



OPEN

Early diagnosis of *Cladosporium fulvum* in greenhouse tomato plants based on visible/near-infrared (VIS/NIR) and near-infrared (NIR) data fusion

Xuerong Zhao¹, Yuanyuan Liu², Zongbao Huang³, Gangao Li¹, Zilin Zhang¹, Xiuhan He¹, Huiling Du⁴, Meiqin Wang²✉ & Zhiwei Li^{1,3}✉

Plant diseases can inflict varying degrees of damage on agricultural production. Therefore, identifying a rapid, non-destructive early diagnostic method is crucial for safeguarding plants. *Cladosporium fulvum* (*C. fulvum*) is one of the major diseases in tomato growth. This work presents a method of data fusion using two hyperspectral imaging systems of visible/near-infrared (VIS/NIR) and near-infrared (NIR) spectroscopy for the early diagnosis of *C. fulvum* in greenhouse tomatoes. First, hyperspectral images of samples at health and different times of infection were collected. The average spectral data of the image regions of interest were extracted and preprocessed for subsequent spectral datasets. Then different classification models were established for VIS/NIR and NIR data, optimized through various variable selection and data fusion methods. The principal component analysis-radial basis function neural network (PCA-RBF) model established using low-level data fusion achieved optimal results, achieving accuracies of 100% and 99.3% for calibration and prediction, respectively. Moreover, both the macro-averaged F1 (Macro-F1) values reached 1, and the geometric mean (G-mean) values reached 1 and 1, respectively. The results indicated that it was feasible to establish a PCA-RBF model by using the hyperspectral technique with low-level data fusion for the early detection of *C. fulvum* in greenhouse tomatoes.

Keywords Data fusion, Early diagnosis, Hyperspectral imaging, *Cladosporium fulvum*, Tomato leaf

Plant diseases can be categorized into infectious diseases and non-infectious diseases¹, with infectious diseases (caused by biotic factors) posing significant threats that require particular attention. Infectious diseases are primarily caused by pathogenic organisms. These pathogenic organisms encompass a broad spectrum of micro-organisms, spanning bacteria, fungi, viruses, and nematodes^{1,2}. They adversely affect crop growth, development, and yield. Early disease diagnosis is highly important in agriculture. First, early disease diagnosis enables timely isolation, disinfection, and application of disease control agents, thereby enhancing agricultural production efficiency and preventing disease spread from causing losses to crops. Second, this technology can lower control costs, facilitate precision application, mitigate environmental pollution risks, and promote sustainable agricultural development. Moreover, early disease diagnosis enhances the quality and safety of agricultural products, thereby ensuring food safety³⁻⁵.

With the increasing population and the growing pursuit of high-quality living, traditional farming methods are no longer able to meet market demands. Consequently, greenhouse technology has become an important development direction in the modern agricultural field⁶. Tomatoes grown in greenhouses are crucial crops worldwide and have consistently maintained high production value^{7,8}. However, greenhouse tomatoes are frequently plagued by various diseases, significantly reducing their yields, with *Cladosporium fulvum* (*C. fulvum*) being among the most detrimental. *C. fulvum*, commonly known as powdery mildew, poses a significant threat

¹College of Agricultural Engineering, Shanxi Agricultural University, Jinzhong 030801, China. ²College of Plant Protection, Shanxi Agricultural University, Jinzhong 030801, China. ³College of Information Science and Engineering, Shanxi Agricultural University, Jinzhong 030801, China. ⁴Department of Basic Sciences, Shanxi Agricultural University, Jinzhong 030801, China. ✉email: mqwang@sxau.edu.cn; lizhiweitong@163.com

to tomato production as a fungal disease. It is characterized by the development of grayish-white mold on tomato leaves. The grayish-white mold is composed of mycelium and spores, creating a powdery white substance on the leaf surface, thus earning *C. fulvum* the moniker "powdery mildew". While the disease primarily impacts tomato leaves, severe cases may affect tomato stems, flowers, and fruits. The initial symptoms of *C. fulvum* include the emergence of small white spots on leaves, which subsequently enlarge into circular or irregular spots. Disease manifestations include yellowing, bending, or even death of plant leaves, significantly impairing plant growth and development. Severe cases may result in the infestation and subsequent rotting of tomato stems, flowers, and fruits, thereby greatly compromising tomato yield and quality^{9,10}. *C. fulvum* is a globally prevalent disease, that is particularly prevalent in hot and humid environments where it proliferates more frequently. Therefore, research on the early diagnosis of greenhouse tomato diseases is urgently needed.

Spectroscopy plays a pivotal role in the early detection of crop diseases, leveraging its high sensitivity, resolution, and remote sensing capabilities to become a potent diagnostic tool^{11–13}. Among these techniques, hyperspectral imaging stands out as an advanced technology that integrates optics, spectroscopy, and image processing. It relies mainly on the reflection of objects at various wavelengths to acquire comprehensive spectral data and construct a 3D dataset of the target area^{14–16}. These datasets facilitate the analysis and identification of various substances, as well as the detection of vegetation health and environmental changes^{17,18}. Huang et al.¹⁹ utilized a hyperspectral imager to capture 200 visible/near-infrared (VIS/NIR) hyperspectral images of both healthy and early-stained blueberries. Following spectral data extraction, they constructed partial least squares discriminant analysis (PLS-DA) models for the full spectrum and the characteristic spectrum. The findings revealed that the characteristic spectrum yielded superior classification outcomes, achieving 100% and 99% identification rates for healthy and early-stage diseased blueberries, respectively. Ugarte et al.²⁰ used black Sigatoka disease in bananas to establish a partial least squares penalized logistic regression (PLS-PLR) model using a hyperspectral imaging system to collect images of healthy and diseased plants. The results showed that 98% accuracy could be achieved using the PLS-PLR model, and wavelengths of 577–651 nm and 700–1019 nm had the greatest influence on the classification results.

Spectral data fusion entails amalgamating data from various bands or sensors to enhance the comprehensiveness and accuracy of information. Feng et al.²¹ employed VIS/NIR hyperspectral imaging spectroscopy, mid-infrared (MIR) spectroscopy, and laser-induced breakdown spectroscopy to acquire leaf spectral data from healthy and diseased rice species, namely, rice bacterial leaf blight, rice blast, and stripe blight. The spectral data were fused at low, medium, and high levels to generate three new datasets. These datasets, along with the original spectral data, underwent feature extraction via principal component analysis (PCA) and autoencoder (AE), after which support vector machine (SVM), logistic regression (LR), and convolution neural network (CNN) models were established for recognition. The findings indicated that except for low-level data fusion which did not improve the accuracy of the classification models, both medium-level and high-level data fusion enhanced the model's recognition performance. High-level data fusion performed better, achieving 100% accuracy particularly on the calibration and prediction sets. Xiao et al.²² acquired hyperspectral images of *Astragalus* from various regions using a VIS/NIR spectroscopy hyperspectral imaging system and a near-infrared (NIR) spectroscopy hyperspectral imaging system. They extracted image spectra and conducted feature extraction using PCA and CNN, followed by low-level and medium-level spectral data fusion to construct SVM, LR, and CNN recognition models. The results demonstrated that all models achieved accuracies exceeding 98%. The model combining hyperspectral imaging with CNN and data fusion yielded the highest accuracy. Yu et al.²³ employed hyperspectral imaging to study tilapia fillets' TVB-N content, using various feature extraction methods and data fusion techniques. The results indicated that low-level data fusion outperformed individual data blocks, and medium-level fusion combined with competitive adaptive reweighted sampling (CARS) achieved the highest performance.

The VIS and NIR wavelengths possess a unique ability to reflect the health status of plants²⁴. By integrating data from these two bands, it becomes possible to acquire more detailed spectral information, thereby enhancing the sensitivity and accuracy of detecting *C. fulvum* in tomato plants. The integrated data can effectively highlight changes in vegetation, facilitating the detection of infected plants. Currently, while numerous studies have explored early disease diagnosis using hyperspectral techniques, the study of *C. fulvum* in tomatoes only have disease classification²⁵, without the early diagnosis of *C. fulvum* by inoculating tomato leaves with pathogens and using VIS/NIR hyperspectral imaging technology and NIR hyperspectral imaging technology combined with machine learning and data fusion methods.

Therefore, the aim of this study is to use hyperspectral imaging technology for the early detection of *C. fulvum* in greenhouse tomatoes. We investigated the effectiveness of VIS/NIR and NIR spectroscopy in distinguishing between healthy and diseased samples at various stages of infection. Feature extraction methods and data fusion techniques were applied to improve classification accuracy, and the optimal model for rapid and non-destructive detection of early-stage *C. fulvum* in greenhouse tomatoes was determined. The flow chart is shown in Fig. 1.

Materials and methods

Experimental design and sample collection

This experiment was carried out between May and August 2023 in the research greenhouse (37°25' N, 112°34' E). Tomato seedlings were selected from Provence tomatoes with high yield, excellent quality, and strong stress resistance. One hundred and twenty Provence tomato plants were grown in pots with coco coir as a substrate and fertilized with self-formulated water-soluble fertilizer: calcium nitrate (1216 mg/L), urea (131.67 mg/L), calcium ammonium nitrate (42.1 mg/L), potassium nitrate (395 mg/L), potassium dihydrogen phosphate (208 mg/L), potassium sulfate (393 mg/L), magnesium sulfate (466 mg/L). The temperature of the greenhouse was maintained at 20 °C to 25 °C, the humidity was maintained at 60% to 70%, and the light time was 12 h. The tomato plants were experimented with after reaching the seedling stage. Since the leaves below the plant canopy were infected

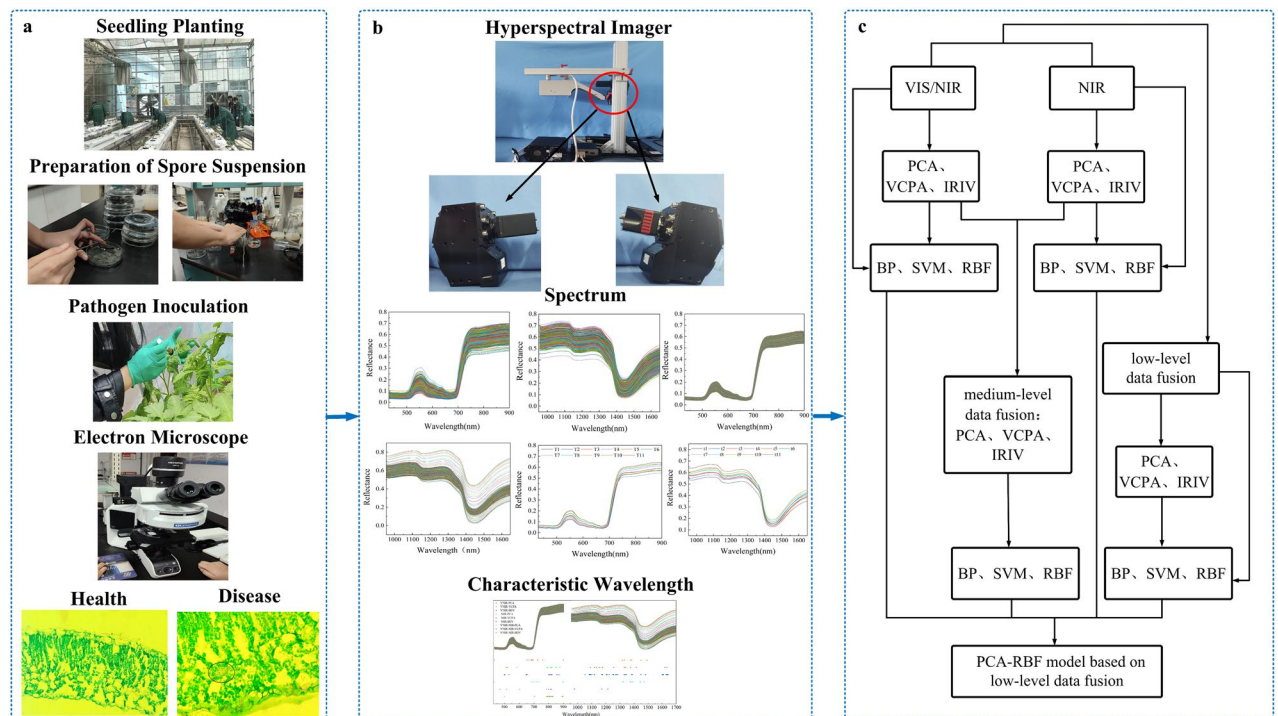


Fig. 1. Analysis process flow chart; **a** A series of preparations; **b** A series of spectra of visible/near-infrared and near-infrared; **c** A series of classification recognition models built by visible/near-infrared and near-infrared spectral data.

more quickly than those in the plant canopy²⁶, leaves from below the plant canopy were selected to capture hyperspectral images. The collection of plant material complies with relevant institutional, national, and international guidelines and legislation. Initially, hyperspectral images in the visible-near infrared and near-infrared range were captured for 60 healthy leaves using a hyperspectral imaging device. Subsequently, all leaves were inoculated with *C. fulvum* pathogens. Spore suspensions for *C. fulvum* were provided by students from the College of Plant Protection. Following inoculation, bacterial invasion was observed using a microscope at 2 h intervals. Upon observing bacterial invasion, a large number of leaves from below the plant canopy were promptly collected and returned to the laboratory for hyperspectral imaging. Subsequently, a hyperspectral imager was used to dynamically monitor the evolution of bacterial invasion at 12 h intervals. Eventually, 1374 valid samples were acquired, comprising 687 VIS/NIR samples and 687 NIR samples. The numbers of leaves exhibiting symptomatic disease progression from 12 to 120 h postinoculation were 61, 60, 60, 60, 62, 63, 61, 63, 71, and 66 leaves, respectively. The VIS/NIR and NIR samples collected during periods of health and disease were labeled T1 to T11 and t1 to t11, respectively, while the combined VIS/NIR and NIR datasets were denoted T and t, respectively. To balance the distribution of sample categories, a subset of 55 samples (totaling 605) was randomly chosen from the VIS/NIR samples in each class to facilitate model calibration and prediction. The NIR samples underwent the same selection process as the VIS/NIR samples.

Hyperspectral imaging acquisition

Leaf hyperspectral images were captured using a VIS/NIR spectroscopy hyperspectral imaging system and a NIR spectroscopy spectroscopy hyperspectral imaging system (*Headwall Photonics, USA*). In the hyperspectral imaging systems of VIS/NIR spectroscopy, the spectral resolution is 0.727 nm, spanning from 380 to 1000 nm, with a total of 856 bands. In the hyperspectral imaging systems of NIR spectroscopy, the spectral resolution is 4.715 nm, spanning from 900 to 1700 nm, comprising a total of 172 bands. To prevent radiation counter saturation or low signal-to-noise ratio during spectral measurements, spectra near the system range need to be excluded. Specifically, the spectral range of 430 to 900 nm within the VIS/NIR spectroscopy, totaling 646 bands, is selected; while the spectral range of 950 to 1650 nm within the NIR spectroscopy, totaling 148 bands, is chosen. The scanning speed of the moving platform within both ranges is set at 2.829 mm/s, with a scan range of 100 mm. Black-and-white correction of hyperspectral images minimizes noise and background interference. A standard white correction plate reflects incident light uniformly. During the white correction stage, the plate is scanned to acquire a complete white calibration image (R_w). In the black correction phase, the camera lens cap is covered and scanned, allowing the system to capture background noise in the absence of light, thus obtaining the all-black calibrated image (R_b). Equation (1) below presents the formula for black and white correction of the original hyperspectral image.

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

where R is the corrected hyperspectral image; R_0 is the original hyperspectral image; R_w is an all-white calibrated image of a standard white calibration plate (reflectance close to 99.9%); and R_b is the all-black calibrated image with the lens cap closed (reflectivity close to 0%).

SpectralView software (*Headwall Photonics, USA*) was used for secondary development with Visual Basic (Version 6.0, Microsoft, USA), which was used to extract the region of interest (ROI) within the hyperspectral image, capture the spectral feature data of the ROI and perform batch processing for averaging. The selection of the ROI in the hyperspectral images of tomato leaves followed the method described by Zhao et al.²⁷ for ROI extraction. The SpectralView software was used to import images and actively extract the reflectivity information based on the coordinate matrix. For each VIS/NIR and NIR hyperspectral image, 15,041-pixel and 617-pixel points, respectively, were extracted. The averaged spectral data (as described in formula 2) derived from this process served as the foundational dataset for subsequent data processing tasks.

$$A_k = \frac{1}{n}(A_1 + A_2 + A_3 + \dots + A_n) = \frac{1}{n} \sum_{i=1}^n A_i \quad (2)$$

Where A_k is the average spectral data of the k -th sample; n is the number of sampling pixel points; and A_i is the spectral data of each sampling pixel point.

Spectral preprocessing and characteristic wavelength extraction

The data were processed and analyzed using MATLAB software (*Mathworks, MATLAB, R2016b*). Before data processing and analysis, 605 samples were randomly permuted using the `randperm` function and subsequently divided into training and prediction sets at a ratio of 3:1.

During spectral acquisition, numerous noise sources, including instrument noise, environmental interference, and background disturbances such as light source leakage and scattering, can compromise the accuracy and reliability of spectral data. Hence, moving average smoothing (MAS) and discrete wavelet transform (DWT) were applied to preprocess the spectral data in this study.

Spectral data comprise numerous bands, and the data dimension is high, potentially causing computational challenges and redundancy issues. Feature extraction techniques aim to preserve essential information, reduce data dimensionality, simplify processing complexity, and enhance classification accuracy. Hence, in this study, PCA, variable combination population analysis (VCPA), and IRIV algorithms were employed to downscale the hyperspectral data.

PCA is a widely utilized technique in multivariate statistical analysis^{28–31}. It serves not only to elucidate the structure and patterns inherent in the data but also to extract its paramount information. Initially, it translates the coordinate system's origin to the data center to eliminate data translational effects. Subsequently, the covariance of the data is computed, and eigenvalue decomposition of the covariance matrix is performed to derive the eigenvalues and corresponding eigenvectors. Based on the eigenvalue magnitude, the K largest eigenvalues and their corresponding eigenvectors are identified as the primary components. Eventually, the original data are projected onto these selected K principal components to reduce the dimensions to K . The foremost feature extracted from the original feature space encapsulates the essential information within the original dataset, thus representing the pivotal feature derived from the original feature set. In this study, the number of principal components was determined using the criterion of an eigenvalue greater than 0.01.

The VCPA is a feature extraction method grounded in model cluster analysis^{32,33}. Initially, variables are sampled via binary matrix sampling, and the variable set with the smallest cross-validation (CV) values is selected as the initial feature. Subsequently, it computes the probability of each wavelength point corresponding to the measured value, and wavelength ranges or variables with lower probabilities are filtered iteratively using a decay function. This process is iterated for retained variables, and the remaining variables are combined to yield the set of characteristic wavelength variables. The maximum feature number of this study was 15, the fold number of cross-validations was 5, the preprocessing method was center, the number of Bayesian model selection (BMS) runs was 1000, the number of exponential decline function (EDF) runs was 85, and the scale factor was 0.1.

The IRIV algorithm is an unsupervised learning method used for data reduction and feature selection^{34,35}. Initially, it normalizes the raw data to have a mean of 0 and a standard deviation of 1 for each feature. Subsequently, it defines an initial feature subset, projects and reconstructs this subset using PCA or other orthogonal decomposition techniques, and calculates the reconstruction error. Features are selected based on their informativeness, and if the chosen subset is not sparse, additional optimization can be performed using L1 regularization or other techniques to increase sparsity. This process iterates until a final, sparsest subset with minimal reconstruction error is obtained. The maximum number of principal components for cross-validation in this test was 10, the number of folds for cross-validation was 5, and the pretreatment method was the center.

Classification models and evaluation criteria

The back propagation neural network (BP) is a widely used fully connected feedforward neural network^{36–38}. It propagates input data from the input layer to the output layer through forward propagation, computing output, and error at each node. Subsequently, backpropagation is employed to pass the error backward through the network, updating weights by calculating gradients between each node and the next layer. This iterative process continues until the error falls within an acceptable range or a preset number of training iterations is reached. In this study, a feed-forward neural network model with "6" hidden nodes was created, the maximum number of

iterations was set to "1000", the target training error was "1e-6", and the learning rate was "0.01", and the learning rate was "0.01".

The SVM is a binary classification model^{39,40}. It determines the classification boundary by identifying an optimal hyperplane in the feature space, with sample points closest to the hyperplane serving as support vectors. Additionally, SVM employs kernel functions to map features to higher-dimensional spaces, enabling it to address linearly inseparable problems. For multiclassification tasks, SVM constructs multiple submodels based on a one-vs-rest or one-vs-one approach⁴¹. This allows the segmentation and categorization of multiple classes by distinguishing each category from all others and assigning samples accordingly. In this study, the kernel function of the model was selected as the "RBF" function, the gamma parameter of the kernel function was "0.01", and the parameter of the penalty factor was set as "10.0".

The radial basis function neural network (RBF) is a single hidden layer feedforward neural network model employing a radial basis function^{42–44}. During training, the K-means clustering algorithm is utilized to categorize samples into k clusters, with each cluster's center serving as an implicit node to enhance model generalizability. Subsequently, the node weights of the output layer are adjusted via an error backpropagation algorithm to minimize the objective function. For this study, the radial basis function's expansion speed was set to "100".

This study employed three evaluation metrics to assess the classification model's effectiveness: accuracy, the macro-averaged F1 (Macro-F1), and the geometric mean (G-mean). All the values fall within the [0,1] range, with higher values indicating better performance. The accuracy represents the proportion of correctly predicted samples to the total number of samples, regardless of the individual category^{45,46}. Macro-F1 computes the average F1 score across all categories, serving as a generalized performance indicator^{47,48}. The G-mean evaluates the model's multiclassification performance, considering the importance of each category^{49,50}. The corresponding formulas are presented below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

$$Macro - F1 = 2 \times \frac{Pr\ ecision_{ma} \times Recall_{ma}}{Pr\ ecision_{ma} + Recall_{ma}} = \frac{1}{n} \sum_{i=1}^n F1_i \quad (4)$$

$$Micro - F1 = 2 \times \frac{Pr\ ecision_{mi} \times Recall_{mi}}{Pr\ ecision_{mi} + Recall_{mi}} \quad (5)$$

$$G-mean = \sqrt{Macro - F1 \times Micro - F1} \quad (6)$$

where *TP* denotes the successful prediction of a positive sample as positive; *TN* denotes the successful prediction of a negative sample as negative; *FP* denotes the incorrect prediction of a negative sample as positive; *FN* indicates that a positive sample is incorrectly predicted as negative; *n* indicates the number of categories; and *Micro-F1* indicates the micro-averaged F1 score of the sample; *Precision* indicates the proportion of successful predicted positive samples among all samples predicted as positive; *Recall* indicates the proportion of successful predicted positive samples among all actual positive samples.

Data fusion

Spectral data fusion integrates data from various spectral bands to enhance the comprehensiveness and accuracy of information. This approach mitigates the limitations of individual bands and enhances spectral resolution. The study employs low-level and medium-level data fusion techniques.

Low-level data fusion directly integrates raw data. It combines VIS/NIR and NIR spectra to create a new dataset. Each row in the dataset represents the spectral data for one sample, including both VIS/NIR and NIR wavelengths. Each column contains spectral values for these wavelengths.

Medium-level data fusion is the fusion of data after feature extraction. The method extracts more discriminative and interpretable features while preserving most of the original data information. The study employed multiple fusion techniques by integrating the VIS/NIR and NIR sample data extracted using PCA, VCPA, and IRIV feature extraction methods, respectively, to create new datasets. This fusion process aimed to enrich the information content and enhance the discriminative capability of the fused datasets.

Confusion matrix

The confusion matrix is a common method for evaluating the performance of a classification model, depicting the relationship between the true attributes of the sample data and the model's recognition results. It uses a matrix to represent the correspondence between true class and prediction class, helping us intuitively understand the model's performance across different categories.

Results

Early spectral characteristics of *Cladosporium fulvum* in tomatoes

Figure 2 shows plots of the raw spectral data, preprocessed spectral data, and averaged spectra for each period of the 605 samples in the VIS/NIR and NIR regions. The raw spectral data exhibit some degree of fluctuation, possibly due to environmental influences, instrumental errors, or other interfering factors, complicating the data analysis. A comparison of spectral maps before and after preprocessing reveals that employing the MAS-DWT

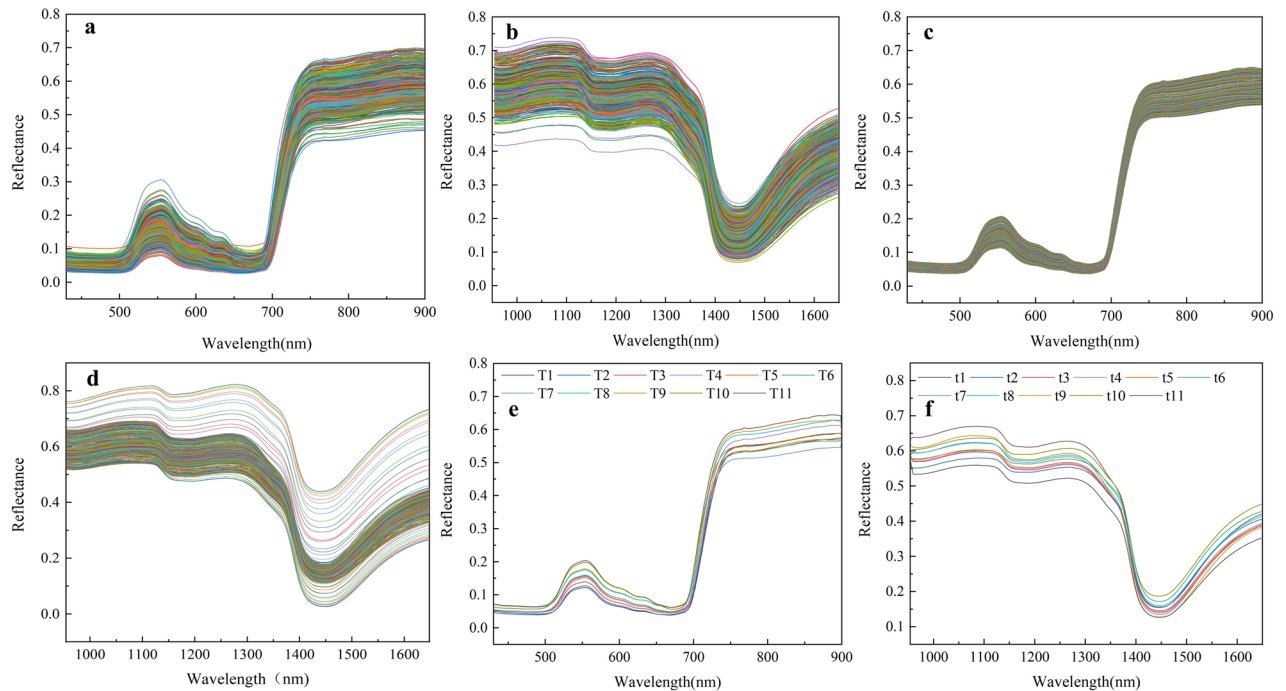


Fig. 2. Raw spectrograms, preprocessed spectrograms, and average spectrograms for each period of the sample's VIS/NIR and NIR; **a** Raw spectrogram in the VIS/NIR; **b** Raw spectrogram in the NIR; **c** VIS/NIR spectrogram of the sample after preprocessing; **d** NIR spectrogram of the sample after preprocessing; **e** Average VIS/NIR spectrograms for each period; **f** Average NIR spectrograms for each period.

preprocessing technique effectively diminishes data fluctuations, eliminating high-frequency noise and clutter signals, thereby enhancing the signal-to-noise ratio of the spectral data.

In the VIS/NIR spectral range, the spectral reflectance curves of the tomato leaves exhibited a consistent trend throughout the different periods. The spectral reflectance changes slowly before the wavelength 700 nm, followed by a sharp increase near 700 nm. There are two absorption peaks near the 510 nm and 690 nm wavelengths and a reflection peak at 550 nm. In the wavelength range below 700 nm, leaf spectral reflectance increases with infection duration. Similarly, in the NIR spectral range, the spectral reflectance curves of tomato leaves demonstrated a consistent pattern across different times. The reflectance slowly changes before 1300 nm, sharply decreases after 1300 nm, and subsequently increases after 1450 nm. Reflection peaks are observed near 1120 nm and 1270 nm, with absorption peaks near 1160 nm and 1440 nm.

Feature extraction results

After preprocessing, there are 646 spectral bands in the VIS/NIR range and 148 bands in the NIR range, totaling 794 bands suitable for low-level data fusion. Hence, in this study, PCA, VCPA, and IRIV algorithms are employed to reduce the dimensionality of the spectral data in the VIS/NIR, NIR, and low-level fusion, preserving critical feature variables to enhance modeling efficiency and accuracy. For the VIS/NIR spectral data, PCA, VCPA, and IRIV extracted 13, 14, and 20 feature variables, respectively. Similarly, in the NIR spectral data, PCA, VCPA, and IRIV extracted 6, 12, and 30 feature variables, respectively. Furthermore, PCA, VCPA, and IRIV extracted 14, 11, and 28 feature variables from the low-level fusion spectral data, respectively. The specific feature variables are shown in Fig. 3.

Classification results for a single data block

For the VIS/NIR spectral data, the 605-sample data from periods T1 to T11 were randomly split into calibration and prediction sets at a 3:1 ratio, resulting in 12 classification models including BP, PCA-BP, VCPA-BP, IRIV-BP, SVM, PCA-SVM, VCPA-SVM, IRIV-SVM, RBF, PCA-RBF, VCPA-RBF, and IRIV-RBF, as detailed in Table 1. Comparative analysis reveals that the RBF classification model performs the best, followed by BP, while the SVM model exhibits the lowest performance. Feature extraction enhances the performance of all models, with the PCA-RBF model demonstrating the highest efficacy. Specifically, the PCA-RBF model achieves accuracies of 99.8% and 98% for the calibration and prediction sets, respectively, with both Macro-F1 values up to 1 and both G-mean values up to 1, respectively. The evaluation metrics for the calibration set match those of the RBF model, while for the prediction set, the accuracy, Macro-F1, and G-mean values are 2.6%, 0.1, and 0.1 higher than those of the RBF model, respectively, with 635 fewer features than the RBF model.

For the NIR spectral data, the 605-sample dataset spanning periods t1 to t11 underwent similar processing. Comparative analysis indicates that the RBF classification model performs the best, followed by BP, with the SVM model displaying the poorest performance. Except for the IRIV-SVM and PCA-RBF models, all models exhibit improved performance after spectral data feature extraction, with the IRIV-RBF model delivering the

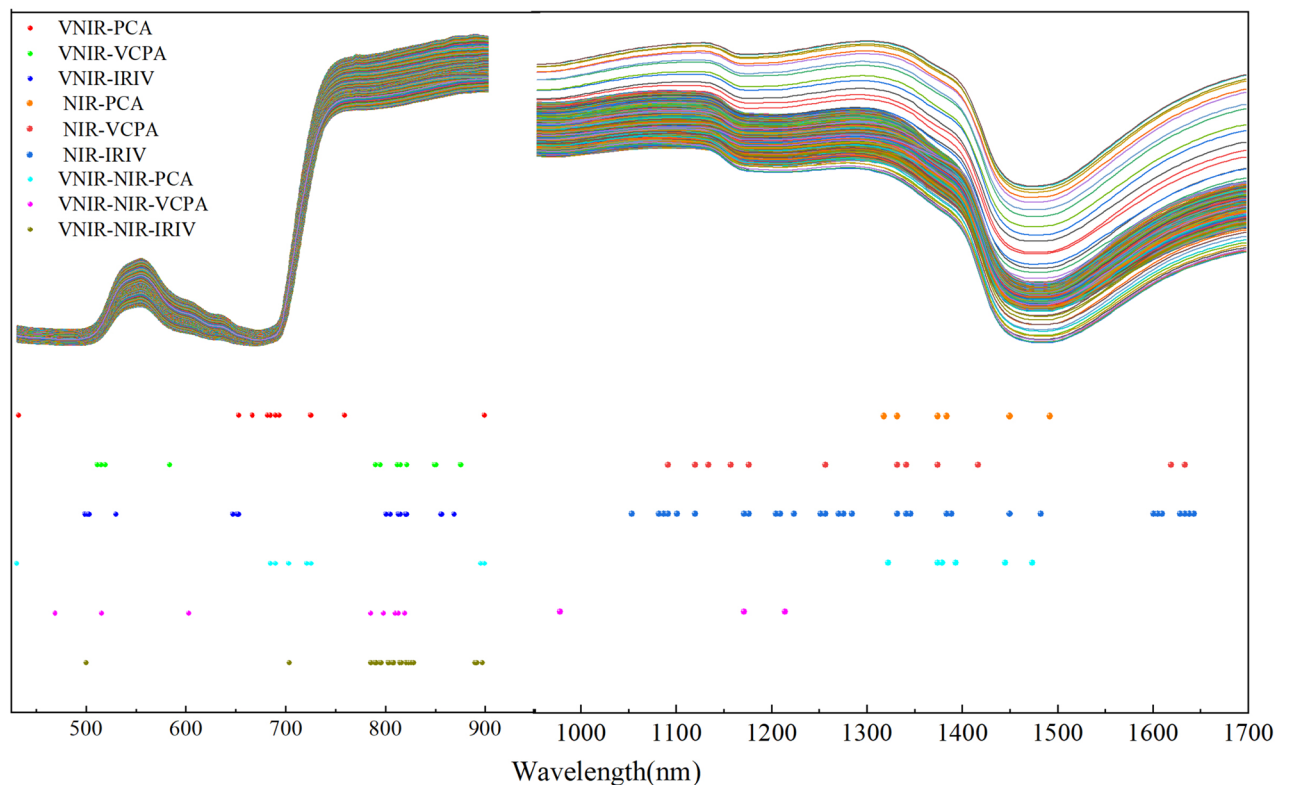


Fig. 3. Feature extraction for visible/near-infrared, and near-infrared spectral data and low-level data fusion based on the principal component analysis, variable combination population analysis, and iterative retention of informative variables algorithms.

most favorable outcomes. Specifically, the IRIV-RBF model achieves accuracy rates of 98.9% and 98.7% for the calibration and prediction sets, respectively, with both Macro-F1 values up to 1, and both G-mean values up to 1, respectively, and the evaluation metric values of the correction set were the same as the evaluation metric values of the RBF model. The accuracy of the prediction set is improved by 2.6%, compared with those of the RBF model, and the number of features is 118 fewer than the number of features of the RBF model.

Classification results of low-level data fusion

In this study, a 794-dimensional spectral dataset was obtained by directly merging the VIS/NIR and NIR spectral data through low-level data fusion. Concurrently, feature dimensionality reduction using PCA, VCPA, and IRIV was applied to the low-level fusion spectral data, followed by the establishment of classification models such as BP, SVM, and RBF for both the low-level fusion spectral data and the spectral data after feature extraction, as detailed in Table 2. A comparison of Tables 1 and 2 shows that the performances of the BP class and RBF class models constructed using low-level data fusion surpass those of models constructed using a single data block. Specifically, in the PCA-RBF model, the accuracy, Macro-F1, and G-mean of the calibration set reach 100%, 1, and 1, respectively. For the prediction set, the accuracy, Macro-F1, and G-mean reach 99.3%, 1, and 1, respectively, indicating excellent performance. Additionally, SVM class models constructed using low-level data fusion outperform those constructed using NIR data. Notably, only the SVM, PCA-SVM, and VCPA-SVM models constructed using low-level data fusion are superior to their counterparts constructed using VIS/NIR data.

Classification results for medium-level data fusion

The characteristic wavelength data extracted from PCA, VCPA, and IRIV in the VIS/NIR and NIR bands were combined using a medium-level data fusion method, resulting in 17-, 25-, and 50-dimensional spectral data, as presented in Table 3. This approach combines feature extraction and data fusion techniques to accurately capture correlations between VIS/NIR and NIR data, thereby providing more precise and valuable information. Comparing Tables 1 and 3, it is evident that the models constructed based on medium-level data fusion outperform those built based on a single block of data in all cases. Specifically, the accuracy, Macro-F1, and G-Mean values in the RBF classification model, which is based on the medium-level fusion of PCA-extracted feature data, reach 100%, 1, and 1 for the calibration set. For the prediction set, the accuracy, Macro-F1, and G-Mean values are 99.3%, 1, and 1, respectively.

Comparing Tables 2 and 3, it can be observed that the SVM class, VCPA-RBF, and IRIV-RBF models constructed using medium-level data fusion, outperform their counterparts built on low-level data fusion. However, the BP class models constructed based on medium-level data fusion do not perform as well as their low-level

Spectrum	Model	NFS	Calibration			Prediction		
			Accuracy	Macro-F1	G-mean	Accuracy	Macro-F1	G-mean
VIS/NIR	BP	646	72	1	0.8	70.4	1	0.8
	PCA-BP	11	91.4	1	0.9	84.9	1	0.9
	VCPA-BP	13	88.5	1	0.9	87.5	1	0.9
	IRIV-BP	20	88.1	1	0.9	82.9	1	0.9
	SVM	646	69.1	0.7	0.7	61.8	0.6	0.6
	PCA-SVM	11	87.6	0.9	0.9	87.5	0.9	0.9
	VCPA-SVM	13	66.9	0.7	0.7	64.5	0.6	0.6
	IRIV-SVM	20	72	0.7	0.7	65.8	0.7	0.7
	RBF	646	99.8	1	1	95.4	0.9	0.9
	PCA-RBF	11	99.8	1	1	98.0	1	1
	VCPA-RBF	13	98.2	1	1	97.4	1	1
	IRIV-RBF	20	98.7	1	1	98	1	1
NIR	BP	148	66.9	0.8	0.7	63.2	0.7	0.7
	PCA-BP	6	74.2	0.7	0.7	71.7	0.5	0.6
	VCPA-BP	12	81.5	1	0.9	78.9	1	0.9
	IRIV-BP	30	81.5	1	0.9	73.7	0.9	0.8
	SVM	148	32.9	0.3	0.3	28.3	0.3	0.3
	PCA-SVM	6	53	0.5	0.5	52	0.5	0.5
	VCPA-SVM	12	42.4	0.4	0.4	31.6	0.3	0.3
	IRIV-SVM	30	30.0	0.3	0.3	28.3	0.3	0.3
	RBF	148	98.9	1	1	96.1	1	1
	PCA-RBF	6	95.4	1	1	94.1	1	1
	VCPA-RBF	12	97.4	1	1	97.4	1	1
	IRIV-RBF	30	98.9	1	1	98.7	1	1

Table 1. Evaluation metrics for BP, SVM, and RBF classification models before and after feature extraction with PCA, VCPA, and IRIV based on individual data blocks. NFS: Number of key spectral.

Model	NFS	Calibration			Prediction		
		Accuracy	Macro-F1	G-mean	Accuracy	Macro-F1	G-mean
BP	794	87.2	1	0.9	82.9	0.9	0.9
PCA-BP	14	96.7	1	1	94.7	1	1
VCPA-BP	11	89.6	1	0.9	88.2	1	0.9
IRIV-BP	28	93.6	1	1	92.1	1	1
SVM	794	70	0.7	0.7	65.1	0.7	0.7
PCA-SVM	14	90.5	0.9	0.9	90.1	0.9	0.9
VCPA-SVM	11	73.1	0.7	0.7	66.4	0.7	0.7
IRIV-SVM	28	57.8	0.6	0.6	56.6	0.6	0.6
RBF	794	99.8	1	1	97.4	1	1
PCA-RBF	14	100	1	1	99.3	1	1
VCPA-RBF	11	98	1	1	98.0	1	1
IRIV-RBF	28	99.1	1	1	98.7	1	1

Table 2. Evaluation metrics for BP, SVM, and RBF classification models before and after feature extraction with PCA, VCPA, and IRIV based on low-level fusion of data.

counterparts. Although the PCA-RBF model constructed using medium-level data fusion performs comparably to the model constructed using low-level data fusion, the latter model has fewer features. Overall, the PCA-RBF classification model constructed using low-level data fusion yields the best results with a required number of feature spectra of 14.

Confusion matrix of optimal model

The confusion matrix plots for the calibration set and prediction set of the PCA-RBF classification model based on low-level data fusion are shown in Fig. 4. From the figure, it can be seen that in the calibration set, the model correctly classifies all samples without any errors. In the prediction set, the model also correctly classifies the

Model	NFS	Calibration			Prediction			
		Accuracy	Macro-F1	G-mean	Accuracy	Macro-F1	G-mean	
PCA	BP	17	90.5	1	1	86.8	1	0.9
	SVM	17	95.4	1	1	94.1	0.9	0.9
	RBF	17	100	1	1	99.3	1	1
VCPA	BP	25	88.5	1	0.9	86.8	1	0.9
	SVM	25	71.3	0.7	0.7	69.1	0.7	0.7
	RBF	25	99.3	1	1	98.0	1	1
IRIV	BP	50	91.6	0.9	0.9	90.8	0.8	0.9
	SVM	50	72.6	0.7	0.7	67.1	0.7	0.7
	RBF	50	99.8	1	1	99.3	1	1

Table 3. Evaluation metrics for BP, SVM, and RBF classification models based on medium-level data fusion.

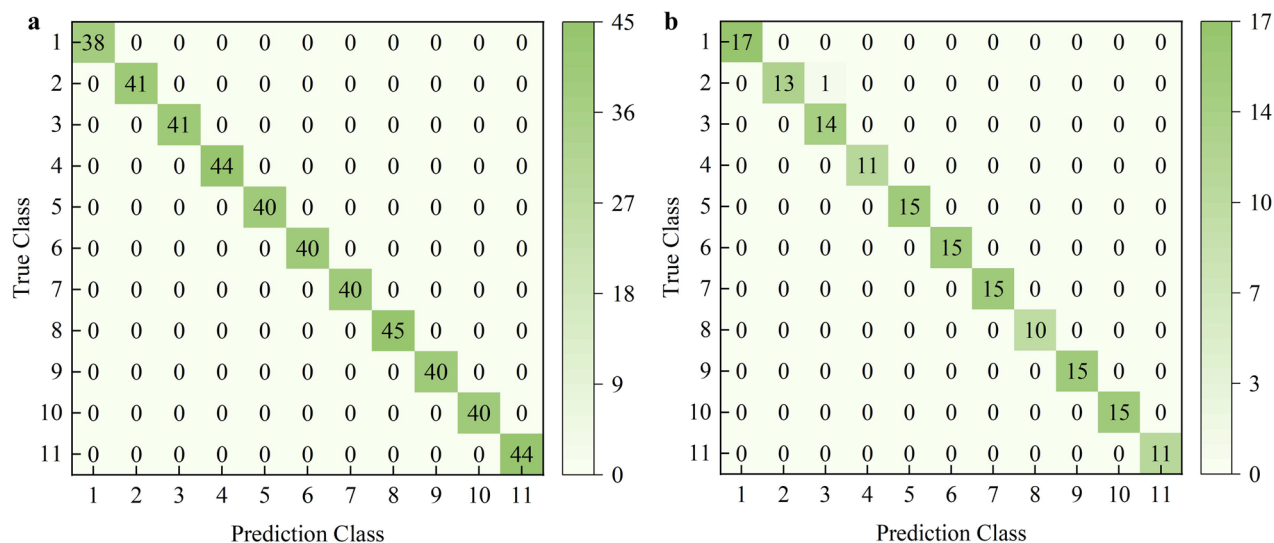


Fig. 4. Confusion matrices of the calibration set and prediction set for the PCA-RBF classification model based on low-level data fusion; (a) Confusion matrix of the calibration set; (b) Confusion matrix of the prediction set.

vast majority of samples, with only one exception: one sample, which is actually of class 2, is incorrectly classified as class 3. Despite this minor error, the model's overall performance on the prediction set remains excellent.

Discussion

The spectral reflectance of tomato leaves is primarily influenced by the content of photosynthetic molecules, the cellular structure, the water content, and diseases⁵¹. The main photosynthetic molecules included chlorophyll and carotenoids. Chlorophyll and carotenoid molecules in the leaves strongly absorb visible (VIS) light, resulting in very low reflectance of tomato leaves in the VIS spectral range. In contrast, in the NIR spectral range, the reflectance of leaves is influenced by the cellular structure and moisture content. Structures such as cell walls, stomata, and chloroplasts scatter NIR light, leading to greater reflectance of leaves in this range.

There are absorption peaks near 510 nm and 690 nm, and a reflection peak at 550 nm. This is primarily due to the influence of chlorophyll, which has a high absorption rate for blue and red light and a low reflectance rate for green light. Reflection peaks are observed near 1120 nm and 1270 nm, mainly due to the effect of leaf structure, while absorption peaks near 1160 nm and 1440 nm are primarily attributed to water absorption⁵².

When inoculated with *C. fulvum* spore suspension, spores attach to the surface of the leaves, and when the stomata of the leaves are opened, these spores change into the interior of the leaves and form sclerotium inside the leaves⁵³. With the growth of *C. fulvum*, the chlorophyll molecules in the leaves are damaged, leading to increased reflectance of the leaves in the VIS spectrum below 700 nm. With increasing invasion time, the reflectivity of the leaves continued to increase. Additionally, since *C. fulvum* invades leaf cells and secretes enzymes like cellulase, hemicellulase and pectinase that degrade the plant cell walls⁵⁴, the cell structure is damaged. These enzymes also disrupt the cell membrane⁵⁵, making it difficult for the leaf to maintain internal water. Once the pathogen penetrates the cell membrane, important components like proteins and nucleic acids in the cytoplasm are also damaged, affecting DNA stability and gene expression⁵⁶, which leads to cell death. The pathogen also causes the rupture of vacuoles⁵⁷, which are crucial for water storage and ion balance regulation⁵⁸. The destruction of vacuoles results in water loss and affects water regulation. Since leaf reflectance in the NIR range is related to

factors such as cell structure and water content⁵⁹, the combined effects of these factors result in irregular spectral reflectance patterns.

The RBF classification model achieved the best performance among the models built on a single block of data. The accuracies of the calibration and prediction sets for the VIS/NIR spectral data model reached 99.8% and 95.4%, respectively. For the NIR spectral data model, the accuracies of the calibration and prediction sets reached 98.9% and 96.1%, respectively.

PCA, VCPA, and IRIV methods were employed to select characteristic wavelengths for VIS/NIR and NIR data. The model constructed from the extracted features performed significantly better than the original model, indicating the beneficial impact of feature extraction on enhancing the model's classification accuracy. In the case of the RBF model, the accuracy of the prediction set increased by 2.6%, 2%, and 2.6% in the VIS/NIR spectral data, respectively. However, in the NIR spectral data, the PCA feature extraction method led to a degradation in the prediction performance of the original classification model. This could be attributed to the limited number of feature wavelengths extracted by PCA, which failed to fully capture the essential information in the original spectral data. In contrast, the other two methods improved the accuracy of the prediction set of the model by 1.3% and 2.6%, respectively.

Comparing the classification models constructed through low-level data fusion with those built using individual data blocks, it is evident that, apart from the IRIV-SVM model, all other models derived from low-level data fusion outperform those established from individual data blocks. This indicates that low-level fusion models exhibit greater potential for the early diagnosis of *C. fulvum* in tomatoes than single data models. In particular, the PCA-RBF model constructed based on low-level data fusion yields the most favorable outcomes, achieving accuracies of 100% and 99.3% for the calibration and prediction sets, respectively, with a Macro-F1 of 1, a G-mean of 1, respectively, and 14 feature wavelengths.

Comparing the classification model constructed through the medium-level fusion of data with the model built from a single data block, it is evident that the performance of the former surpasses that of the latter, indicating that the medium-level fusion model holds greater promise for early diagnosis of *C. fulvum* in tomatoes compared to a single data model. Specifically, the RBF class classification model constructed based on the medium-level fusion of PCA-extracted feature data yields the most favorable outcomes, achieving an accuracy, Macro-F1, and G-mean of 100%, 1, and 1, respectively, for the calibration set. For the prediction set, the corresponding values are 99.3%, 1, and 1, respectively, with the model utilizing 17 feature wavelengths.

A comparison between the classification models constructed via medium-level fusion and low-level fusion of the data indicates that the SVM class, VCPA-RBF, and IRIV-RBF models obtained through medium-level fusion outperform their low-level fusion counterparts, whereas the BP class models exhibit better performance in the low-level fusion scenario. This highlights the importance of selecting appropriate fusion strategies based on different models and problem domains. For the PCA-RBF model, both fusion strategies yield similar performances; however, the low-level fusion model has fewer feature spectra than the medium-level fusion model. The 14 feature wavelengths extracted for the PCA-RBF model via low-level fusion are 430.538, 684.918, 690.006, 703.088, 721.258, 725.619, 895.69, 899.324, 1322.39, 1374.25, 1378.97, 1393.11, 1444.97, and 1473.26 nm. The spectral reflectances at 430.538, 684.918, and 690.006 nm are primarily correlated with the chlorophyll and carotenoid contents in the leaves, whereas the reflectances at 703.088, 721.258, 725.619, 895.69, and 899.324 nm are associated with cellular structures such as the cell wall, stomata, and chloroplasts and water content⁶⁰. The reflectances at 1322.39, 1374.25, 1378.97, 1393.11, 1444.97, and 1473.26 nm are mainly influenced by the water content in the cells⁶¹.

Conclusions

This study investigated the use of VIS/NIR and NIR hyperspectral imaging combined with data fusion for the early diagnosis of *C. fulvum* in greenhouse tomatoes using both healthy and inoculated diseased tomato leaves. The study compared the effects of different feature extraction methods and different machine learning models on the early diagnosis of *C. fulvum* in tomatoes. Low-level and medium-level data fusion were also performed on spectral data in the VIS/NIR and NIR to compare the effects of individual data blocks, low-level fusion, and medium-level fusion on the performance of the model for the early diagnosis of *C. fulvum*. Comparisons revealed that:

- The RBF classification model was the most effective for early diagnosis, followed by the BP model, with the SVM model performing the worst.
- The PCA-RBF classification model achieved the best results for models built on VIS/NIR spectral data, while the IRIV-RBF classification model achieved the best results for models built on NIR spectral data.
- Models built based on low-level data fusion outperformed those built based on NIR spectral data. The BP class, RBF class, SVM class, PCA-SVM, and VCPA-SVM models built based on low-level data fusion outperformed those built based on VIS/NIR data.
- Modeling based on medium-level data fusion consistently outperformed modeling based on individual data blocks.
- The SVM class, VCPA-RBF, and IRIV-RBF models built based on medium-level data fusion all performed better than their low-level fusion counterparts, whereas the BP class models showed the opposite trend. The PCA-RBF model built based on medium-level data fusion yielded similar performance to that built on low-level data fusion, but the PCA-RBF model based on low-level data fusion required fewer features.

In conclusion, the study demonstrated that the PCA-RBF classification model established using hyperspectral imaging combined with low-level data fusion can effectively enable the early diagnosis of *C. fulvum* in greenhouse

tomatoes, providing a fast and nondestructive approach for early disease detection, with implications for the early detection of other plant diseases. For hyperspectral images, besides spectral information, spatial information also contains important features. The neglect of spatial information in the study might lead to less precise disease diagnosis. Therefore, future research could focus on integrating both spectral and spatial information from samples to enhance the accuracy and precision of disease diagnosis. Additionally, corresponding preventive measures can be implemented for plants at different stages of disease infection to control disease spread and minimize crop losses.

Data availability

Available upon request from the corresponding author.

Received: 12 May 2024; Accepted: 26 August 2024

Published online: 30 August 2024

References

- Nazarov, P. A., Baleev, D. N., Ivanova, M. I., Sokolova, L. M. & Karakozova, M. V. Infectious plant diseases: etiology, current status, problems and prospects in plant protection. *J. Acta Nat.* **12**, 46 (2020).
- Savary, S. *et al.* The global burden of pathogens and pests on major food crops. *J. Nat. Ecol. Evolut.* **3**, 430–439 (2019).
- Abdulridha, J., Batuman, O. & Ampatzidis, Y. UAV-based remote sensing technique to detect citrus canker disease utilizing hyperspectral imaging and machine learning. *J. Remote Sens.* **11**, 1373 (2019).
- Li, L., Zhang, S. & Wang, B. Plant disease detection and classification by deep learning—a review. *J. IEEE Access.* **9**, 56683–56698 (2021).
- Nguyen, C. *et al.* Early detection of plant viral disease using hyperspectral imaging and deep learning. *J. Sens.* **21**, 742 (2021).
- Nouri, N. M., Abbood, H. M., Riahi, M. & Alagheband, S. H. A review of technological developments in modern farming: Intelligent greenhouse systems. *AIP Conf. Proc. AIP Publ.* **2631**, 1 (2023).
- Magalhães, S. A. *et al.* Evaluating the single-shot multibox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse. *J. Sens.* **21**, 3569 (2021).
- Zhang, S., Griffiths, J. S., Marchand, G., Bernards, M. A. & Wang, A. Tomato brown rugose fruit virus: An emerging and rapidly spreading plant RNA virus that threatens tomato production worldwide. *J. Mol. Plant Pathol.* **23**, 1262–1277 (2022).
- Iida, Y. *et al.* Evaluation of the potential biocontrol activity of *Dicyma pulvinata* against *Cladosporium fulvum*, the causal agent of tomato leaf mould. *J. Plant Pathol.* **67**, 1883–1890 (2018).
- Wang, Y. Y., Yin, Q. S., Qu, Y., Li, G. Z. & Hao, L. Arbuscular mycorrhiza-mediated resistance in tomato against *Cladosporium fulvum*-induced mould disease. *J. Phytopathol.* **166**, 67–74 (2018).
- Zahir, S. A. D. M., Omar, A. F., Jamlos, M. F., Azmi, M. A. M. & Muncan, J. A review of visible and near-infrared (Vis-NIR) spectroscopy application in plant stress detection. *J. Sens. Actuat. A: Phys.* **338**, 113468 (2022).
- Sankaran, S., Mishra, A., Ehsani, R. & Davis, C. A review of advanced techniques for detecting plant diseases. *J. Comput. Electron. Agricult.* **72**, 1–13 (2010).
- Martinelli, F. *et al.* Advanced methods of plant disease detection. *A Rev. Agronomy Sustain. Dev.* **35**, 1–25 (2015).
- Terentev, A., Dolzhenko, V., Fedotov, A. & Eremenko, D. Current state of hyperspectral remote sensing for early plant disease detection: a review. *J. Sens.* **22**, 757 (2022).
- Wan, L. *et al.* Hyperspectral sensing of plant diseases: principle and methods. *J. Agronomy* **12**, 1451 (2022).
- Feng, Z. H. *et al.* Hyperspectral monitoring of powdery mildew disease severity in wheat based on machine learning. *J. Front. Plant Sci.* **13**, 828454 (2022).
- Zhang, N. *et al.* A review of advanced technologies and development for hyperspectral-based plant disease detection in the past three decades. *J. Remote Sens.* **12**, 3188 (2020).
- Lu, B., Dao, P. D., Liu, J., He, Y. & Shang, J. Recent advances of hyperspectral imaging technology and applications in agriculture. *J. Remote Sens.* **12**, 2659 (2020).
- Huang, Y., Wang, D., Liu, Y., Zhou, H. & Sun, Y. Measurement of early disease blueberries based on vis/nir hyperspectral imaging system. *J. Sensors.* **20**, 5783 (2020).
- Ugarte Fajardo, J. *et al.* Early detection of black Sigatoka in banana leaves using hyperspectral images. *J. Appl. plant Sci.* **8**, e11383 (2020).
- Feng, L. *et al.* Investigation on data fusion of multisource spectral data for rice leaf diseases identification using machine learning methods. *J. Front. Plant Sci.* **11**, 5770636 (2020).
- Xiao, Q., Bai, X., Gao, P. & He, Y. Application of convolutional neural network-based feature extraction and data fusion for geographical origin identification of radix astragali by visible/short-wave near-infrared and near infrared hyperspectral imaging. *J. Sens.* **20**, 4940 (2020).
- Yu, H. D. *et al.* Hyperspectral imaging in combination with data fusion for rapid evaluation of tilapia fillet freshness. *J. Food Chem.* **348**, 129129 (2021).
- Khaled, A. Y. *et al.* Early detection of diseases in plant tissue using spectroscopy—applications and limitations. *J. Appl. Spectroscopy Rev.* **53**, 36–64 (2018).
- Zhang, X., Wang, Y., Zhou, Z., Zhang, Y. & Wang, X. Detection method for tomato leaf mildew based on hyperspectral fusion terahertz technology. *J. Foods* **12**, 535 (2023).
- Babadoost, M. Leaf mold (*Fulvia fulva*), a serious threat to high tunnel tomato production in Illinois. *In III Int. Symposium on Tomato Dis.* **914**, 93–96 (2010).
- Zhao, J. *et al.* Simultaneous quantification and visualization of photosynthetic pigments in *Lycopersicon esculentum* Mill under different levels of nitrogen application with Visible-Near Infrared Hyperspectral Imaging Technology. *J. Plants.* **12**, 2956 (2023).
- Hasan, B. M. & Abdulazeez, A. M. A review of principal component analysis algorithm for dimensionality reduction. *J. Soft Comput. Data Mining.* **2**(1), 20–30 (2021).
- Zhao, H., Zheng, J., Xu, J. & Deng, W. Fault diagnosis method based on principal component analysis and broad learning system. *J. IEEE Access* **7**, 99263–99272 (2019).
- Zhang, D., Zou, L., Zhou, X. & He, F. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *J. Ieee Access* **6**, 28936–28944 (2018).
- Jafarzadegan, M., Safi-Esfahani, F. & Beheshti, Z. Combining hierarchical clustering approaches using the PCA method. *J. Expert Syst. Appl.* **137**, 1–10 (2019).
- Fan, Y., Zhang, C., Liu, Z., Qiu, Z. & He, Y. Cost-sensitive stacked sparse auto-encoder models to detect striped stem borer infestation on rice based on hyperspectral imaging. *J. Knowledge-Based Syst.* **168**, 49–58 (2019).
- Wan, G. *et al.* Feature wavelength selection and model development for rapid determination of myoglobin content in nitrite-cured mutton using hyperspectral imaging. *J. Food Eng.* **287**, 110090 (2020).

34. Wei, L., Yuan, Z., Yu, M., Huang, C. & Cao, L. Estimation of arsenic content in soil based on laboratory and field reflectance spectroscopy. *J. Sens.* **19**, 3904 (2019).
35. Yin, C. *et al.* Method for detecting the pollution degree of naturally contaminated insulator based on hyperspectral characteristics. *J. High Voltage* **6**, 1031–1039 (2021).
36. Fan, B. *et al.* Evaluation of mutton adulteration under the effect of mutton flavour essence using hyperspectral imaging combined with machine learning and sparrow search algorithm. *J. Foods* **11**, 2278 (2022).
37. Liu, C. *et al.* A discriminative model for early detection of anthracnose in strawberry plants based on hyperspectral imaging technology. *J. Remote Sens.* **15**, 4640 (2023).
38. Wang, H., Zhu, H., Zhao, Z., Zhao, Y. & Wang, J. The study on increasing the identification accuracy of waxed apples by hyperspectral imaging technology. *J. Multimedia Tools Appl.* **77**, 27505–27516 (2018).
39. Chen, Y. N., Thaipisutikul, T., Han, C. C., Liu, T. J. & Fan, K. C. Feature line embedding based on support vector machine for hyperspectral image classification. *J. Remote Sens.* **13**, 130 (2021).
40. Guo, Y., Yin, X., Zhao, X., Yang, D. & Bai, Y. Hyperspectral image classification with SVM and guided filter. *J. EURASIP J. Wireless Commun. Netw.* **2019**, 1–9 (2019).
41. Ding, S., Zhao, X., Zhang, J., Zhang, X. & Xue, Y. A review on multi-class TWSVM. *J. Artif. Intell. Rev.* **52**, 775–801 (2019).
42. Deng, Y. *et al.* New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloalketones in tap water. *J. Sci. Total Environ.* **772**, 145534 (2021).
43. He, Q. *et al.* Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. *J. Sci. Total Environ.* **663**, 1–15 (2019).
44. Li, X. & Sun, Y. Application of RBF neural network optimal segmentation algorithm in credit rating. *J. Neural Comput. Appl.* **33**, 8227–8235 (2021).
45. Hassan, S. M., Maji, A. K., Jasiński, M., Leonowicz, Z. & Jasińska, E. Identification of plant-leaf diseases using CNN and transfer-learning approach. *J. Electron.* **10**, 1388 (2021).
46. Kaya, A. *et al.* Analysis of transfer learning for deep neural network based plant classification models. *J. Comput. Electron. Agricult.* **158**, 20–29 (2019).
47. Karayığit, H., Acı, Ç. İ & Akdağlı, A. Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods. *J. Exp. Syst. Appl.* **174**, 114802 (2021).
48. Takahashi, K., Yamamoto, K., Kuchiba, A. & Koyama, T. Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores. *J. Appl. Intell.* **52**, 4961–4972 (2022).
49. Bi, J. & Zhang, C. An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *J. Knowledge-Based Syst.* **158**, 81–93 (2018).
50. Karim, F., Majumdar, S., Darabi, H. & Harford, S. Multivariate LSTM-FCNs for time series classification. *J. Neural Netw.* **116**, 237–245 (2019).
51. Liu, L. Y. & Huang, W. J. Detection of internal leaf structure deterioration using a new spectral ratio index in the near-infrared shoulder region. *J. Integrat. Agricult.* **13**, 760–769 (2014).
52. De Wit, P. J. *et al.* The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *J. PLoS Genet.* **8**, e1003088 (2012).
53. Ökmen, B. Identification and characterization of novel effectors of *Cladosporium fulvum* in *Wageningen University and Research* (2013).
54. Kubicek, C. P., Starr, T. L. & Glass, N. L. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. *J. Ann. Rev. Phytopathol.* **52**, 427–451 (2014).
55. Dodds, P. N. *et al.* Effectors of biotrophic fungi and oomycetes: pathogenicity factors and triggers of host resistance. *J. New Phytologist* **183**, 993–1000 (2009).
56. Lartey, R., & Citovsky, V. Nucleic acid transport in plant-pathogen interactions. *J. Genetic Eng.: Principles and Methods.* 201–214(1997).
57. Gross, P., Julius, C., Schmelzer, E. & Hahlbrock, K. Translocation of cytoplasm and nucleus to fungal penetration sites is associated with depolymerization of microtubules and defence gene activation in infected, cultured parsley cells. *EMBO J.* **12**, 1735–1744 (1993).
58. WINK, M. The plant vacuole: a multifunctional compartment. *J. Exp. Botany.* 231–246(1993).
59. Slaton, M. R. Estimating near-infrared leaf reflectance from leaf structural characteristics. *Am. J. Botany* **88**, 278–284 (2001).
60. Marin-Ortiz, J. C., Gutierrez-Toro, N., Botero-Fernández, V. & Hoyos-Carvajal, L. M. Linking physiological parameters with visible/near-infrared leaf reflectance in the incubation period of vascular wilt disease. *Saudi J. Biol. Sci.* **27**, 88–699 (2020).
61. Xu, H. R., Ying, Y. B., Fu, X. P. & Zhu, S. P. Near-infrared spectroscopy in detecting leaf miner damage on tomato leaf. *J. Biosyst. Eng.* **96**, 447–454 (2007).

Acknowledgements

The cultivation of *Cladosporium fulvum* and preparation of spore suspension were completed by a teacher and a student from the College of Plant Protection, Shanxi Agricultural University. Here we sincerely thank Professor M.W. for assisting.

Author contributions

X.Z.: Conceptualization, Methodology, Investigation, Writing—original draft, Writing-review and editing. Y.L.: Methodology, Investigation. Z.H.: Investigation. G.L.: Investigation. Z.Z.: Investigation. X.H.: Investigation. H.D.: Investigation. M.W.: Writing-review and editing. Z.L.: Writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by the Major Special Projects of National Key R&D, grant number 2021YFD1600301-4; Major Special Projects of National Key R&D, grant number 2017YFD0701501; Major Special Projects of Shanxi Province Key R&D, grant number 201903D211005; Central Government Guides Local Funds for Scientific and Technological Development, grant number YDZJSX20231A042; Construction Project of Shanxi Modern Agricultural Industry Technology System, grant number 2024CYJSTX08-18; Major Projects of Shanxi Province Key R&D, grant number 2022ZDYF119.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.W. or Z.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024