Original Article
Food Science and Technology

ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.123322

Hyperspectral imaging with a band matrix reduction method to detect early drought stress in tomato

Gen ZHANG[1]*, Jianping HUANG[1], Yongjun MA[1], Jingzong ZHANG[1], Yi ZHANG[1]

Abstract

Drought stress is one of the key abiotic stresses affecting plant growth, crop yield and food quality. The main objective of this study is to investigate the potential effectiveness of hyperspectral imaging with band selection method for the rapid detection of the early drought stress of tomatoes. First, the unsupervised algorithm - K-means and statistical histogram are used to extract samples representing each experimental treatment group. Then, to solve problems related to the high redundancy and correlation of hyperspectral data, band matrix reduction method (BMRM) based on recursive feature elimination theory is proposed to determine the optimal band subset. The band matrix is constructed according to the band ranking obtained by the discrimination coefficient -$Coef(i)$, which is calculated from the average spectral curve and the first-derivative spectrum. Finally, the effectiveness of waveband selection algorithms was validated by comparison with successive projections algorithm, competitive adaptive reweighted sampling, recursive feature elimination with cross-validation and full spectrum. The results demonstrated that BMRM achieved higher classification accuracy with fewer bands selected, and the amount of calculation is not greatly improved. The proposed method provides a more accurate, and effective way of detecting early drought stress.

Keywords: hyperspectral imaging; drought stress; band selection; band matrix reduction method; classification.

Practical Application: Rapid and non-destructive testing of food quality.

1 Introduction

Drought stress is a phenomenon in which plant growth is significantly inhibited by the lack of available water due to drought. The available water of plants is usually expressed by water potential and measuring it before dawn is a direct method to assess plant water stress (Govender et al., 2009). Of course, indirect measurements can also be used to detect plant water stress, commonly used techniques include measuring the relative water content of leaves, chlorophyll pigment content of plants and canopy temperature, which correspond to the changes of plants under drought stress.

Hyperspectral imaging (HSI) technology can obtain continuous images of hundreds of bands, and each image pixel can extract a spectral curve. It combines the advantages of machine vision and near-infrared spectroscopy technology, which not only contains the spatial shape information of the measured sample, but also has spectral information highly related to the internal chemical composition. Therefore, it can not only detect the physical structure parameters of the sample, but also obtain its internal quality information. Further combined with the image and spectral interaction analysis method, the spatial distribution of the chemical composition and quality parameters of the sample can also be obtained. Compared with traditional measurement methods, HSI technology is a fast, accurate, sensitive and convenient nondestructive testing technology, it has great development potential to conduct multi-information fusion nondestructive testing in the comprehensive quality of samples.

Drought stress affects the physiological behavior of plants, leading to differences in reflex patterns, thus providing potential for remote sensing diagnosis of vegetation stress (Martin & Aber, 1997). With the potentially high spatial and spectral resolution and sensitivity of hyperspectral data, the technique is an excellent tool for distinguishing subtle differences in spectral reflectance that indicate early symptoms of stress (Delalieux et al., 2009).

However, the large amount of data collected by hyperspectral sensors brings challenges in analytical implementations. The redundancy problems are linked to the multicollinearity of bands and the curse of dimensionality imposes high computational costs on analytical pipelines (Burger & Gowen, 2011; Sun et al., 2022). The fundamental solution to these problems is to select appropriate methods to reduce dimensionality (Bruce et al., 2002).

At present, scholars have developed a series of band selection methods, the most popular and widely used methods are still the traditional spectral feature selection algorithms (Zheng et al., 2022), such as inter-correlation analysis (Chen et al., 2022), successive projections algorithm (SPA) and uninformative variable elimination (UVE). How to determine the number of band selection is still one of the challenges and future development trends of band selection.

Krishna et al. (2019) indicated that partial least squares regression (PLSR) is a robust technique for identification of water deficit stress in the crop. Zovko et al. (2019) used

Received 06 Dec., 2022
Accepted 21 Jan., 2023
[1]*Northeast Forestry University – NEFU, Harbin, Heilongjiang, China*
*Corresponding author: zhanggen0721@163.com

PLS-Discriminant Analysis to analyze the images of four water treatments, and treatments were classified using PLS-SVM. PLS-SVM demonstrated the capability to determine levels of grapevine drought or irrigated treatments with an accuracy of more than 97%. The selection of characteristic wavelengths by UVE and competitive adaptive reweighted sampling (CARS) is based on PLSR coefficients, but the calculation principles are different. Studies show that CARS can achieve better results than UVE (Li et al., 2014; Jun et al., 2019). Chen et al. showed that the PLSR prediction model established after extracting the sensitive bands by the CARS is better than the two-band exponential and PLSR models in both prediction accuracy and modeling accuracy, this method can provide reference for accurate and rapid monitoring of winter wheat drought and irrigation decisions (Chen et al., 2020). But the CARS extracts characteristic bands while retaining the bands with larger variable weights caused by noise (Qin et al., 2020). However, the SPA is a forward feature variable selection method, through the projection analysis of the vector, the collinearity among the selected spectral variables is minimized, and the redundant information is the least (Jian et al., 2019). For each stage of barley drought stress, Behmann et al. used the wrapper method to identify a distinctive set of most relevant VIs, which detected drought stress ten days earlier than using normalized difference vegetation index (NDVI) (Behmann et al., 2014). Recursive feature elimination (RFE) is a representative algorithm of wrapper method, the essence of it is a process of iteratively building models until the optimal feature subset is selected (Wu et al., 2017).

This study mainly aims to use HSI technology to detect different degrees of drought stress in the early stage, however, there are some difficulties. First, we focused on invisible early stage of drought stress, which is slow and complex, how to obtain data representing each group at the pixel scale is crucial. Second, to improve classification accuracy, and to reduce dimensionality, a band selection algorithm is needed to determine the optimal band subset to build the model, which can find the better balance between the validity and fastness of algorithm.

2 Materials and methods

2.1 Experimental design

This study takes tomato plants as the experimental object and the tomato was the pink crown series, and the seedlings were raised in a professional seedling base. After 40 days, choose tomato plants with the same growth and size and transplant them into flowerpots, one plant in a pot. The depth of each pot is 12.5 cm and the diameter is 13.5 cm, and there are multiple holes at the bottom for drainage. The soil adopts general organic nutrient soil, which is mainly composed of northeast virgin forest peat, vermiculite and perlite, and the maximum water holding capacity is 28%.

The drought stress experiment was carried out in the greenhouse of Northeast Forestry University in Harbin, China, the indoor photoperiod was 14 h/10 h (day/night), the ambient temperature was 24 °C/14 °C (day/night), and the air relative humidity was 60%. The experimental plants were divided into two groups, the well-watered group and the no-watered group. The no-watered group includes 2, 3, 4, 5, 6 and 7 day groups, there are seven groups in total, with three repetitions in each group. No more groups are set because there is a great difference between the shape of the leaves on the eighth day of no watering and the normal ones, which can be judged only by the naked eye. The plants were fertilized biweekly with 20-20-20 N-P-K water-soluble fertilizer, at a rate of 4 kg of N (Ihuoma & Madramootoo, 2019). To avoid the influence of other factors, the same amount of fertilizer was applied for each treatment. The soil moisture sensor is used to monitor the soil moisture in the non-watering group, which measure three times at different locations and take the average value as the actual value. When the soil moisture reaches 60% of the maximum water holding capacity, record the time as the start time of the experiment.

2.2 Hyperspectral data acquisition and preprocessing

Labscanner dual-light source scanning HSI system (spectral imaging Ltd., Oulu, Finland) is used in this study. As shown in Figure 1, it is composed of hyperspectral camera, light source, motion platform and control computer. The model of the hyperspectral camera is SPECIM FX 10, its spectral range is 400 nm-1000 nm, a total of 224 bands, and the spectral resolution is 5.5 nm. Before starting the acquisition, the system has been turned on for 30 minutes to achieve the thermal and temporal stability of the lighting system and hyperspectral camera (Erkinbaev et al., 2017). The scanner2017 software in the control computer is used to set relevant parameters. The frame rate of the camera is set to 20 frames per second, the moving speed of the platform is 20mm/s, the scanning speed is 5.5mm/s, and the exposure time is set to 10ms to make the camera not exposed.

To ensure the accuracy and repeatability of the results obtained by the HSI system, it is necessary to perform black and white correction on the acquired images. The system automatically closes the shutter to measure the black reference, while for the white reference, it scans the 99% Spectralon reflectance white

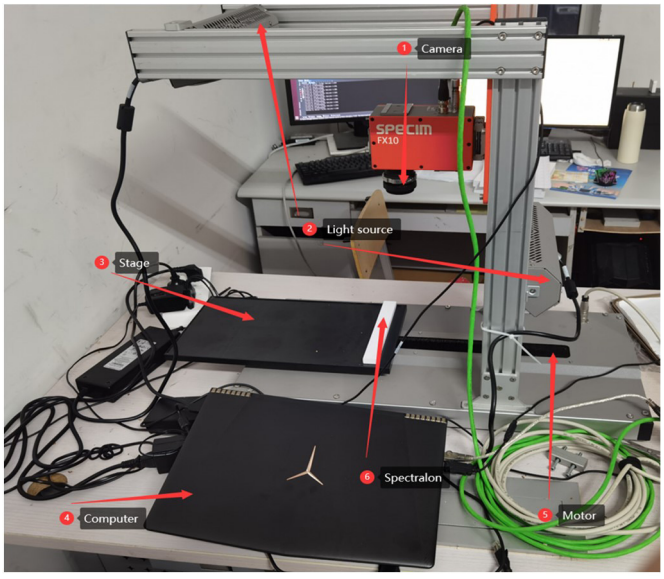

Figure 1. Schematic diagram of hyperspectral imaging system.

board (Labsphere, North Sutton, NH). The calculation formula of relative reflectance R is Equation 1 (Zou et al., 2022a):

$$R = \frac{R_0 - R_d}{R_w - R_d} \quad (1)$$

where, R_0 is the measurement image, R_d is the black reference image and R_w is the white reference image.

The regions of interest in the hyperspectral images corrected by black and white is extracted by using ENVI5.3 (Exelis Visual Information Solutions Inc., USA) software. (Wang et al., 2021; Zou et al., 2022b), which can basically eliminate shadow interference compared with using NVDI to extract the sample pixel, and then the data was exported to Excel. The data of the first 20 bands were removed to avoid the damage of noisy data in the spectral boundary region. The spectral reflectance contains a large amount of interfering noise due to atmospheric water absorption, signal background and light scattering effects, the Savitzky Golay filter is used to smooth the reflection curve to reduce the noise (Ma et al., 2022), the spectral curve after SG smoothing is shown in Figure 2.

2.3 Method

Clustering

The occurrence process of water stress is complex, which is a natural phenomenon with slow change, and we can't visualize hyperspectral high-dimensional data. Therefore, an unsupervised algorithm combined with statistical histogram is used to extract samples representing each degree of stress. K-means clustering, which has simple principle, fast convergence speed and better clustering effect, is suitable for problems with large amount of data.

Band selection

NDVI reflects crop growth and nutrition information by measuring the reflectance difference between near-infrared (strong reflectance of vegetation) and red light (absorption of vegetation). Shadrin et al. proposed a new discriminative coefficient ($DiscriminativeCoef(i,j)$) similar in structure to the NDVI to determine the optimal bandwidth for apple tree disease detection (Shadrin et al., 2020), the definition is Equation 2:

$$DiscriminativeCoef(i,j) = \frac{|AUC_1(i,j) - AUC_2(i,j)|}{AUC_1(i,j) + AUC_2(i,j)} \quad (2)$$

where AUC - area under an averaged spectrum of waveband, i - wavelength from each waveband started and j - width of waveband, the red area in Figure 3 represents $|AUC_1(i,j) - AUC_2(i,j)|$.

In Shadrin's study, the discrimination coefficients of all wavelengths and all possible band widths are calculated. To determine the specific band, in this study, j is taken as $2*di$ and the band range is $[i-di, i+di]$, and then AUC becomes Equation 3.

$$\begin{aligned} AUC_1(i,j) &= 2y_1(i-1)di + (y_1(i+1) - y_1(i-1))di \\ AUC_2(i,j) &= 2y_2(i-1)di + (y_2(i+1) - y_2(i-1))di \end{aligned} \quad (3)$$

where $y_1(i-1), y_2(i-1)$ represent the corresponding spectral reflectance at band i-1.

In addition, the original discrimination coefficient has a disadvantage that when the reflectance at band i is too high, compared with the band with low reflectance, it needs a large area difference to obtain the same discrimination coefficient, which is unfair to determine the importance of band. The ratio of the area difference at band i to the total area difference should be used as the discrimination coefficient. In this way, a new discrimination coefficient $Coef(i)$ is obtained, as shown in Equation 4, we can obtain desired bands according to the original spectral curve and the 1-Der spectrum combined with the threshold. The 1-Der spectrum is a preprocessing of the original data, in addition, the preprocessing also includes standard normal variation (SNV), multiple scatter correction (MSC) and etc (Zhang et al., 2022).

$$Coef(i) = \frac{AUC_i}{\sum_{k=1}^{n} AUC_k} = \frac{\left| y_1(i-1) + \frac{1}{2}y_{1(i,1st)} - \left(y_2(i-1) + \frac{1}{2}y_{2(i,1st)}\right)\right|}{\sum_{k=1}^{n}\left| y_1(k-1) + \frac{1}{2}y_{1(k,1st)} - \left(y_2(k-1) + \frac{1}{2}y_{2(k,1st)}\right)\right|} \quad (4)$$

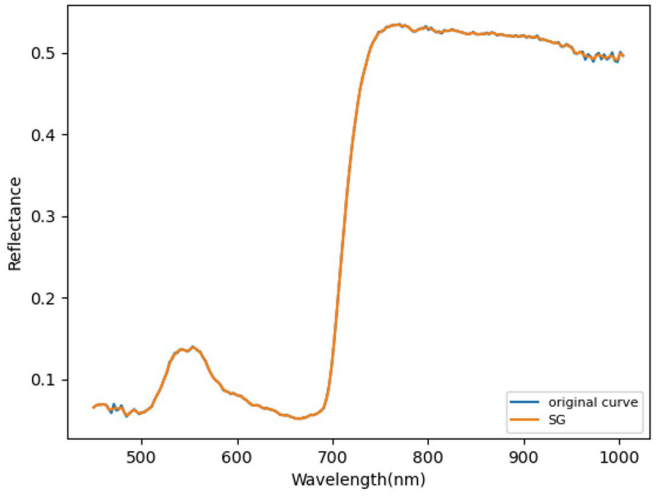

Figure 2. The spectral curve after SG smoothing.

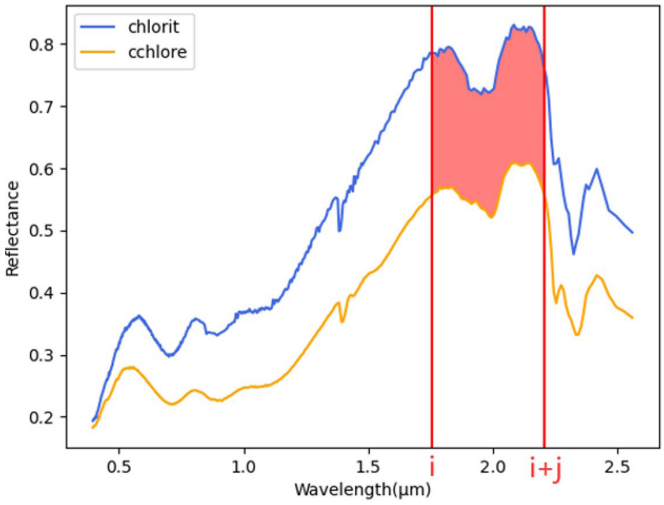

Figure 3. Two average spectral curves in the object spectral library.

where $y_{1(i,1st)}$, $y_{2(i,1st)}$ represent the spectral values of the spectrum at the wavelength i after the 1-Der with a difference width of 1.

Select optimal band subset

Using the band selection method can obtain the ordering of all bands, but cannot eliminate multicollinearity, and the optimal subset of bands cannot be determined. The most direct and effective way to determine the optimal band subset is to compare all possible band combinations, but in this way, with the increase of the number of bands, the amount of computation will increase exponentially, it is necessary to further reduce the number of bands. In the convolutional neural network, the filter or kernel will be used for convolution calculation to reduce the size of the original image and extract relevant features. With this idea, this paper proposes a new method, the band matrix reduction method (BMRM).

The essence of the BMRM is to select the best among the best. First, we number all bands (starting from 0, assuming there are $n \times m$ bands in total), then use the method in Section Band selection to get the ordering of all bands. According to the sorting order, m band numbers are formed into a row in turn, in this way, we can get the matrix of $n \times m$, which is called the band matrix, the matrix elements correspond to the bands one by one, the importance of the band in the same row is similar, but different rows are quite different. Next, it is necessary to reduce the size of the matrix to reduce the number of bands, and use the idea of Recursive Feature Elimination (RFE) to select the rows and columns of the matrix. Assuming that the size of the matrix after the first reduction is $p \times q$, we select any p row band numbers from the original matrix to form a new feature combination, use the results of Cross Validation (CV) to evaluate the p-line feature combinations, the classification algorithm adopts support vector machine (SVM), which has been proven to be a robust classification for handling massive datasets (Behmann et al., 2014). Comparing all possible row combinations, it is easy to know that there are C_n^p groups in total, the group with the maximum classification accuracy is selected to form a new matrix, the above operation is called row elimination.

Correspondingly, columns of the new matrix can be selected through column elimination after row elimination, so that the optimal reduced band matrix can be obtained through row elimination and column elimination. Finally, row elimination and column elimination are repeated until a suitable number of matrix elements is obtained, then the optimal band subset is determined by using CV results to compare all band combinations. The steps of the BMRM are shown in Figure 4.

For the band matrix with large size, it is necessary to set an appropriate reduction step and repeat the reduction several times until the appropriate number of bands are left. Since the number of rows and columns of the original matrix is large, setting a large reduction step will increase the calculation amount, it is appropriate to set the step to 1 or 2 at the beginning, but it can be appropriately increased later because $C_n^k = C_n^{n-k}$.

On the other hand, the number of elements of the final matrix needs to be determined. The number of calculations for different numbers of remaining elements is $\sum_{i=1}^{N} c_N^i$, where N represents the number of remaining elements, it can be known that 7-11 is more appropriate. For the size of the final band matrix, as with convolutional neural networks, determining the optimal convolution kernel size requires experience and experimentation, but the construction process of band matrix determines that selecting more rows is preferable to more columns.

3 Results and discussion

3.1 Clustering results

In the normal watering group, using the elbow method, we can determine that the number of clusters is 4, and then get the cluster center of each cluster. It is necessary to combine the statistical histogram to find the cluster that can represent the group, the specific results are shown in Figure 5.

In the normal watering group, although some leaf pixels of drought stress and edge areas will be included, the normal ones must account for a large part. It can be seen from the histogram that the black cluster and the pink cluster are the most, the number gap between the two clusters is small, but the gap with the other two clusters is large. From the above analysis, it is reasonable to use black and pink clusters to represent the normal watering group. Like the normal watering group, according to the clustering results, combined with the statistical histogram, the average spectral curves of each non-watering group as shown in Figure 6 can be obtained, which represent different degrees of drought stress.

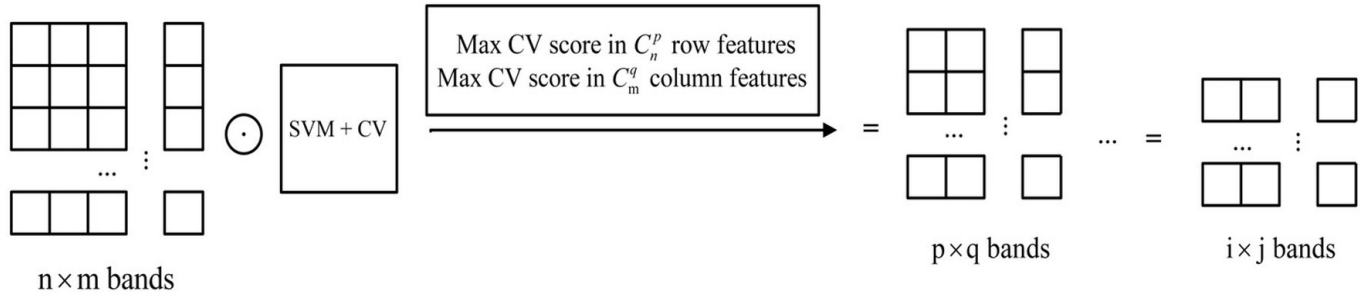

Figure 4. The specific implementation steps of the BMRM.

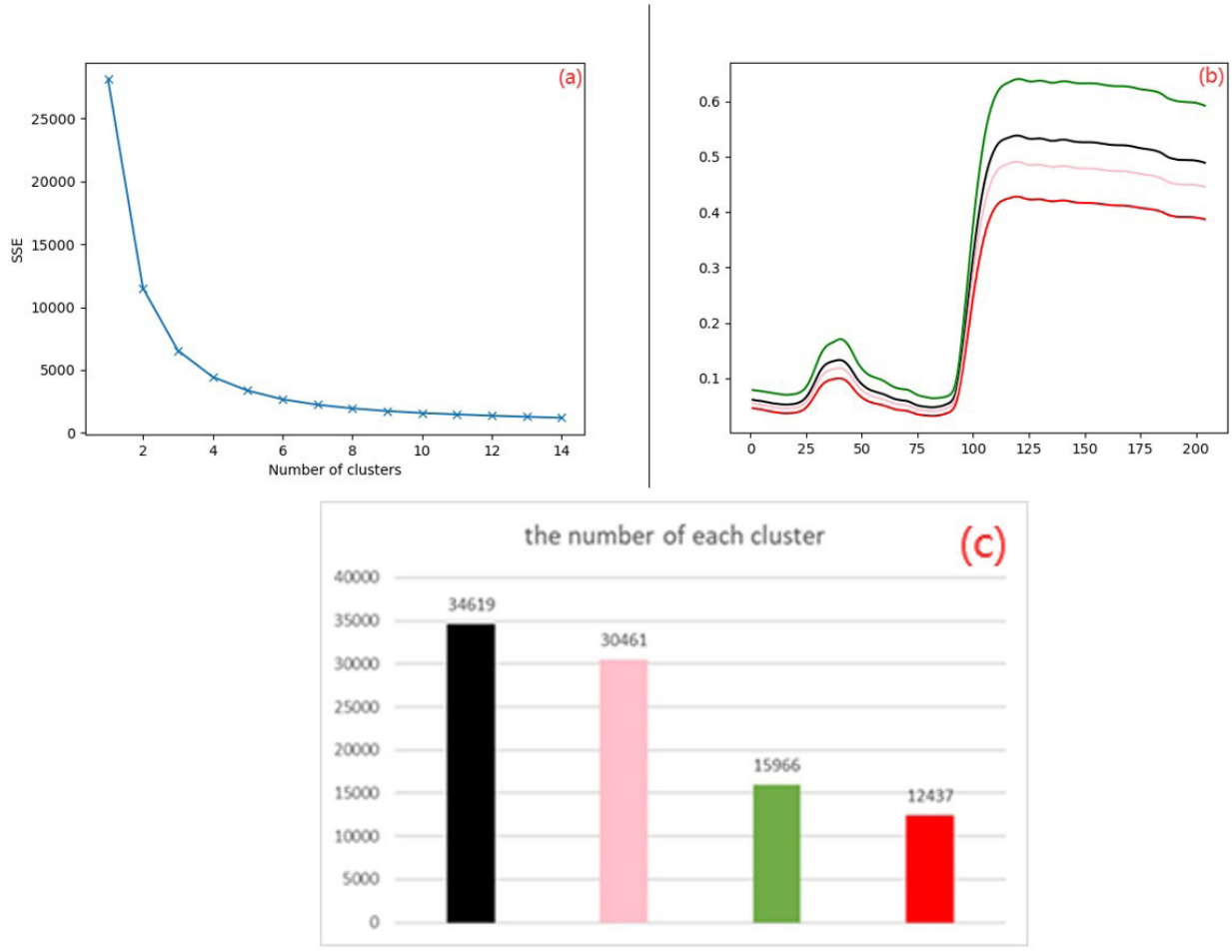

Figure 5. Clustering results and statistical histogram of normal watering group. (a) Determining the number of clusters (b) Cluster centers (c) The number of each cluster.

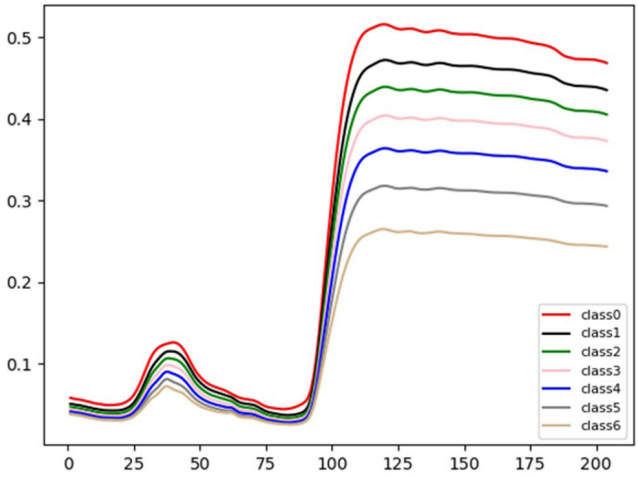

Figure 6. Average spectral curve of each experimental group.

3.2 Band selection

To verify the effectiveness of band selection, three different classification models are applied to compare our method with SPA, CARS, and full band. The results in Table 1 which takes the distinction between class 0 and class 1 as an example show the average value of classification accuracy that is obtained by five-fold cross-validation.

Table 1. Comparison results of different algorithms for all bands (Note: the number of bands selected by the algorithm is in brackets).

Algorithm / Model	SPA (3)	Full band	CARS (37)	Coef
Logistic regression	0.973	0.992	0.986	0.994(108),1(111)
Decision tree	0.984	0.977	0.968	0.977(145),1(200)
SVM	0.994	0.99	0.987	0.994(103),1(201)

It can be seen from Table 1 that although the new method can obtain the same effect as other algorithms, it requires more bands to model. In addition, from the average spectral curves of class 0 and 1, the reflectance of latter bands does not change significantly, and the difference between them is large, full band has good accuracy under different classification models.

Therefore, the first 110 bands are selected to repeat the experiment, and the results in Table 2 are obtained. Except for the

decision tree, modeling using the first few bands of the ranking can achieve same classification accuracy as SPA and CARS, and the effect of logistic regression is even better than SPA, while the amount of computation is negligible. This shows that the algorithm proposed in this paper has strong practicability when the collinearity is small, and the spectral curve gap is not large.

3.3 BMRM

We can learn from Section 3.2 that the first 110 bands have a lot of room for improvement. The effect of using SVM is better than other models, which further proves that SVM is a robust classifier, we will use SVM as the base model to select the optimal subset of bands among the first 110 bands.

Figure 7 shows the results obtained by using different RFECVs. Figure 7a is the normal RFECV, when 2 bands are selected, the CV score is the largest, which is 0.938. In normal SVM-RFECV, a linear kernel needs to be used to provide the band importance, if the selected bands are modeled with the Gaussian kernel in Figure 7b, the result is 0.946, which is slightly better than SPA. Figure 7b is the SVM-RFECV result obtained by directly using the CV results to evaluate the features, the optimal number of features is 7, and the maximum CV score is 0.9594. Compared with SPA and normal RFECV, it is greatly improved. However, for the feature elimination of 110 bands, 6104 times of five-fold cross-validation are required, the number of bands participating in modeling is proportional to computing times, and the number of multi-bands participating in modeling accounts for a large part.

Table 2. Comparison results of different algorithms for the first 110 bands.

Algorithm / Model	SPA (4)	Full band	CARS (20)	Coef
Logistic regression	0.93	0.934	0.88	0.94(1),0.946(3)
Decision tree	0.938	0.924	0.944	0.938(20)
SVM	0.945	0.927	0.942	0.945(5)

Although the above methods have achieved good results, the amount of calculation is huge, and the application of BMRM is intended to find a balance between the effect and the amount of calculation. First, 110 bands are grouped in groups of 10 according to the sorting results in Section 3.2 to form a 11*10 band matrix, since the number of rows and columns of the original matrix is large, at the beginning, it is more appropriate to set the step to 1 and set it to 2 at the end.

After comprehensively considering various factors, the change process of the band matrix size in this study is

$11\times10 \rightarrow 10\times9 \rightarrow 9\times8 \rightarrow 8\times7 \rightarrow 7\times6 \rightarrow 6\times5 \rightarrow 5\times4 \rightarrow 4\times2$, the final matrix is left with eight bands, an appropriate number obtained according to Section 2.3. Then calculate all possible combinations of these eight bands to obtain the maximum CV score corresponding to different number of band combinations as shown in Figure 8 the optimal number of features is 5, and the corresponding maximum CV score is 0.959, which is almost the same as the result in Figure 7b, the total number of calculations is only 255, and most of the bands involved in modeling are less than 8. Figure 9 shows the results obtained by constructing the band matrix using the band sorting of RFECV, the best number of features is 6, and the corresponding maximum CV score is 0.956, the gap is small, but the calculation is relatively complex.

To further determine the effect of matrix size on the results, the step was changed to make final band matrix size and number of elements different, all results are summarized in Table 3. Although the matrices of 4×3 and 6×2 achieve slightly better results than those in Figure 7b, the computations for a total of 12 bands are also heavy, reaching more than 4000 times. The effect of 5×2 is better than 4×2, but the difference is only 0.04%, which can be ignored, comprehensively comparing the calculation times, 4×2 is the best choice. In addition, the smaller the number of rows, the worse the effect. When the number of rows is fixed, increasing the number of columns does not change the overall effect, the final band matrix size must have enough rows, and a smaller number of columns has no major impact, which is consistent with the analysis in section 2.3.

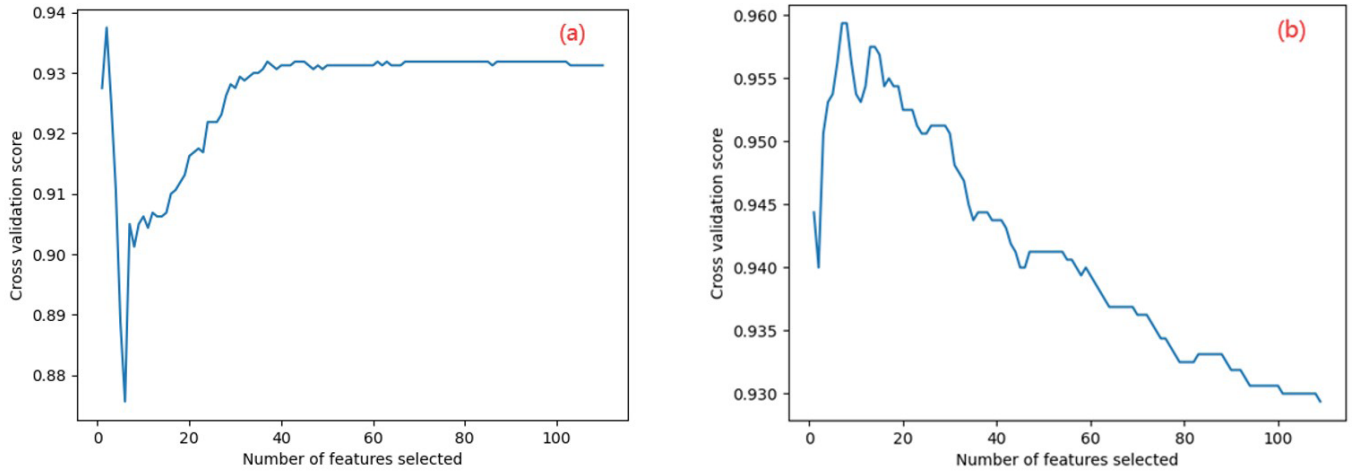

Figure 7. The curve of CV score with the number of selected bands under different RFECVs. (a) Normal RFECV (b) The RFECV obtained by using the results of CV to evaluate the feature.

Considering all band combinations is subject to the limitation of the number of bands. In the above experiment, although only 8 bands are reserved, the amount of calculation is still very large. Table 4 shows the running time of the program under different conditions, the running time of SPA and RFECV is similar, and the effect is similar, the running time of 8 bands is more than twice that of SPA, which is not counting the time of band matrix reduction. The runtime of 6 bands is very little, but further reducing the number of bands will not achieve good results. In addition, if the total number of bands is too large or no better results in the above matrix size, it is also necessary to increase the number of final matrix elements. After increasing the size of the final band matrix, it is more appropriate to evaluate the features with the results of CV, which can greatly reduce the amount of calculation. As shown in Figure 10, the matrix size is 4×3, and there are 12 bands left. The result is consistent with Figure 7b, the optimal number of features is 7, and the maximum CV score is 0.9594, which is very close to Table 3. A total of $\frac{12\times11}{2}-1=65$ calculations are required, and the running time is 24.62s, increasing the final band matrix size will also reduce the matrix reduction time, which is 36.24s, the total sum of the two is 60.86s, which is slightly higher than SPA and RFECV, but the effect improvement is obvious.

Table 3. Summary of results for different final band matrix sizes.

Size	4 × 3	6 × 2	5 × 2	2 × 5	2 × 4
Optimal number of bands	5	7	6	4	4
Max CV score	0.96	0.961	0.9594	0.955	0.955

Table 4. Program running time in different situations (Python).

Types	RFECV	SPA	8bands	9bands	6bands
Program running time	44.65s	47.33s	103.65s	154.99s	17.88s

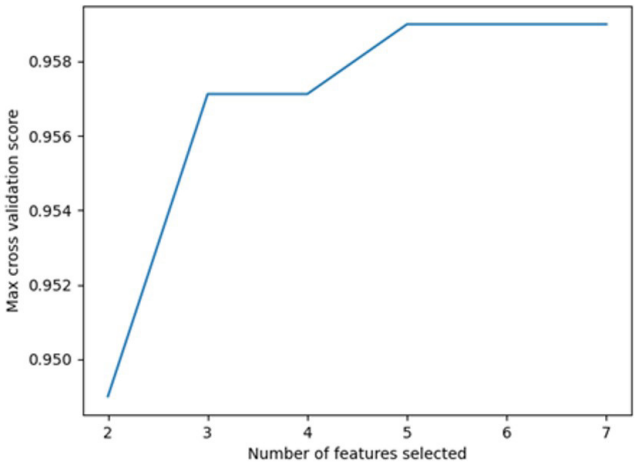

Figure 8. Maximum CV scores corresponding to different number of band combinations

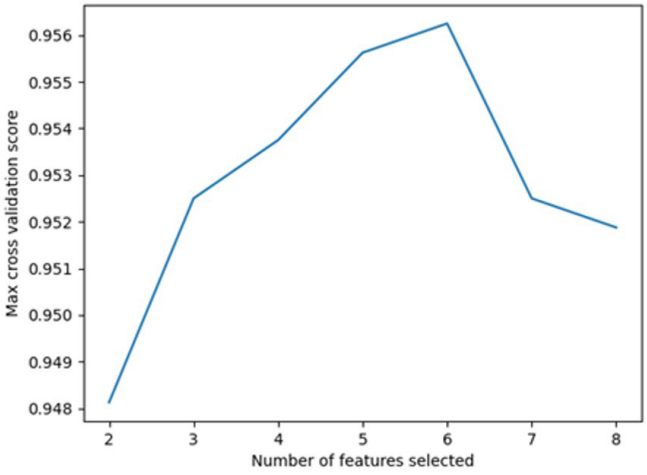

Figure 9. Maximum CV scores corresponding to different number of band combinations obtained by using band sorting of RFECV.

3.4 The effect of BMRM in distinguishing other classes

The results in Sections 3.2 and 3.3 are based on the classification of class 0 and 1. In this paper, it is necessary to classify different degrees of drought stress, a total of seven classes. For multi-class problems, there are two approaches: one-to-many and one-to-one, the one-to-one approach is used to further verify the band matrix reduction proposed in this study. Samples far away from the cluster center were removed to avoid the influence of some extreme points, and then 1000 samples were randomly selected for each class. From the average spectral curve and Table 5, there is a large gap between the two types of samples that are not adjacent to each other, and the classification accuracy close to 1 or equal to 1 can be obtained by modeling all the first 110 bands. BMRM is used to solve the classification problem of adjacent classes, the size of final band matrix is 4×3, and the CV results are used to evaluate the features directly.

Applying SPA, CARS, RFECV and our method to the first 110 bands in Sections 3.2 and 3.3 shows that SPA is better than

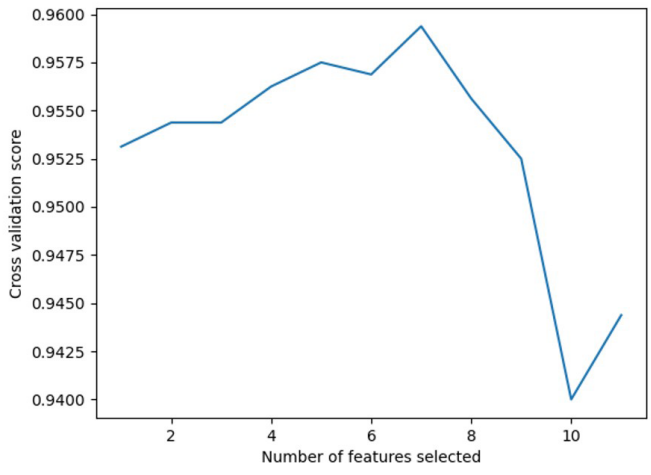

Figure 10. After the reduction of the band matrix, the CV results are directly used to evaluate the features.

Table 5. Classification accuracy of non-adjacent class using full band.

Class	0/2	1/3	2/4	3/5	4/6	5/7	6/8	7/9
Accuracy	0.991	0.994	0.991	0.997	0.997	1	1	1

Table 6. Statistical table of classification of adjacent classes.

Class	1/2	2/3	3/4	4/5	5/6
Optimal number of bands	4	5	5	7	8
Accuracy of full bands	0.85	0.843	0.895	0.917	0.907
Accuracy of SPA	0.915(4)	0.902(5)	0.937(6)	0.936(9)	0.923(10)
Accuracy of BMRM	0.942	0.945	0.963	0.958	0.944
Accuracy difference	0.092	0.102	0.068	0.041	0.037

CARS in both dimensionality reduction and classification accuracy, which is almost the same as RFECV. SPA has improved a little accuracy of classification upon full band, and BMRM is better than SPA, but the overall effect is not obvious, this is because using full band has achieved better classification accuracy, close to 93%, it is relatively difficult to increase on this basis. Therefore, in the following analysis, the first 100 bands are used for experiments to obtain the results in Table 6, focusing on the comparison of full band and SPA with the best effect in the experiment.

From Table 6, we can know that the classification accuracy has been greatly improved, the lowest accuracy difference is 3.7%, which corresponds to the higher full-band classification accuracy, and the highest is close to 11%, which corresponds to the worst full-band classification accuracy. Moreover, the number of bands involved in the final modeling is equal to or less than SPA, and the value range is [4,8], indicating that the method in this paper has a good dimensionality reduction effect.

4 Conclusion

In this work, it was demonstrated that HSI is a promising rapid and nondestructive technique for the detection of water stress in tomato plants. The proposed method can distinguish different degrees of drought stress from spectral reflectance by a data-driven method that combines clustering and optimal band subset selection, even before visible changes occur to the naked eye. The experimental results showed that the average spectral curve contains information about sensitive bands, the band ordering obtained from the average spectral curve has good effect in the case of small collinearity, and the amount of computation is negligible. The optimal band subset selection method - BMRM proposed by the idea of recursive feature elimination is a method of selecting the best from the best, which is an extreme wrapper approach. Compared to CARS, SPA and RFECV wavelength selection method, the BMRM can reduce the model complexity, extract the more significant information related to water stress and make the model more stable. The results of this study illustrated that hyperspectral data coupled with BMRM was a powerful tool for evaluating the water stress of plants.

HSI technology has been widely used in the relevant fields of food industry, such as the detection of internal and external defects of food, quality classification, chemical composition detection, etc. Then the application of the band selection method proposed in this paper will greatly improve the efficiency of the HSI detection system, which can realize the online, rapid and non-destructive testing of food comprehensive quality and safety.

References

Behmann, J., Steinrücken, J., & Plümer, L. (2014). Detection of early plant stress responses in hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 98-111. http://dx.doi.org/10.1016/j.isprsjprs.2014.03.016.

Bruce, L. M., Koger, C. H., & Li, J. (2002). Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10), 2331-2338. http://dx.doi.org/10.1109/TGRS.2002.804721.

Burger, J., & Gowen, A. (2011). Data handling in hyperspectral image analysis. *Chemometrics and Intelligent Laboratory Systems*, 108(1), 13-22. http://dx.doi.org/10.1016/j.chemolab.2011.04.001.

Chen, M., Ni, Y., Jin, C., Liu, Z., & Xu, J., 2022. Spectral inversion model of the crushing rate of soybean under mechanized harvesting. *Food Science and Technology*, 42, e123221. https://doi.org/10.1590/fst.123221.

Chen, X. Q., Yang, Q., Han, J. Y., Lin, L., & Shi, L. S. (2020). Estimation of winter wheat leaf water content based on leaf and canopy hyperspectral data. *Spectroscopy and Spectral Analysis*, 40(3), 891. https://doi.org/10.3964/j.issn.1000-0593(2020)03-0891-07.

Delalieux, S., Somers, B., Verstraeten, W. W., van Aardt, J. A. N., Keulemans, W., & Coppin, P. (2009). Hyperspectral indices to diagnose leaf biotic stress of apple plants, considering leaf phenology. *International Journal of Remote Sensing*, 30(8), 1887-1912. http://dx.doi.org/10.1080/01431160802541556.

Erkinbaev, C., Henderson, K., & Paliwal, J. (2017). Discrimination of gluten-free oats from contaminants using near infrared hyperspectral imaging technique. *Food Control*, 80, 197-203. http://dx.doi.org/10.1016/j.foodcont.2017.04.036.

Govender, M., Govender, P. J., Weiersbye, I. M., Witkowski, E. T. F., & Ahmed, F. (2009). Review of commonly used remote sensing and ground-based technologies to measure plant water stress. *Water S.A.*, 35(5). http://dx.doi.org/10.4314/wsa.v35i5.49201.

Ihuoma, S. O., & Madramootoo, C. A. (2019). Sensitivity of spectral vegetation indices for monitoring water stress in tomato plants. *Computers and Electronics in Agriculture*, 163, 104860. http://dx.doi.org/10.1016/j.compag.2019.104860.

Jian, H., Li, Y. Z., Cao, Z. M., Qiang, L., & Mou, H. W. (2019). Water content prediction for high water-cut crude oil based on SPA-PLS using near infrared spectroscopy. *Spectroscopy and Spectral Analysis*, 39(11), 3452-3458.

Jun, H., Yande, L., Aiguo, O., & Hongliang, L. (2019). Mid-infrared spectroscopy detection of methanol content in methanol gasoline based on CARS band screening. *Laser & Optoelectronics Progress*, 56(23), 233002. http://dx.doi.org/10.3788/LOP56.233002.

Krishna, G., Sahoo, R. N., Singh, P., Bajpai, V., Patra, H., Kumar, S., Dandapani, R., Gupta, V. K., Viswanathan, C., Ahmad, T., & Sahoo, P. M. (2019). Comparison of various modelling approaches for water deficit stress monitoring in rice crop through hyperspectral remote

sensing. *Agricultural Water Management*, 213, 231-244. http://dx.doi.org/10.1016/j.agwat.2018.08.029.

Li, J. B., Peng, Y. K., Chen, L. P., & Huang, W. Q. (2014). Near-infrared hyperspectral imaging combined with cars algorithm to quantitatively determine soluble solids content in "Ya" pear. *Guang pu xue yu guang pu Fen Xi = Guang Pu*, 34(5), 1264-1269. PMid:25095419.

Ma, X., Luo, H., Liao, J., Zhu, L., Zhao, J., & Gao, F., (2022). Study on the detection of apple soluble solids based on fractal theory and hyperspectral imaging technology. *Food Science and Technology (Campinas)*, 43, e96722. https://doi.org/10.1590/fst.96722.

Martin, M. E., & Aber, J. D. (1997). High spectral resolution remote sensing of forest canopy lignin, nitrogen, and ecosystem processes. *Ecological Applications*, 7(2), 431-443. http://dx.doi.org/10.1890/1051-0761(1997)007[0431:HSRRSO]2.0.CO;2.

Qin, L. F., Zhang, X., & Zhang, X. Q. (2020). Early detection of cucumber downy mildew in greenhouse by hyperspectral disease differential feature extraction. *Transactions of the Chinese Society for Agricultural Machinery*, 51(11), 212-220. http://dx.doi.org/10.6041/j.issn.1000-1298.2020.11.023.

Shadrin, D., Pukalchik, M., Uryasheva, A., Tsykunov, E., Yashin, G., Rodichenko, N., & Tsetserukou, D. (2020). Hyper-spectral NIR and MIR data and optimal wavebands for detection of apple tree diseases. *arXiv*, arXiv:2004.02325, 1-6. https://doi.org/10.48550/arXiv.2004.02325.

Sun, J., Hu, Y., Zou, Y., Geng, J., Wu, Y., Fan, R., & Kang, Z. (2022). Identification of pesticide residues on black tea by fluorescence hyperspectral technology combined with machine learning. *Food Science and Technology (Campinas)*, 42, 42. http://dx.doi.org/10.1590/fst.55822.

Wang, X., Xing, X., Zhao, M., & Yang, J. (2021). Comparison of multispectral modeling of physiochemical attributes of greengage: Brix and pH values. *Food Science and Technology (Campinas)*, 41(Suppl 2), 611-618. http://dx.doi.org/10.1590/fst.21320.

Wu, C., Liang, J., Wang, W., & Li, C. (2017). Random forest algorithm based on recursive feature elimination method. *Statistics & Decisions*, 21, 60-63. https://doi.org/10.13546/j.cnki.tjyjc.2017.21.014.

Zhang, Y., Wang, J., Luo, H., Yang, J., Wu, X., Wu, Q., & Zhong, Y. (2022). Rapid prediction of Yongchuan Xiuya tea quality by using near infrared spectroscopy coupled with chemometric methods. *Food Science and Technology (Campinas)*, 43, e101122. https://doi.org/10.1590/fst.101122.

Zheng, Z., Liu, Y., He, M., Chen, D., Sun, L., & Zhu, F. (2022). Effective band selection of hyperspectral image by an attention mechanism-based convolutional network. *RSC Advances*, 12(14), 8750-8759. http://dx.doi.org/10.1039/D1RA07662K. PMid:35424797.

Zou, Z., Wang, L., Chen, J., Long, T., Wu, Q., & Zhou, M. (2022a). Research on peanut variety classification based on hyperspectral image. *Food Science and Technology*, 42, e18522. https://doi.org/10.1590/fst.18522.

Zou, Z., Wu, Q., Chen, J., Long, T., Wang, J., Zhou, M., Zhao, Y., Yu, T., Wang, Y., & Xu, L. (2022b). Rapid determination of water content in potato tubers based on hyperspectral images and machine learning algorithms. *Food Science and Technology (Campinas)*, 42, e46522. http://dx.doi.org/10.1590/fst.46522.

Zovko, M., Žibrat, U., Knapič, M., Kovačić, M. B., & Romić, D. (2019). Hyperspectral remote sensing of grapevine drought stress. *Precision Agriculture*, 20(2), 335-347. http://dx.doi.org/10.1007/s11119-019-09640-2.