorithm based on orthogonal projection. It constructs a subset of wavelengths that is minimally redundant and maximally correlated with the target variable by progressively selecting the most correlated wavelength variable.[23]

PLS is a multivariate statistical data analysis method that enables regression modeling (multivariate linear regression), data structure simplification (principal component analysis), and correlation analysis (typical correlation analysis) between two sets of variables to be performed simultaneously under a single algorithm. The PLS method is used to obtain mutually orthogonal eigenvectors for the independent variables and the dependent variable, respectively, by projecting their high-dimensional data spaces into the corresponding low-dimensional spaces. The mutually orthogonal eigenvectors of the independent variable and the dependent variable are obtained, and then the one-way linear regression relationship between the eigenvectors of the independent variable and the dependent variable is established. Not only can it overcome the problem of covariance, but it also emphasizes the explanatory and predictive roles of the independent variables on the dependent variable when selecting the feature vectors and removes the influence of unhelpful noise on the regression, resulting in the model containing a minimum number of variables.

SVM is a supervised learning model mainly used for classification and regression analysis. The basic idea is to maximize the spacing between different classes of data points by finding an optimal hyperplane (or hypersurface). This optimal hyperplane, called the maximum spacing hyperplane, maximizes the separation of data points of different categories and thus achieves good classification results.[24]

The PSO-BP neural network is an algorithm that combines PSO and the BP neural network and is mainly used for nonlinear function fitting and regression problems. The basic idea is to use the PSO algorithm to optimize the weights and thresholds of the BP neural network to improve the performance and convergence speed of the BP neural network. PSO is a heuristic algorithm

that simulates the foraging behavior of a flock of birds for searching, and it is able to achieve the search of the global optimal solution.[25]

RF is an integrated learning method that is defined as a linear classifier with maximum intervals on the feature space. RF constructs multiple decision trees during training and outputs classes as patterns of classes (classification) or average predictions (regression). The learning strategy of RF is interval maximization, which can be formalized as solving a convex quadratic programming, which is also equivalent to minimizing a regularized hinge loss function.[26]

The average coefficient of determination for the three enzymes and the average root mean square error of the three enzymes of the two spectral selection algorithms mentioned above are used to measure the degree of model fit to the data and the prediction error, respectively. The larger the average coefficient of determination is and the smaller the average root mean square error is, the better the predictive performance of the model is. These two metrics are used to compare the advantages and disadvantages of the GA and SPA algorithms and choose the better algorithm. Then, the better algorithm in combination with PLS, SVM, PSO-BP, and RF is used to observe the prediction results.

The establishment of predictive models is a key step in spectral analysis, which allows for the use of spectral data to predict the nature or composition of a target substance. Using spectra selected by GA as inputs, four predictive models were built: GA-PLS, GA-SVM, GA-PSO-BP, and GA-RF; they are based on PLS regression, SVM, particle swarm optimization-backpropagation neural network (PSO-BP), and RF algorithms, respectively.

3 Results

3.1 Extraction of Hyperspectral Image Data From Leaves and Stems of Tomato Samples

This paper focuses on the spectral characteristics of tomato plants, specifically their leaves and stems. To observe these characteristics, a hyperspectral imaging system was used to photograph the plants, resulting in a clear image (Fig. 3). The image allows for the distinction of different colors and shapes of the tomato leaves and stems. To analyze the spectral information, the image was processed using ENVI software. This produced a spectral plot of the average of the spectra of each point of all leaves in different wavelength ranges. Figures 4 and 5 show the results, with the horizontal axis representing the wavelength and the vertical axis representing the intensity of light received by the detector. Upon comparison of these graphs, it is evident that they share similar trends in their spectral curves, which are linked to their physiological and chemical properties.

(a) (b)

Fig. 3 Photos of tomato plants in the NIR and VIS bands: (a) photos of tomato plants in the NIR band and (b) photos of tomato plants in the VIS light band.

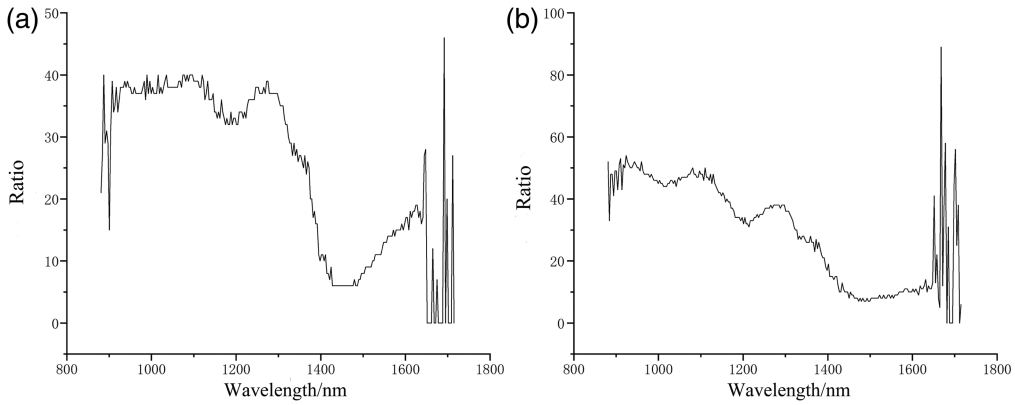

Fig. 4 Reflectance of tomato stems and leaves in the NIR spectrum: (a) reflectance spectra of tomato stems in the NIR band and (b) reflectance spectra of tomato leaves in the NIR band.

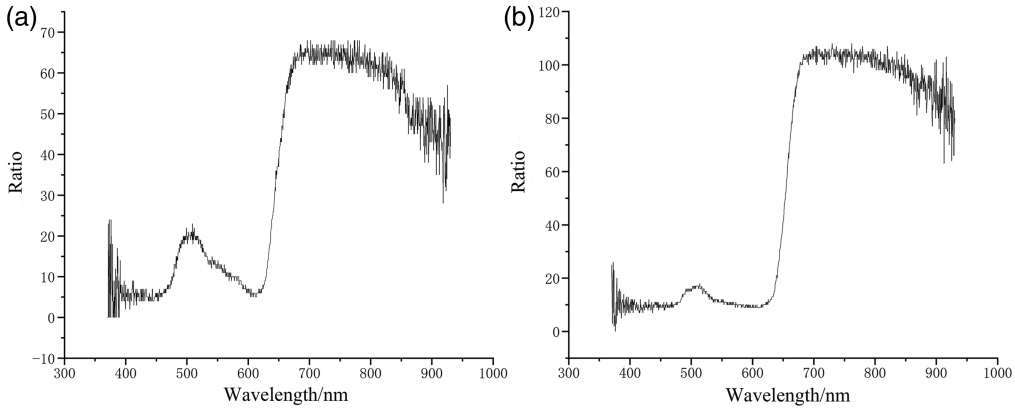

Fig. 5 Reflectance of tomato stems and leaves in the VIS spectrum: (a) reflectance spectra of tomato stems in the VIS band and (b) reflectance spectra of tomato stems in the NIR band.

However, it is evident in Figs. 4 and 5 that there is also some noise in the raw image and spectral data, which may affect the analysis results. Therefore, some preprocessing of the data is needed to remove or reduce the effect of noise before proceeding to subsequent studies. In the lower spectral range of 370 to 430 nm and in the higher spectral ranges of 880 to 940 nm and 1640 to 1715 nm, the noise is high and shows high-frequency features. These noises may come from the instrument, the environment, or the sample itself. To prevent these noises from interfering with the study, it was decided to exclude these bands in the preprocessing stage and keep only the low spectral range 430 to 880 nm and the high spectral range 940 to 1640 nm.

3.2 Preprocessing of Data

Data preprocessing is an important step in spectral analysis as it eliminates noise, baseline drift, scattering effects, and other interfering factors in the spectral data. This improves the quality and comparability of spectral signals, enhancing the correlation between the spectra and the target substances. This, in turn, provides more accurate and effective data for subsequent model building and analysis.[27] Three data preprocessing methods were used: variable normalization, truncation operations, and removal of outliers.

Variable standardization is a frequently used data preprocessing technique. It removes differences in magnitude and scale between variables, resulting in each variable having a mean of 0 and a standard deviation of 1.[28] This approach prevents the model from being influenced by variables that are too large or too small and enhances the stability and generalization of the model. The reflectance at each wavelength was normalized based on the variables to enable comparison at the same scale. The truncation operation selects a suitable wavelength range by removing irrelevant or redundant information, reducing data dimension and computational

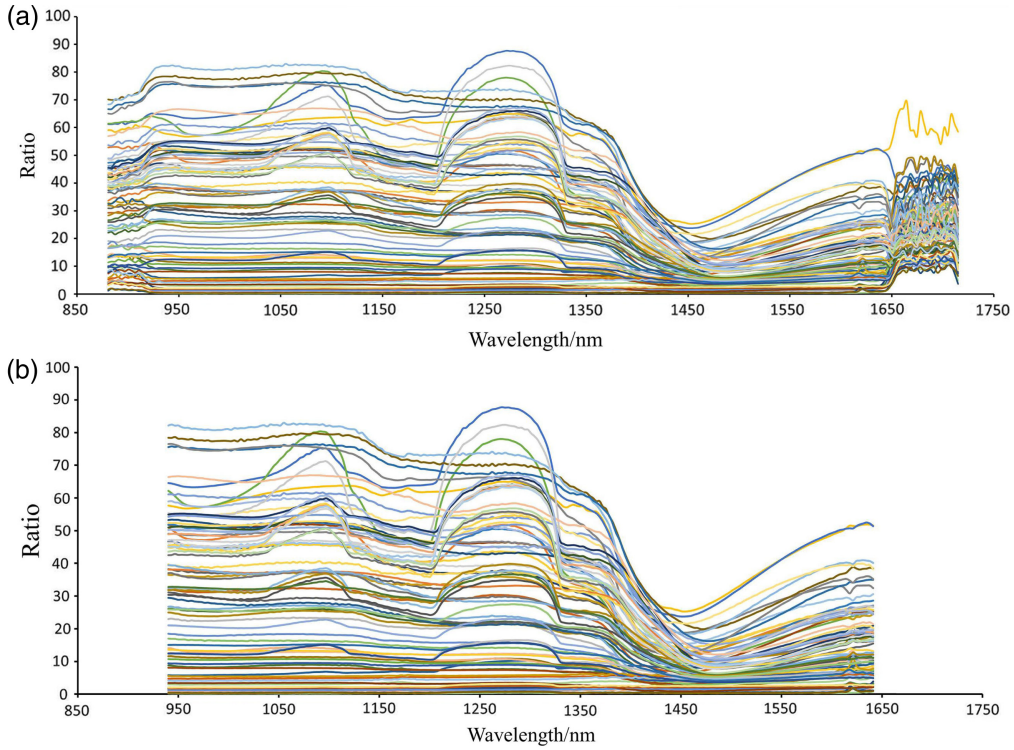

Fig. 6 Spectral information of unpretreated and pretreated tomato stems: (a) spectral information of unpretreated tomato stems in the NIR light band and (b) spectral information of pretreated tomato stems in the NIR band.

complexity. The above two data preprocessing methods were used to obtain clearer and more concise spectral data, laying the foundation for subsequent spectral analyses.

These remove noise and interference and improve the quality and credibility of the data.[29] After preprocessing, two spectral graphs are obtained, as shown in Fig. 6; these represent the reflectance of tomato leaves at different wavelengths.

3.3 Spectral Band Selection

Parameters such as population size, crossover probability, and mutation probability were set, and several iterations were run to finally obtain an optimal solution for a leaf and stem variable containing 31 and 26 wavelengths in the VIS range and 35 and 28 wavelengths in the NIR band range.

The SPA was used to select the best subset of wavelengths to be used as input variables for the PLS regression. The number of wavelengths to be selected was set, and the SPA algorithm was run; the result was an optimal solution containing 27 and 20 wavelength variables in the VIS range and 25 and 19 wavelength variables in the NIR range for both leaf and stem.

The relationship between the variable importance and the variable index was obtained as follows.

The importance of each band of the leaf of tomato plants in the NIR band was obtained by the GA model, as shown in Fig. 7, and the screened spectral bands were determined based on the order of importance as well as the cumulative contribution, which was taken as the 179th (1460 nm)-214th (1574 nm), totaling 35 bands. Similarly, the bands screened under the other seven models can be obtained. The calibration and validation sets are first assigned based on the available data.

Significance refers to the degree of contribution of each NIR band to the enzyme activity in the GA model. To reduce the dimensionality of the data analysis and to improve the analytical relevance of the characteristic spectra, the variables with more than 80% importance were taken for the analysis in this paper, i.e., the bands from 179 to 214.

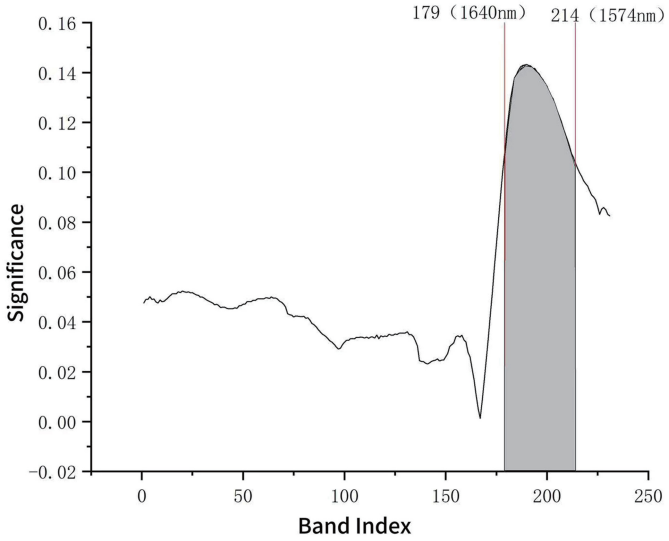

Fig. 7 Significance of each band of leaves in the NIR range under the GA model.

3.4 Allocation of Calibration and Validation Sets

Allocation was done by the random division method, i.e., the dataset was randomly divided into the test set and validation set in the ratio of 11:3, with the validation set being used to verify the accuracy of the constructed model. All 100 tomato plants used in this experiment were examined by spectral analysis and then sent for chemical examination. 21 were used as the validation set and the remaining 79 were used as the test set.

This paper combines GA and SPA with the PLS, SVM, PSO-BP, and RF algorithms and conducts a comparative analysis on different parts of tomatoes (leaves and stems) and different spectral bands (VIS and NIR bands), respectively. Table 2 shows an example with GA-PLS and SPA-PLS. Here, LEAF-NIR represents the average determination coefficient and average root mean square error of the analysis of leaves under NIR, LEAF-VIS represents the average determination coefficient and average root mean square error of the analysis of leaves under VIS light, STEM-NIR represents the average determination coefficient and average root mean square error of the analysis of stems under NIR, and STEM-VIS represents the average determination coefficient and average root mean square error of the analysis of stems under VIS light.

To compare the effectiveness of the two spectral selection algorithms, GA and SPA, two spectral prediction models were developed by combining them with PLS regression, respectively,

Table 2 Comparison of the accuracy of the GA and SPA spectral selection methods.

	Test set		Validation set	
Prediction model	$\overline{R_p}$	$\overline{\text{RMSE}}$	$\overline{R_p}$	$\overline{\text{RMSEP}}$
GA-PLS-LEAF-NIR	0.834	1.564	0.815	1.659
SPA-PLS-LEAF-NIR	0.793	1.674	0.755	1.946
GA-PLS-STEM-NIR	0.787	1.986	0.774	2.032
SPA-PLS-STEM-NIR	0.705	2.543	0.687	2.633
GA-PLS-LEAF-VIS	0.812	1.585	0.783	1.754
SPA-PLS-LEAF-VIS	0.775	1.741	0.733	2.003
GA-PLS-STEM-VIS	0.746	2.001	0.711	2.113
SPA-PLS-STEM-VIS	0.699	2.723	0.655	2.722

to predict the prediction accuracies of the three enzymes in the VIS and NIR ranges for tomato leaves and stems. The average root mean square error ($\overline{\text{RMSEP}}$) and the average coefficient of determination ($\overline{R}_p$) for the three enzymes were used as evaluation metrics. The results showed that the validation set's $\overline{\text{RMSEP}}$ and $\overline{R}_p$ of the GA-PLS model for tomato leaves were 1.659 and 0.815 in the NIR and 1.754 and 0.783 in the VIS range, respectively. The validation set's $\overline{\text{RMSEP}}$ and $\overline{R}_p$ for tomato stems in the NIR band were 1.986 and 0.787, respectively, and in the VIS range were 2.113 and 0.711, respectively. By contrast, the validation set of SPA-PLS model for tomato leaves and stems had a higher $\overline{\text{RMSEP}}$ and lower $\overline{R}_p$ compared with the GA-PLS model, which indicates that the GA-PLS model has a higher prediction accuracy and stability than the SPA-PLS model.

In summary, two spectrum selection algorithms, GA and SPA, were used and combined with PLS regression to build a spectrum prediction model. The results show that the GA-PLS model is better than the SPA-PLS model, so GA is chosen as the spectrum selection algorithm.

3.5 Prediction Results

Four prediction models, GA-PLS, GA-SVM, GA-PSO-BP, and GA-RF, were developed to predict the enzyme activities of stems and leaves of the tomato plant in the NIR and VIS wavelengths, respectively.

The average root means square error of the three enzymes and average coefficient of determination for the three enzymes were used as evaluation metrics. The results obtained are presented in Table 3.

To visualize the prediction of tomato enzyme concentration, GA-PLS-LEAF-NIR was used as an example. The statistical analysis of the predicted and chemically detected values is presented in Table 4.

Table 3 Comparison of the accuracy of the four prediction models.

		Test set		Validation set	
Prediction model	Research object	$\overline{R}_p$	$\overline{\text{RMSE}}$	$\overline{R}_p$	$\overline{\text{RMSEP}}$
GA-PLS	LEAF-NIR	0.834	1.564	0.815	1.659
	STEM-NIR	0.787	1.986	0.774	2.032
	LEAF-VIS	0.812	1.585	0.783	1.754
	STEM-VIS	0.746	2.001	0.711	2.113
GA-SVM	LEAF-NIR	0.805	1.588	0.799	1.773
	STEM-NIR	0.782	1.865	0.756	1.866
	LEAF-VIS	0.788	1.792	0.776	1.911
	STEM-VIS	0.757	1.954	0.732	1.995
GA-PSO-BP	LEAF-NIR	0.865	1.005	0.832	1.559
	STEM-NIR	0.772	2.006	0.759	1.992
	LEAF-VIS	0.794	1.883	0.764	1.865
	STEM-VIS	0.702	1.954	0.694	2.131
GA-RF	LEAF-NIR	0.798	1.887	0.786	1.859
	STEM-NIR	0.783	1.873	0.752	1.975
	LEAF-VIS	0.734	2.113	0.710	2.221
	STEM-VIS	0.715	1.874	0.701	2.201

Table 4 Descriptive statistical analysis of the prediction of the three enzymes by the GA-PLS-LEAF-NIR model.

		Real value	Predicted value		Real value	Predicted value		Real value	Predicted value
Average		4.972666667	4.854095233		7.394828	7.306207		34.66433	38.18133
Standard error		0.070492411	0.068011376		0.129417	0.108856		1.10219	1.055575
Standard deviation	POD	0.386102839	0.372513647	PAL	0.69693	0.586207	α-AMS	6.036944	5.781625
Variance		0.149075402	0.138766417		0.485712	0.343639		36.44469	33.42718
Minimum value		4.33	4.11769		6.25	6.38		26.84	30.32
Maximum value		5.84	5.72769		8.81	8.55		44.36	47.96

3.6 Validation of a Predictive Model for Tomato Enzyme Content

The samples not involved in the modeling were randomly selected as the validation set, and the three enzyme contents of 21 samples were predicted by the GA-PLS-LEAF-NIR model. The scatter plots of the predicted and measured values are shown in Fig. 8, where real value represents the chemically detected enzyme activity and predicted value represents the enzyme activity predicted by the GA-PLS-LEAF-NIR model.

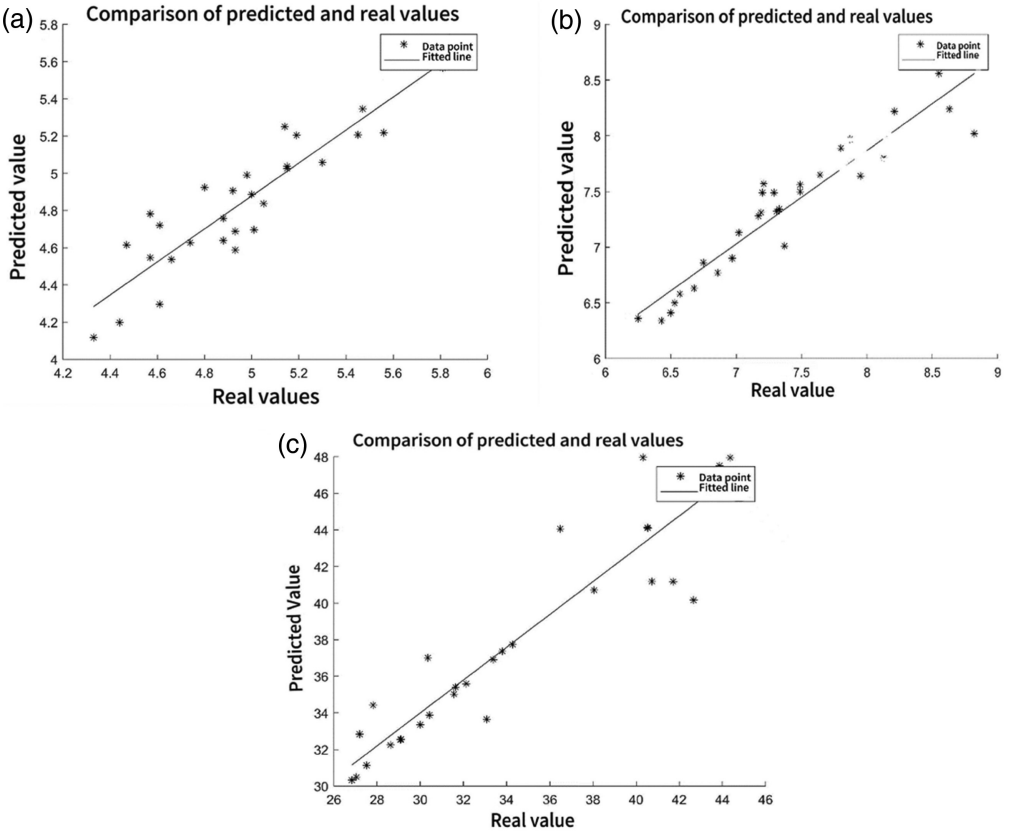

Fig. 8 Prediction effect of the validation set of the three enzymes: (a) comparison chart of POD enzyme content validation set data, (b) comparison chart of PAL enzyme content validation set data, and (c) comparison chart of α-AMS enzyme content validation set data.

The $\overline{R}_p$ of GA-PLS in the NIR leaf model was 0.815 and $\overline{\text{RMSEP}}$ was 1.659, which had good fit and prediction and helped to improve the model accuracy.

To compare the effectiveness of the four prediction models, they were applied to the problem of enzyme content prediction in tomato plants. The results showed that the GA-PLS model has the best prediction effect overall. Moreover, the effect of leaf prediction is better than that of stem for the prediction of enzyme content in tomato plants, and the effect of prediction in the NIR band is better than that in the VIS range. This also indicates that the GA-PLS model has a higher prediction accuracy and stability than the other three models.

4 Discussion

In this study, the sensitivity of spectral data to reflectance of enzymatic activity was analyzed on the spectral reflectance properties of tomato leaves and stems. In this study, the extraction of characteristic wavelengths was optimized through outlier elimination, sample set division, selection of pretreatment methods, and optimization, as well as the use of four algorithms to construct a prediction model of tomato enzyme activity, which provides a comprehensive non-destructive method for the detection of tomato enzyme activity.

Upon examining the spectral response to varying enzyme activities, it was observed that the spectral curve trends across different tomato plants exhibited similarities. The study identified that the spectra of leaves and stems with differing enzyme activities displayed an overall increasing trend in the 450 to 510 nm region, a smooth trend in the 700 to 800 nm range, and a descending trend in the 1350 to 1420 nm region. The average spectral reflectance of 140 tomato samples was extracted using the hyperspectral imaging system. Through outlier elimination, sample set partitioning, truncation operations, variable normalization, and two preprocessing methods applied to the original spectrum, a GA-PLS model was constructed for comparison. The results and graphical changes indicate that the preprocessing improves the modeling results compared with the raw spectra modeling. This enhancement can be attributed to the smoother spectral curve of the hyperspectral image preprocessing, increased model robustness, and a reasonable reduction in the effect of noise.[30]

In this study, GA was used for feature wavelength selection. Some researchers have also used competitive adaptive reweighted sampling to extract feature wavelengths to build a multiple linear regression prediction model for tomato leaves.[31] Others used SPA to extract the characteristic wavelengths to model the enzyme activity of leaves.[32] Both achieved optimal performance. This suggests that not all indicators are suitable for feature wavelength extraction using the GA. This study establishes four regression models based on GA: PLS, RF, PSO-BP, and SVM; it was found that all four models can detect enzyme activity. The performance of the model calibration set is superior to the validation set, but the difference is not significant, indicating the stable performance of these four methods without overfitting issues. Among them, the GA-PLS model performs optimally. It can effectively deal with multicollinearity problems in multiple regression. The correlation between predictors is reduced by extracting the principal components of the predictors, thereby enhancing the stability and prediction accuracy of the model.

Future research could be conducted in the field by expanding the sample size and integrating data over several years for comprehensive analysis, which could be practically applied for real-time, rapid, and non-destructive monitoring in the field. Because deep learning has strong learning and feature extraction capabilities, more algorithms can be explored to improve the accuracy of the model in future work. Using a more accurate and stable model, online monitoring equipment for the enzyme activity of other crops can be developed.

5 Conclusions

Among the prediction models, the best spectral selection method obtained is GA, the best model for predicting plants is GA-PLS, and the best study subject is tomato leaves in the NIR band. The prediction set of the three enzymes of this model $\overline{R}_p$ is 0.815 and $\overline{\text{RMSEP}}$ is 1.659, showing that the prediction is better. GA-PLS had the best prediction; GA-SVM, GA-PSO-BP, and GA-RF were not as good. The PLS regression method reduces the data dimension and noise, whereas the GA spectral selection algorithm filters out the most relevant wavelength subsets to improve the

search ability and robustness. The GA-PLS model has the highest accuracy and stability in predicting enzyme content in tomatoes. This is why the GA-PLS algorithm is superior to other algorithms. The results of the study were compared using two spectral options and four predictive models. It provides a theoretical basis for how to measure the enzyme content in tomato plants in a rapid and non-destructive manner. This paper also provides an effective method for phytochemical analysis using hyperspectral imaging.

Code and Data Availability

The research data and code are confidential in the industry. Due to the principle of confidentiality, the data cannot be shared.

Acknowledgments

We are grateful for external financial support, this study was supported by the Science and Technology Development Programme of Jilin Province, China (Grant No. YDZJ202301ZYTS241) and (Grant No. D21009) carried out under the auspices of the Discipline Innovation and Intelligence Programme of Higher Education Institutions (111 Programme); China Ministry of Education Belt and Road High-end Talent Program (Grant No. DL2023009002L).

References

1. M. Steinhauser et al., "Enzyme activity profiles during fruit development in tomato cultivars and solanum pennellii," *Plant Physiol.* **153**(1), 80–98 (2010).
2. B. Raza, A. Hameed, and M. Y. Saleem, "Fruit nutritional composition, antioxidant and biochemical profiling of diverse tomato (Solanum lycopersicum L.) genetic resource," *Front. Plant Sci.* **13**, 1035163 (2022).
3. M. Quinet et al., "Tomato fruit development and metabolism," *Front. Plant Sci.* **10**, 1554 (2019).
4. R. Thangaraj et al., "Artificial intelligence in tomato leaf disease detection: a comprehensive review and discussion," *J. Plant Dis. Prot.* **129**(3), 469–488 (2021).
5. F. Yasar and O. Uzal, "Oxidative stress and antioxidant enzyme activities in tomato (Solanum lycopersicum) plants grown at two different light intensities," *Gesunde Pflanzen* **75**(3), 479–485 (2022).
6. Z. Yongyang et al., "Genome wide analysis of the phenylalanine ammonia lyase (PAL) gene family from Senna tora," *Hubei Agric. Sci.* **62**(6),181–187 (2023).
7. L. Igor et al., "Identification of substances from diffuse reflectance spectra of a broadband quantum cascade laser using Kramers–Kronig relations," *Opt. Eng.* **59**(6) 061621 (2020).
8. L. Fimognari et al., "Simple semi-high throughput determination of activity signatures of key antioxidant enzymes for physiological phenotyping," *Plant Methods* **16**(1), 42 (2020).
9. J. Boeckx et al., "Kinetic modelling: an integrated approach to analyze enzyme activity assays," *Plant Methods* **13**(1), 69 (2017).
10. W. Fan et al., "Nondestructive determination of lycopene content based on visible/near infrared transmission spectrum," *Chin. J. Anal. Chem.* **46**(09), 1424–1431 (2018).
11. C. Li et al., "Spectral fusion based on hyperspectral imaging technology for discrimination of rice varieties," *Preprints*, 2024030886 (2024).
12. L. Yande, C. Mengjie, and H. Yong, "Application of spectral diagnoses technology in determination of agricultural products quality," *J. East China Jiaotong Univ.* **35**(4), 1–7 (2018).
13. Y. Xiang et al., "Deep learning and hyperspectral images based tomato soluble solids content and firmness estimation," *Front. Plant Sci.* **13**, 860656 (2022).
14. N. Fang-peng et al., "Hyperspectral estimation model of soil organic carbon content based on genetic algorithm fused with continuous projection algorithm," *Spectrosc. Spectral Anal.* **43**(7), 2232–2237 (2023).
15. A. ShiRong and L. M. MuHua, "Nondestructive measurement of acidity in strawberry using genetic algorithm and NIR spectroscopy," *J. Jiangxi Agric. Univ.* 633–636 (2010).
16. X. Chunhui et al., "Application of hyperspectral imaging technology in nondestructive testing of agricultural products," *Sci. Technol. Cereals, Oils Foods* **31**(1), 109–122 (2023).
17. W. Qiaohua, M. Yixiao, and F. Dandan, "Progress of non-destructive detection of poultry egg internal quality based on spectroscopy," *J. Huazhong Agric. Univ.* **40**(6), 220–230 (2021).
18. B. Park and R. Lu, *Hyperspectral Imaging Technology in Food and Agriculture*, Springer (2015).
19. F. Hui et al., "Detection of activity of POD in tomato leaves based on hyperspectral imaging technology," *Spectrosc. Spectral Anal.* **32**(8), 2228–2233 (2012).
20. P. Guo et al., "Evaluating calibration and spectral variable selection methods for predicting three soil nutrients using VIS-NIR spectroscopy," *Remote Sens.* **13**(19), 4000 (2021).

21. Y. Zhang et al., "Accurate and nondestructive detection of apple brix and acidity based on visible and near-infrared spectroscopy," *Appl. Opt.* **60**(13), 4021–4028 (2021).
22. X. G. Zhuang et al., "New induced mutation genetic algorithm for spectral variables selection in near infrared spectroscopy," *J. Appl. Spectrosc.* **87**(2), 260–266 (2020).
23. Y. Li et al., "SPA combined with swarm intelligence optimization algorithms for wavelength variable selection to rapidly discriminate the adulteration of apple juice," *Food Anal. Methods* **10**(6), 1965–1971 (2016).
24. M. A. Hearst, "Support vector machines," *IEEE Int. Syst. Appl.* **13**(4), 18–28 (1998).
25. M. Arumugasamy and A. Antonidoss, "An enhanced framework for categorization of fruits based on ripeness using ensemble PSO model," *Int. J. Electron. Commun. Eng.* **10**, 76–84 (2023).
26. W. Zhen et al., "Estimation of chlorophyll relative content in winter wheat by red edge parameters of canopy spectrum combined with random forest machine learning," *Trans. CSAE* **40**, 1–12 (2023).
27. J. Qingliang et al., "Spectral pre-processing based on convolutional neural network," *Spectrosc. Spectral Anal.* **42**(1), 292–297 (2022).
28. L. Lu et al., "Nonlinear normalization for non-uniformly distributed data," *Comput. Sci.* **43**(4), 264–269 (2016).
29. Y. Zhou, X. Li, and J. Cui, "High-efficiency hyperspectral unmixing based on band selection," in *Third Global Congr. Intell. Syst.*, Wuhan, China, pp. 140–143 (2012).
30. J. Hou, J. Tian, and J. Liu, "Spatial image filtering based on wavelet thresholding denoising," *Proc. SPIE* **6044**, 60440F (2005).
31. Y. F. Fei et al., "Identification and degree discrimination analysis of waterlogging stress in winter wheat based on hyperspectral remote sensing," *J. Smart Agric.* **3**(2), 35–44 (2021).
32. W. Jingyong et al., "Hyperspectral imaging inversion method of chlorophyll content and water content of maize leaves under drought stress," *Smart Agric.* **5**(3), 142–153 (2023).

Ao Li is a student at Changchun University of Science and Technology. He is interested in hyperspectral application and analysis, data processing, and image analysis and has participated in many experimental projects.

Jin-Shi Cui has been teaching since December 2019. Her research field is spectral analysis and image processing. As the project leader, she is currently implementing the project of "Research on Real-time Detection System of Tomato Leaf Enzyme Activity from Multiple Angles and Spectra" by Jilin Provincial Department of Science and Technology. As the first author, she has published a number of SCI, Scopus indexed papers.