Hindawi
Journal of Spectroscopy
Volume 2019, Article ID 9753829, 15 pages
https://doi.org/10.1155/2019/9753829

Research Article

Identification of *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici* Based on Near-Infrared Spectroscopy

Yaqiong Zhao,[1] Feng Qin,[1] Fei Xu,[2] Jinxing Ma,[1] Zhenyu Sun,[3] Yuli Song,[2] Longlian Zhao,[4] Junhui Li,[4] and Haiguang Wang[1]

[1]*College of Plant Protection, China Agricultural University, Beijing 100193, China*
[2]*Institute of Plant Protection, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*
[3]*Institute of Plant Protection, Gansu Academy of Agricultural Sciences, Lanzhou 730070, China*
[4]*College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*

Correspondence should be addressed to Haiguang Wang; wanghaiguang@cau.edu.cn

Received 14 February 2019; Revised 10 June 2019; Accepted 9 July 2019; Published 24 July 2019

Academic Editor: Alessandra Durazzo

Identifying plant pathogens for disease diagnosis and disease control strategy making is of great significance. In this study, based on near-infrared spectroscopy, a method for identifying three kinds of pathogens causing wheat smuts, including *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici*, was investigated. Based on the acquired near-infrared spectral data of the teliospore samples of the three pathogens, pathogen identification models were built in different spectral regions using distinguished partial least squares (DPLS), backpropagation neural network (BPNN), and support vector machine (SVM). Satisfactory identification results were achieved using the DPLS, BPNN, and SVM models built in each of the 22 spectral regions. By contrast, the modeling effects of DPLS and SVM were better than those of BPNN. The modeling ratio of the training set to the testing set affected the identification results of the BPNN models more than those obtained using the DPLS and SVM models. In this study, a rapid, accurate, and nondestructive method was provided for plant pathogen identification, and some basis was provided for disease diagnosis, pathogen monitoring, and disease control. Moreover, some methodological references and supports were provided for identification of quarantine wheat smut fungi in plant quarantine.

1. Introduction

The occurrence and epidemics of plant diseases can cause severe yield losses and quality decline of agricultural products and sometimes lead to serious social problems. It is crucial to prevent and control plant diseases in a timely manner for ensuring normal agricultural production and food security. Accurate and rapid identification of plant diseases and the causal agents is the most important prerequisite of disease monitoring and early warning, which are the bases for prevention and control of plant diseases. Nowadays, morphological identification methods and molecular marker-based identification techniques are commonly used for pathogen identification. For example, morphological identification methods [1–5] and molecular biological identification methods [2–4, 6–11] have been developed for identification of the causal agents of many plant smuts with important economic significance in agricultural production. However, morphological identification methods highly depend on personnel experience, and it is difficult to clearly observe very small pathogens using ordinary optical microscopes, and thus identification errors easily occur. Molecular biological identification methods are highly accurate but generally require professional instruments and reagents, professional personnel to operate, and a relatively long time to achieve identification results [12, 13]. In addition, nondestructive identification of pathogens cannot be implemented using molecular biological methods. Therefore, a new method to solve the problems in identification of plant pathogens is urgently required. To ensure the quality and safety of agricultural products and the safe production of wheat, a rapid, accurate,

and nondestructive identification method for three kinds of wheat smut fungi based on near-infrared spectroscopy (NIRS) was explored in this present study.

NIRS, a rapid, low-cost, pollution-free, and nondestructive analytical technique, can be used to conduct both qualitative analysis and quantitative analysis of samples using material information contained in near-infrared spectra and has been widely applied in agriculture, food, petroleum, chemical, pharmacy, and other industries [14–24]. Since NIRS is an indirect analytical technique, when it is applied to perform qualitative or quantitative analysis of an unknown sample, it is necessary to obtain near-infrared spectral data of a large number of known samples and then to establish a qualitative or quantitative identification model using chemometric methods.

In recent years, NIRS has been applied to detection and identification of insect pests [25, 26] and plant diseases [27–29]. Wu et al. [27] built an early detection model for gray mold (caused by *Botrytis cinerea*) on eggplant leaves based on acquired visible and near-infrared spectra using backpropagation neural network (BPNN), and an identification accuracy of 88% was achieved. For the early and rapid detection of soybean pod anthracnose (caused by *Colletotrichum truncatum*), integrating the successive projections algorithm and least square support vector machine (LS-SVM), a disease detection model with an accuracy of 95.45% was built based on visible/near-infrared spectra by Feng et al. [28]. To realize nondestructive detection of citrus huanglongbing (caused by *Candidatus liberibacter*), Liu et al. [29] acquired the near-infrared spectra (4000–9000 cm^{-1}) of three categories of citrus leaves including healthy leaves, nutrient-deficient leaves, and huanglongbing leaves, subsequently preprocessed the original spectral data using the methods including first derivative transformation, smoothing, and multiple scattered correction, and then built a LS-SVM model with an accuracy of 100% for identifying the three categories of leaves. At present, there are still few reports on the use of NIRS for identification and early monitoring of plant diseases and identification of plant pathogens.

In our previous studies, NIRS was used to implement early diagnosis and detection of wheat stripe rust caused by *Puccinia striiformis* f. sp. *tritici* (*Pst*) [30–32], qualitative discrimination of plant disease species [30, 31], disease severity assessment [33], and qualitative identification and quantitative analysis of *Pst* and *P. recondita* f. sp. *tritici* (*Prt*) based on NIRS [32, 34]. In combination with a molecular biology technique, real-time polymerase chain reaction, quantitative determination of the relative contents of *Pst* DNA in wheat leaves in incubation period, was implemented based on NIRS [32]. In addition, the research on quantitative determination of germinability of *Pst* urediospores based on NIRS was conducted [35], and the effect of ultraviolet B radiation on NIRS-based identification of *Pst* was investigated [36]. In particular, in the study of qualitative identification and quantitative analysis of *Pst* and *Prt*, based on the acquired near-infrared spectra of pure and mixed samples of the *Pst* and *Prt* urediospores, a qualitative identification model was built using distinguished partial least squares (DPLS) and a quantitative determination model was built using quantitative partial least squares [34]. The two models demonstrated good performance, and the qualitative identification and quantitative analysis of *Pst* and *Prt* were realized based on NIRS.

Based on the studies mentioned above, the rapid and accurate identification of three kinds of wheat smut fungi including *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici* causing common bunt, loose smut, and flag smut on wheat, respectively, was investigated using NIRS technology in this study. Based on the acquired near-infrared spectra of the teliospores of *T. foetida*, *Ustilago tritici*, and *Urocystis tritici*, pathogen identification models were built with different modeling ratios of training sets and testing sets in different modeling spectral regions using DPLS, BPNN, and support vector machine (SVM). The aim of this study was to provide a rapid, accurate, and nondestructive method for the identification of multiple kinds of plant pathogens and to provide some basis for disease diagnosis, pathogen monitoring, and disease control measure making. In addition, some methodological references were provided for identifying quarantine wheat smut fungi in plant quarantine.

2. Materials and Methods

2.1. Materials. *T. foetida*, *Ustilago tritici*, and *Urocystis tritici* were used as experimental materials. The specimens of *T. foetida* and *Urocystis tritici* were collected from wheat growing areas in Henan Province, China in 2014 and 2015. The specimens of *Ustilago tritici* were collected from wheat growing areas in Gansu Province, China, in 2015. For *T. foetida* and *Urocystis tritici*, diseased wheat panicles were collected. For *Ustilago tritici*, diseased wheat stems, leaves, and leaf sheaths with typical symptoms of wheat flag smut were collected.

For *T. foetida*, each of the bunt ball from the diseased wheat panicles was cracked with an inoculation needle, and the released black powdery teliospores of approximately 7–12 mg were treated as a sample. There were 1071 bunt balls obtained in this study. For *Ustilago tritici*, each diseased wheat panicle was clamped with sterile forceps, then the black powdery teliospores were shacked off and landed on a piece of a smooth-surface weighing paper. The teliospores of 6-7 mg were treated as a sample, and a total of 389 teliospore samples of *Ustilago tritici* were obtained. For *Urocystis tritici*, 10 mg of teliospores from the diseased wheat stems, leaves, and leaf sheaths were taken as a sample. A total of 30 teliospore samples of *Urocystis tritici* were obtained.

2.2. Acquisition of Near-Infrared Spectral Data. Using a FT-NIR MPA spectrometer (Bruker, Germany), the near-infrared spectra of teliospore samples of *T. foetida*, *Ustilago tritici*, and *Urocystis tritici* were acquired. Each teliospore sample was placed into a sample cup (4 mm in diameter) for near-infrared spectral measurement using the integrating sphere diffuse reflectance method. Especially, the tightness of teliospores in the sample cup was kept as consistent as possible to reduce the errors that may be induced by

different tightness. The spectral resolution and the number of scan processes in the measurement system were set as 8 cm^{-1} and 32, respectively. The near-infrared spectra in the range of 4000 to 12000 cm^{-1} were collected. Each teliospore sample was used for acquisition of only one near-infrared spectrum.

Totally, 1490 teliospore samples were measured and 1490 near-infrared spectra were acquired including 1071, 389, and 30 spectra of *T. foetida*, *Ustilago tritici*, and *Urocystis tritici*, respectively. After averaging the spectral data of *T. foetida*, *Ustilago tritici*, and *Urocystis tritici* according to their categories, the spectral curve of each category is shown in Figure 1. There were considerable differences between the average spectra in the range of 4000–12000 cm^{-1} except the range of 6000–7000 cm^{-1}. In particular, obvious differences could be found among the three average spectra in the range of 7200–12000 cm^{-1}.

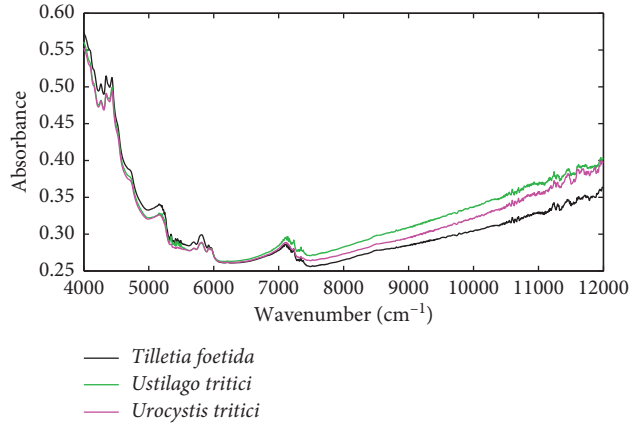

Figure 1: Spectral curves for the three kinds of fungi including *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici*.

2.3. Establishment of Pathogen Identification Models Based on NIRS. The acquired near-infrared spectra had relatively strong random noises in high-frequency regions (Figure 1). To eliminate the effects of random noises in high-frequency regions and to select the suitable spectral region for modeling, the spectral region of 4000–12000 cm^{-1} and 21 spectral regions divided from the region of 4000–10000 cm^{-1}, including 4000–5000, 4000–6000, 4000–7000, 4000–8000, 4000–9000, 4000–10000, 5000–6000, 5000–7000, 5000–8000, 5000–9000, 5000–10000, 6000–7000, 6000–8000, 6000–9000, 6000–10000, 7000–8000, 7000–9000, 7000–10000, 8000–9000, 8000–10000, and 9000–10000 cm^{-1}, were taken as modeling spectral regions. In this study, the acquired near-infrared spectral data were processed using the two following methods: Data treatment method 1 and Data treatment method 2, and then pathogen identification models were built with three modeling ratios of training sets to testing sets (i.e., 3 : 1, 4 : 1 and 5 : 1) in the different spectral regions using three modeling methods including DPLS, BPNN, and SVM. The main steps for establishment of pathogen identification models are shown in Figure 2.

Data treatment method 1 was based on all the acquired near-infrared spectral data. The near-infrared spectra of each kind of pathogen were randomly selected according to modeling ratio (3 : 1, 4 : 1 or 5 : 1) and then were integrated together to form training set and testing set, respectively. The number of near-infrared spectra of each kind of pathogen in training sets and testing sets under different modeling ratios is shown in Table 1. With the modeling ratios of 3 : 1, 4 : 1, and 5 : 1, DPLS, BPNN, and SVM were used to build the pathogen identification models in the different modeling spectral regions, respectively. The identification accuracies of training and testing sets were used to evaluate the models. While building a DPLS model, the number of principal components was set as 1, 2, 3, . . ., or 10. The number of neurons in the hidden layer while building a BPNN model was set as 10, 50, or 100. Each BPNN model was built using the neural network toolbox of the software MATLAB 8.2 (MathWorks, Natick, MA, USA) with the default settings except for changes in the number of neurons in the hidden layer and randomly taking 90% of data from the training set for training and the remaining 10% for validation. Using C-SVM in LIBSVM package developed by the Chih-Jen Lin group from Taiwan, China [37], SVM models for pathogen identification were built with radial basis function as the kernel function. The optimal values of penalty parameter C and kernel function parameter g for each SVM model were searched using grid search algorithm in the range of 2^{-10} to 2^{10} with 0.8 as the searching step. Identification accuracies were calculated by 3-fold cross validation on the training set at all points within the grid. When the identification accuracy of the training set was the highest, the values of C and g were selected as the optimal parameters of the SVM model.

For Data treatment method 2, to reduce the bias that may be induced by the great differences among the numbers of samples of the three kinds of pathogens, the near-infrared spectra of the samples of *T. foetida* were randomly divided into 10 groups that were labeled as TFG1, TFG2, TFG3, . . ., and TFG10, and the near-infrared spectra of the samples of *Ustilago tritici* were also randomly divided into 10 groups that were labeled as UTG1, UTG2, UTG3, . . ., and UTG10. Then, TFG1, UTG1, and all the near-infrared spectra of *Urocystis tritici* formed a data subset that was labeled as data subset 1. TFG2, UTG2, and all the near-infrared spectra of *Urocystis tritici* formed data subset 2. In such a way, a total of 10 data subsets (including data subset 1, data subset 2, data subset 3, . . ., and data subset 10) were formed. The near-infrared spectra of each kind of pathogen in each data subset were randomly selected according to modeling ratio (3 : 1, 4 : 1 or 5 : 1) and then were integrated together to form training set and testing set, respectively. The number of near-infrared spectra of each kind of pathogen in the training and testing sets for each data subset under different modeling ratios is shown in Table 2. Based on each data subset, using DPLS, BPNN, and SVM modeling methods as described above, the pathogen identification models were built in the different modeling spectral regions when the modeling ratios of the training sets to the testing sets were 3 : 1, 4 : 1 and 5 : 1. The average identification accuracies and the standard deviation (SD) of the identification accuracies of training sets and testing sets on the 10 data subsets were used to evaluate the

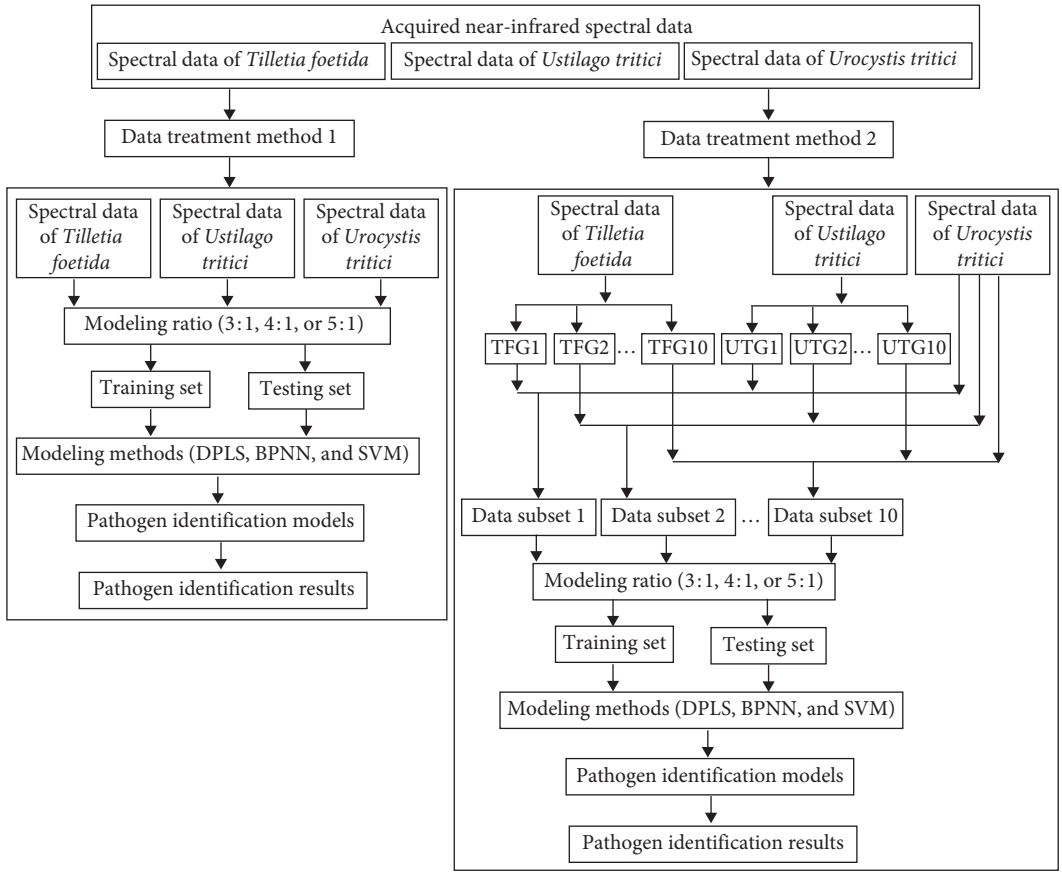

Figure 2: Work flow diagram of establishment of pathogen identification models based on near-infrared spectroscopy.

Table 1: Number of near-infrared spectra of each kind of pathogen in the training and testing sets under different modeling ratios when Data treatment method 1 was used.

Modeling ratio of training set to testing set	Training set				Testing set			
	Tilletia foetida	*Ustilago tritici*	*Urocystis tritici*	Total number of samples	*Tilletia foetida*	*Ustilago tritici*	*Urocystis tritici*	Total number of samples
3 : 1	803	291	22	1116	268	98	8	374
4 : 1	856	311	24	1191	215	78	6	299
5 : 1	892	324	25	1241	179	65	5	249

Note. There were 1490 near-infrared spectra including 1071, 389, and 30 spectra of *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici*, respectively.

modeling effect. The value of SD can be calculated using the following formula:

$$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{n} \left(A_i - \overline{A}\right)^2}{n}}, \quad (1)$$

where SD is the standard deviation of the identification accuracies of training sets or testing sets on n data subsets, A_i is the identification accuracy of the training set or the testing set on the ith data subset, $\overline{A}$ is the average of the identification accuracies of training sets or testing sets on n data subsets, and n is the number of the data subsets. In this study, n was equal to 10.

For each modeling method (DPLS, BPNN, or SVM), when Data treatment method 1 or Data treatment method 2 was used, the optimal modeling results obtained with each modeling ratio of the training set to the testing set in each near-infrared spectral region were selected. When the DPLS models were built for pathogen identification, the optimal modeling results were selected from the identification results obtained with different numbers of principal components in a modeling spectral region, according to the identification accuracies of training sets and testing sets (Data treatment method 1) or the average identification accuracies and the SD values for training sets and test sets on the 10 data subsets (Data treatment method 2). If the same identification accuracies for training sets and testing sets were obtained when different numbers of principal components were used, the modeling effect with the minimum number of principal components was optimal. When the BPNN models were built, the optimal modeling results were selected from the results obtained with different numbers of neurons in the

Table 2: Number of near-infrared spectra of each kind of pathogen in the training and testing sets for each data subset under different modeling ratios when Data treatment method 2 was used.

Modeling ratio of training set to testing set	Data subset	Training set				Testing set			
		Tilletia foetida	*Ustilago tritici*	*Urocystis tritici*	Total number of samples	*Tilletia foetida*	*Ustilago tritici*	*Urocystis tritici*	Total number of samples
3 : 1	1	80	29	22	131	27	10	8	45
3 : 1	2	80	29	22	131	27	10	8	45
3 : 1	3	80	29	22	131	27	10	8	45
3 : 1	4	80	29	22	131	27	10	8	45
3 : 1	5	80	29	22	131	27	10	8	45
3 : 1	6	80	29	22	131	27	10	8	45
3 : 1	7	80	29	22	131	27	10	8	45
3 : 1	8	80	29	22	131	27	10	8	45
3 : 1	9	80	29	22	131	27	10	8	45
3 : 1	10	81	28	22	131	27	10	8	45
4 : 1	1	85	31	24	140	22	8	6	36
4 : 1	2	85	31	24	140	22	8	6	36
4 : 1	3	85	31	24	140	22	8	6	36
4 : 1	4	85	31	24	140	22	8	6	36
4 : 1	5	85	31	24	140	22	8	6	36
4 : 1	6	85	31	24	140	22	8	6	36
4 : 1	7	85	31	24	140	22	8	6	36
4 : 1	8	85	31	24	140	22	8	6	36
4 : 1	9	85	31	24	140	22	8	6	36
4 : 1	10	86	30	24	140	22	8	6	36
5 : 1	1	89	32	25	146	18	7	5	30
5 : 1	2	89	32	25	146	18	7	5	30
5 : 1	3	89	32	25	146	18	7	5	30
5 : 1	4	89	32	25	146	18	7	5	30
5 : 1	5	89	32	25	146	18	7	5	30
5 : 1	6	89	32	25	146	18	7	5	30
5 : 1	7	89	32	25	146	18	7	5	30
5 : 1	8	89	32	25	146	18	7	5	30
5 : 1	9	89	32	25	146	18	7	5	30
5 : 1	10	90	31	25	146	18	7	5	30

Note. There were 1490 near-infrared spectra including 1071, 389, and 30 spectra of *Tilletia foetida*, *Ustilago tritici*, and *Urocystis tritici*, respectively. There were 10 data subsets, and each data subset was composed of 176 spectra.

hidden layer in a modeling spectral region, according to the identification accuracies of training sets and testing sets (Data treatment method 1) or the average identification accuracies and the SD values for training sets and testing sets on the 10 data subset (Data treatment method 2). If the same identification accuracies for training and testing sets were obtained when different numbers of neurons in the hidden layer were used, the modeling effect with the minimum number of neurons was optimal.

Moreover, when the pathogen identification models were built using DPLS, BPNN, or SVM, a comprehensive comparison of the optimal modeling results obtained in different modeling spectral regions was performed. A comparison of the modeling results obtained with Data treatment method 1 and Data treatment method 2 was also conducted.

3. Results

3.1. Pathogen Identification Results of the DPLS, BPNN, and SVM Models Using Data Treatment Method 1. When the acquired near-infrared spectral data were processed with Data treatment method 1, very good performance with the identification accuracies of the training and testing sets of more than 97% was achieved using the DPLS models built with different modeling ratios in each of the 22 modeling spectral regions (Table 3). The identification accuracies of the training and testing sets reached 100% for the DPLS models built with different modeling ratios in the modeling spectral regions, except for the spectral regions of 4000–5000, 6000–7000, 8000–9000, 8000–10000, and 9000–10000 cm^{-1}. When the DPLS models were built in the narrow spectral regions, the identification accuracies of the training and testing sets could reach 100% only for the models with different modeling ratios in the two spectral regions of 5000–6000 and 7000–8000 cm^{-1}. The identification accuracies of the training and testing sets could reach 100% for the DPLS models built with different modeling ratios in the relatively wide spectral regions, except for the spectral region of 8000–10000 cm^{-1}. The results indicated that the width of modeling spectral region had certain effects on the identification results of the DPLS models. In addition, the results demonstrated that relatively low-frequency regions and relatively high-frequency regions in the acquired near-infrared spectra could decrease the identification

Table 3: The optimal results except those with both the identification accuracy of the training set and the identification accuracy of the testing set equal to 100% when the DPLS models for Data treatment method 1 were built with different modeling ratios in each modeling spectral region.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	The number of principal components	Identification accuracy of training set (%)	Identification accuracy of testing set (%)
4000–5000	3 : 1	9	99.91	99.73
4000–5000	4 : 1	9	100	99.67
4000–5000	5 : 1	10	99.92	100
6000–7000	3 : 1	10	99.46	98.66
6000–7000	4 : 1	9	99.33	99.33
6000–7000	5 : 1	9	99.44	98.80
8000–9000	3 : 1	8	98.21	98.13
8000–9000	4 : 1	9	98.40	98.33
8000–9000	5 : 1	9	98.39	97.99
8000–10000	3 : 1	10	99.91	99.47
8000–10000	4 : 1	8	99.92	99.33
8000–10000	5 : 1	9	99.84	99.60
9000–10000	3 : 1	9	98.03	97.33
9000–10000	4 : 1	7	97.98	97.66
9000–10000	5 : 1	9	97.99	97.59

accuracies of the training and testing sets. In particular, high-frequency regions with relatively wide ranges had some effects on the identification results of the models with different modeling ratios.

When the acquired near-infrared spectral data were processed with Data treatment method 1, the BPNN models built with different modeling ratios in each modeling spectral region performed well (Table 4). The optimal identification accuracies of the training and testing sets were 87.66% and 88.29%, respectively, when the BPNN model was built with the modeling ratio equal to 4 : 1 and the number of neurons in the hidden layer equal to 10 in the spectral region of 9000–10000 cm^{-1}. The identification accuracies of the training and testing sets reached more than 90% in other cases. When the BPNN models were built in the spectral region of 4000–12000 cm^{-1}, the identification accuracies of the training and testing sets for the models built with the modeling ratios of 3 : 1 and 5 : 1 reached 100%, and those for the model built with the modeling ratio equal to 4 : 1 and the number of neurons in the hidden layer equal to 50 were 99.83% and 100%, respectively. When the BPNN models were built in the 21 spectral regions from the range of 4000 to 10000 cm^{-1}, the optimal identification accuracies of the training sets and the testing sets for the models with different modeling ratios reached 100%, except for the spectral regions of 4000–6000, 5000–6000, 5000–9000, 6000–7000, 6000–8000, 6000–9000, 6000–10000, 7000–8000, 8000–9000, 8000–10000, and 9000–10000 cm^{-1}. For the BPNN models built in each of the 21 spectral regions from the range of 4000–10000 cm^{-1} except the spectral regions of 6000–7000, 8000–9000, 8000–10000, and 9000–10000 cm^{-1}, at least when one modeling ratio was used, both the identification accuracy of the training set and that of the testing set reached 100%. For the BPNN models built in the narrow spectral regions, 4000–5000, 5000–6000, and 7000–8000 cm^{-1}, both the identification accuracies of the training and testing sets could reach 100%. However, when the BPNN models were built in the spectral regions of 5000–6000 and 7000–8000 cm^{-1}, the modeling ratio affected the modeling results. For the BPNN models built in the spectral region of 5000–6000 cm^{-1}, when the modeling ratios were 3 : 1, 4 : 1, and 5 : 1, the identification accuracies of the training sets were 100%, 97.98%, and 99.92%, respectively, and those of the testing sets were 100%, 97.99%, and 100%, respectively. For the BPNN models built with the modeling ratios of 3 : 1, 4 : 1, and 5 : 1 in the spectral region of 7000–8000 cm^{-1}, the identification accuracies of the training sets were 100%, 97.98%, and 100%, respectively, and those of the testing sets were 100%, 97.99%, and 100%, respectively.

When Data treatment method 1 was used, the SVM models that were built with different modeling ratios in each modeling spectral region presented very good performances, with the optimal identification accuracies of the training and testing sets of more than 99% (Table 5). For the SVM models that were built in the modeling spectral regions, except the spectral region of 9000–10000 cm^{-1}, the identification accuracies of 100% of the training and testing sets were achieved at least when one modeling ratio was used. When the SVM models were built in the narrow spectral regions of 4000–5000, 6000–7000, 7000–8000, and 8000–9000 cm^{-1}, both the identification accuracies of the training sets and those of the testing sets reached 100% when each of the three modeling ratios was used. For the SVM models built in the spectral region of 5000–6000 cm^{-1}, when the modeling ratios were 3 : 1 and 4 : 1, the optimal identification accuracies of the training sets were 99.91% and 100%, respectively, and those of the testing sets were 100% and 99.67%, respectively, and when the modeling ratio was 5 : 1, both the optimal identification accuracy of the training set and that of the testing set were 100%. For the SVM models built in the spectral region of 9000–10000 cm^{-1}, the optimal identification accuracy of the training set was 100% and that of the testing set was 99.73% when the modeling ratio was 3 : 1; when the modeling ratio was 4 : 1, the optimal identification accuracies of the training set and the testing set were 99.92% and 100%, respectively; and when the modeling ratio was 5 : 1,

Table 4: The optimal results except those with both the identification accuracy of the training set and the identification accuracy of the testing set equal to 100% when the BPNN models for Data treatment method 1 were built with different modeling ratios in each modeling spectral region.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	The number of neurons in the hidden layer	Identification accuracy of training set (%)	Identification accuracy of testing set (%)
4000–12000	4 : 1	50	99.83	100
4000–6000	4 : 1	10	97.98	97.99
4000–6000	5 : 1	10	97.99	97.99
5000–6000	4 : 1	10	97.98	97.99
5000–6000	5 : 1	100	99.92	100
5000–9000	3 : 1	10	98.03	97.86
6000–7000	3 : 1	100	98.03	97.59
6000–7000	4 : 1	50	97.98	97.99
6000–7000	5 : 1	10	97.99	97.59
6000–8000	3 : 1	50	98.03	97.86
6000–9000	4 : 1	10	97.98	97.99
6000–10000	3 : 1	10	98.03	97.86
6000–10000	4 : 1	10	97.98	97.99
7000–8000	4 : 1	10	97.98	97.99
8000–9000	3 : 1	50	98.03	97.86
8000–9000	4 : 1	50	97.98	97.99
8000–9000	5 : 1	50	97.99	97.99
8000–10000	3 : 1	100	98.03	97.86
8000–10000	4 : 1	50	97.98	97.99
8000–10000	5 : 1	10	97.99	97.99
9000–10000	3 : 1	50	92.11	91.98
9000–10000	4 : 1	10	87.66	88.29
9000–10000	5 : 1	50	97.58	97.19

Table 5: The optimal results except those with both the identification accuracy of the training set and the identification accuracy of the testing set equal to 100% when the SVM models for Data treatment method 1 were built with different modeling ratios in each modeling spectral region.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	Optimal parameters		Identification accuracy of training set (%)	Identification accuracy of testing set (%)
		C	g		
4000–12000	3 : 1	0.2500	6.9644	100	99.73
4000–6000	4 : 1	1.3195	64.0000	100	99.67
4000–10000	3 : 1	0.4353	6.9644	100	99.73
5000–6000	3 : 1	4.0000	36.7583	99.91	100
5000–6000	4 : 1	1.3195	337.7940	100	99.67
5000–10000	3 : 1	0.4353	12.1257	100	99.73
5000–10000	5 : 1	0.7579	111.4305	100	99.60
6000–8000	3 : 1	6.9644	64.0000	100	99.73
6000–8000	5 : 1	2.2974	588.1336	100	99.20
6000–9000	3 : 1	1.3195	21.1121	100	99.47
6000–9000	5 : 1	1.3195	337.7940	100	99.20
6000–10000	3 : 1	1.3195	12.1257	99.91	99.73
9000–10000	3 : 1	64.0000	111.4305	100	99.73
9000–10000	4 : 1	588.1336	4.0000	99.92	100
9000–10000	5 : 1	12.1257	337.7940	100	99.60

those of the training set and the testing set were 100% and 99.60%, respectively.

As described above, satisfactory results could be obtained using the pathogen identification models built using DPLS, BPNN, and SVM in different spectral regions. Compared with the DPLS and BPNN modeling methods, the identification accuracies of the training and testing sets could reach 100% in more spectral regions when the pathogen identification models were built using SVM. By contrast, the effects of the modeling ratio on the identification results of the models built using BPNN and SVM were greater than those on the identification results of the models built using DPLS.

3.2. Pathogen Identification Results of the DPLS, SVM, and BPNN Models Using Data Treatment Method 2. When Data treatment method 2 was used, very satisfactory modeling results were achieved using the DPLS models built on each data subset with different modeling ratios in each modeling

spectral region. The average identification accuracies of the training sets and the testing sets on the 10 data subsets reached more than 98% (Table 6). Except for the models built with the modeling ratio of 3 : 1 in the spectral region of 6000–7000 cm^{-1} and the models built on each data subset with the three modeling ratios in the spectral region of 8000–9000 or 9000–10000 cm^{-1}, the average identification accuracies of the training sets and the testing sets on the 10 data subsets reached 100% for the DPLS models built on each data subset with different modeling ratios in each of the other modeling spectral regions. For the DPLS models built in the narrow spectral region of 4000–5000, 5000–6000, 6000–7000, or 7000–8000 cm^{-1}, both the average identification accuracies of the training and testing sets on the 10 data subsets reached 100%, when one or more than one modeling ratios were used. By contrast, there was some reduction in the average identification accuracies of the training sets and the testing sets on the 10 data subsets when the DPLS models were built in the narrow high-frequency region of 8000–9000 or 9000–10000 cm^{-1}, demonstrating that high-frequency regions had an impact on the modeling results. Although both the average identification accuracies of the training sets and those of the testing sets on the 10 data subsets did not reach 100% for the models built with the modeling ratio of 3 : 1 in the spectral region of 6000–7000 cm^{-1} and for those built with each modeling ratio in the spectral region of 8000–9000 or 9000–10000 cm^{-1}, the average identification accuracies were still very high, and the SD values of the identification accuracies for the 10 data subsets were very small. The results indicated that satisfactory identification performance could be achieved using the DPLS models based on NIRS and that the models had high stability.

When Data treatment method 2 was used, satisfactory identification results could be obtained using the BPNN models built on each data subset in each modeling spectral region (Table 7), but the modeling ratio had a great influence on the identification results of the models. For the BPNN models built in the spectral region of 4000–8000 cm^{-1}, both the average identification accuracy of the training sets and that of the testing sets on the 10 data subsets reached 100% when the modeling ratio was 4 : 1 or 5 : 1. When the BPNN models were built with the modeling ratio of 5 : 1 in the spectral region of 5000–6000 cm^{-1} on the 10 data subsets, both the average identification accuracy of the training sets and that of the testing sets reached 100%. When the BPNN models were built with the modeling ratio of 3 : 1 in the spectral region of 5000–10000 cm^{-1} on the 10 data subsets, both the average identification accuracy of the training sets and that of the testing sets reached 100%. For the models built with the modeling ratio of 3 : 1 or 5 : 1 in the spectral region of 4000–5000 cm^{-1}, the ones built with the modeling ratio of 4 : 1 in the spectral region of 5000–6000 cm^{-1}, the ones built with the modeling ratio of 3 : 1 or 5 : 1 in the spectral region of 6000–7000 cm^{-1}, and the ones built with the modeling ratio of 3 : 1 in the spectral region of 8000–10000 cm^{-1}, the SD values of the identification accuracies of the training and testing sets were both relatively high, indicating poor modeling effects.

When Data treatment method 2 was used, the SVM models built on the 10 data subsets, with each modeling ratio in each modeling spectral region, presented very good performance with the averages of the identification accuracies of the training sets and the testing sets of more than 98% (Table 8). Very small SD values were obtained for the identification accuracies of the training and testing sets. For the SVM models built on the 10 data subsets with the modeling ratio of 5 : 1 in the spectral region of 4000–7000 cm^{-1}, the average identification accuracies of the training and testing sets were 99.93% and 100%, respectively, and the SD values for the training and testing sets were 0.002040 and 0, respectively. For the SVM models built on the 10 data subsets with the modeling ratio of 5 : 1 in the spectral region of 4000–12000 cm^{-1}, the average identification accuracies of the training and testing sets were 99.86% and 100%, respectively, and the SD values for the training and testing sets were 0.002720 and 0, respectively. For the SVM models built on the 10 data subsets with the modeling ratio of 4 : 1 in the spectral region of 5000–6000 cm^{-1}, the average identification accuracies of the training and testing sets were 99.86% and 100%, respectively, and the SD values for the training and testing sets were 0.004290 and 0, respectively. The identification results of the models in the three cases above were the best. By contrast, the worst identification results were obtained when the models were built on the 10 data subsets with the modeling ratio of 5 : 1 or 3 : 1 in the spectral region of 8000–9000 cm^{-1} and on the 10 data subsets with the modeling ratio of 4 : 1 in the spectral region of 9000–10000 cm^{-1}. For the models built in these three cases, the average identification accuracies of the training sets were 99.73%, 99.77%, and 99.57%, respectively, and the SD values for the training sets were 0.006261, 0.004893, and 0.005715, respectively; the average identification accuracies of the testing sets were 98.33%, 98.00%, and 98.06%, respectively, and the SD values for the testing sets were 0.02236, 0.01554, and 0.03298, respectively.

The results showed that the pathogen identification models built using DPLS performed better than those built using BPNN or SVM when Data treatment method 2 was used. The modeling effect of DPLS on each data subset was more stable than that of BPNN. When the pathogen identification models were built using BPNN in the spectral region of 4000–5000, 5000–6000, 6000–7000, or 8000–10000 cm^{-1}, greater SD values of the identification accuracies of the training and testing sets were obtained. Although the modeling effect of SVM was not as good as that of DPLS, very satisfactory identification accuracies of the training and testing sets were achieved using the SVM models built on each data subset.

3.3. Comparison of Modeling Results Obtained with Data Treatment Methods 1 and 2. The modeling effect using DPLS with Data treatment method 2 was better than that achieved with Data treatment method 1. When Data treatment method 2 was used, the average identification accuracies of the training and testing sets did not simultaneously reach 100% for the DPLS models built on the 10 data subsets with

Table 6: The optimal results except the average identification accuracies and the standard deviations of the identification accuracies for the 10 data subsets for the training sets and the testing sets were 100% and 0 when the DPLS models for Data treatment method 2 were built with different modeling ratios in each modeling spectral region.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	The number of principal components	Training sets		Testing sets	
			Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets	Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets
6000–7000	3 : 1	7	100	0	99.78	0.006660
8000–9000	3 : 1	6	100	0	99.78	0.006660
8000–9000	4 : 1	6	100	0	99.44	0.01112
8000–9000	5 : 1	6	100	0	99.67	0.009990
9000–10000	3 : 1	5	99.70	0.006994	98.45	0.01734
9000–10000	4 : 1	6	100	0	98.61	0.02241
9000–10000	5 : 1	5	99.66	0.01026	99.33	0.01332

each modeling ratio only in two spectral regions (8000–9000 and 9000–10000 cm^{-1}). When Data treatment method 1 was used, both the identification accuracy of the training set and that of the testing set were less than 100% for the DPLS models built with each modeling ratio in five spectral regions (4000–5000, 6000–7000, 8000–9000, 8000–10000, and 9000–10000 cm^{-1}).

When Data treatment method 1 was used, the optimal accuracies of the training and testing sets could reach more than 97% for the BPNN models built in different spectral regions. Among the identification results obtained using the BPNN models built with different modeling ratio in each spectral region, both the optimal accuracy of the training set and that of the testing set were less than 97%, only for the models built with the modeling ratio of 3 : 1 or 4 : 1 in the spectral region of 9000–10000 cm^{-1}. However, when Data treatment method 2 was used, relatively great differences were observed between the average identification accuracies of the training sets and between the average identification accuracies of the testing sets for the BPNN models built on the 10 data subsets in different modeling spectral regions. Among the selected optimal identification results, the range of the average identification accuracies of the training sets was 88.77–100% and the range of the average identification accuracies of the testing sets was 87.67–100%. In addition, when Data treatment method 2 was used, the modeling ratio greatly influenced the identification results of the BPNN models. Relatively large identification accuracy SD values of the training and testing sets were obtained when the BPNN models were built on the 10 data subsets with one or two modeling ratios in some spectral regions (4000–5000, 5000–6000, 6000–7000, and 8000–10000 cm^{-1}). This indicated that the poor modeling effects were achieved on some data subsets using BPNN and that there were relatively great differences both between the identification accuracies of the training sets and between the identification accuracies of the testing sets for the 10 BPNN models built with certain modeling ratio in certain modeling spectral region. For instance, when Data treatment method 2 was used and the BPNN models were built in the spectral region of 4000–5000 cm^{-1} on the 10 data subsets with a ratio of the training set to the testing set of 3 : 1, the optimal identification results (Figure 3) were achieved with the number of neurons in the hidden layer equal to 50. The SD values of the identification accuracies of the training and testing sets on the 10 data subsets were 0.1541 and 0.1508, respectively. Relatively low identification accuracies of the training and testing sets were obtained for the BPNN models built on data subset 1 and data subset 2. For the BPNN model built on data subset 1, the identification accuracies of the training and testing sets were 70.23% and 66.67%, respectively. For the BPNN model built on data subset 2, the identification accuracies of the training and testing sets were 54.20% and 57.78%, respectively. The results indicated that these two models presented poor performance in pathogen identification.

When Data treatment method 1 or Data treatment method 2 was used, the built SVM models presented very good performance with very high identification accuracies of both the training sets and the testing sets. In particular, when Data treatment method 2 was used, for the SVM models built on the 10 data subsets, very small SD values of the identification accuracies of the training and testing sets were obtained. It was demonstrated that very satisfactory identification results could be achieved using the SVM models built on each data subset and that this performance had very high stability.

4. Conclusions and Discussion

The application of NIRS to the identification of three kinds of wheat smut pathogens, including *T. foetida*, *Ustilago tritici*, and *Urocystis tritici*, was investigated in this study. The acquired near-infrared spectral data were processed using Data treatment method 1 and Data treatment method 2 to form the training and testing sets. Then, the DPLS, BPNN, and SVM models for pathogen identification were built in 22 spectral regions. Using Data treatment method 1 or Data treatment method 2, satisfactory identification results were achieved when each of the three modeling methods was used for model-building in different spectral regions. By contrast, the modeling effect of DPLS was the most stable, followed by that of SVM and that of BPNN was the worst. In this study, a rapid, accurate, and nondestructive method for the identification of the three kinds of wheat

Table 7: Pathogen identification results of the BPNN models built with different modeling ratios in each modeling spectral region when Data treatment method 2 was used.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	The number of neurons in the hidden layer	Training sets		Testing sets	
			Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets	Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets
4000–12000	3 : 1	100	99.70	0.006994	99.78	0.006660
4000–12000	4 : 1	10	98.00	0.04237	97.22	0.04648
4000–12000	5 : 1	50	98.02	0.04058	97.67	0.05176
4000–5000	3 : 1	50	92.14	0.1541	92.00	0.1508
4000–5000	4 : 1	100	96.57	0.04367	96.39	0.04488
4000–5000	5 : 1	100	92.88	0.1175	93.33	0.1183
4000–6000	3 : 1	50	99.54	0.009770	99.56	0.01332
4000–6000	4 : 1	50	99.71	0.008580	99.44	0.01668
4000–6000	5 : 1	10	99.59	0.01024	99.33	0.01332
4000–7000	3 : 1	10	98.02	0.04279	97.78	0.04662
4000–7000	4 : 1	50	95.29	0.08590	95.00	0.08766
4000–7000	5 : 1	50	99.59	0.01024	99.67	0.009990
4000–8000	3 : 1	10	99.31	0.01617	99.33	0.01017
4000–8000	4 : 1	50	100	0	100	0
4000–8000	5 : 1	50	100	0	100	0
4000–9000	3 : 1	10	99.92	0.002280	99.78	0.006660
4000–9000	4 : 1	10	99.57	0.009156	99.17	0.01780
4000–9000	5 : 1	50	99.93	0.002040	99.67	0.009990
4000–10000	3 : 1	10	99.77	0.004893	99.78	0.006660
4000–10000	4 : 1	10	99.64	0.007314	99.44	0.01668
4000–10000	5 : 1	10	99.73	0.006261	100	0
5000–6000	3 : 1	50	99.54	0.01143	99.11	0.02038
5000–6000	4 : 1	50	95.14	0.1272	94.44	0.1400
5000–6000	5 : 1	50	100	0	100	0
5000–7000	3 : 1	10	99.62	0.009181	99.33	0.01421
5000–7000	4 : 1	10	99.29	0.01153	98.89	0.01844
5000–7000	5 : 1	10	97.67	0.05708	97.33	0.05925
5000–8000	3 : 1	10	99.85	0.004590	99.56	0.008880
5000–8000	4 : 1	50	99.86	0.004290	99.44	0.01112
5000–8000	5 : 1	10	99.80	0.006150	100	0
5000–9000	3 : 1	100	99.92	0.002280	99.78	0.006660
5000–9000	4 : 1	50	99.79	0.004573	99.44	0.01668
5000–9000	5 : 1	100	99.80	0.004381	99.67	0.009990
5000–10000	3 : 1	50	100	0	100	0
5000–10000	4 : 1	50	100	0	99.72	0.008340
5000–10000	5 : 1	50	99.66	0.01026	99.67	0.009990
6000–7000	3 : 1	50	94.28	0.1015	94.22	0.1010
6000–7000	4 : 1	50	95.72	0.08484	94.72	0.08094
6000–7000	5 : 1	50	88.77	0.1922	87.67	0.1700
6000–8000	3 : 1	50	100	0	99.78	0.006660
6000–8000	4 : 1	50	99.86	0.004290	99.44	0.01112
6000–8000	5 : 1	50	97.88	0.04379	97.00	0.06046
6000–9000	3 : 1	100	98.24	0.04069	98.00	0.04604
6000–9000	4 : 1	50	99.86	0.002840	99.44	0.01112
6000–9000	5 : 1	50	99.93	0.002040	99.67	0.009990
6000–10000	3 : 1	100	99.47	0.01186	99.33	0.01421
6000–10000	4 : 1	50	97.93	0.04233	96.67	0.04779
6000–10000	5 : 1	10	99.04	0.02435	99.33	0.01332
7000–8000	3 : 1	50	99.85	0.003040	99.33	0.01421
7000–8000	4 : 1	50	99.86	0.004290	99.44	0.01668
7000–8000	5 : 1	50	99.93	0.002040	99.00	0.02135
7000–9000	3 : 1	50	98.24	0.04040	97.33	0.04534
7000–9000	4 : 1	50	98.43	0.03805	97.22	0.04648
7000–9000	5 : 1	50	97.95	0.04029	97.33	0.05122
7000–10000	3 : 1	50	95.65	0.08542	95.56	0.08374

Table 7: Continued.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	The number of neurons in the hidden layer	Training sets		Testing sets	
			Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets	Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets
7000–10000	4 : 1	50	99.29	0.01356	98.33	0.02833
7000–10000	5 : 1	50	99.66	0.006296	99.33	0.01332
8000–9000	3 : 1	50	97.79	0.04002	97.11	0.03854
8000–9000	4 : 1	50	98.00	0.03949	96.39	0.05423
8000–9000	5 : 1	50	97.74	0.04030	96.67	0.04945
8000–10000	3 : 1	10	94.12	0.1173	93.56	0.1207
8000–10000	4 : 1	10	96.14	0.06422	95.28	0.05700
8000–10000	5 : 1	10	97.13	0.04351	96.67	0.05579
9000–10000	3 : 1	50	93.89	0.08533	93.56	0.08401
9000–10000	4 : 1	50	97.36	0.05689	96.11	0.07049
9000–10000	5 : 1	50	97.74	0.04076	97.00	0.05261

smut fungi was provided, and methodological references were also provided for the identification of quarantine wheat smut fungi in the processes of plant quarantine.

A variety of wheat diseases are caused by smut fungi, including wheat common bunt, wheat loose smut, wheat flag smut, wheat dwarf bunt, and wheat Karnal bunt. The diseases caused by wheat smut fungi can cause serious yield losses in production and even losses of up to 100% under appropriate environmental conditions [2, 11, 38]. In addition, the flour quality can seriously decline due to contamination by the causal agents. Among wheat smut fungi, *Tilletia controversa* (the causal agent of wheat dwarf bunt) and *T. indica* (the causal agent of wheat Karnal bunt) are two kinds of important entry-exit quarantine pests in many countries. In the processes of import and export wheat quarantine, it is critical to detect and identify the species of wheat smut fungi and the categories of wheat bunt galls carried in wheat. At present, the identification of these pathogens is conducted mainly based on morphological observation [1–5] and molecular biological detection [2–4, 9, 11]. Morphological observation is usually performed by using light microscopy and scanning electron microscopy. Molecular biological methods based on molecular markers, such as intergenic spacer [9], internal transcribed spacer [11], and extension factor [11], have been used to identify wheat smut fungi. NIRS is a rapid and nondestructive analytical technique [14, 15, 18, 22]. In this study, a method based on NIRS was provided for identification of the three kinds of wheat smut fungi including *T. foetida*, *Ustilago tritici*, and *Urocystis tritici*. In future studies, NIRS can be used to identify other pathogens and an NIRS-based identification system can be developed for the identification of multiple kinds of plant pathogens. Moreover, direct detection of bunt gall and healthy wheat seeds using NIRS can be explored to provide technical support for rapid online detection of imported and exported wheat.

The numbers of acquired near-infrared spectra of the three kinds of pathogens obtained in this study were not equal. There was a considerable difference among the numbers, and the spectra of *Urocystis tritici* were fewer than those of the other two pathogens. Although all the acquired near-infrared spectra were divided into different data subsets using Data treatment method 2 and relatively perfect results were achieved using the built DPLS, BPNN, and SVM models, the bias in pathogen identification may still be caused by the unbalanced numbers of samples. In further studies, the collection of *Urocystis tritici* samples and the acquisition of near-infrared spectra of the corresponding samples can be strengthened to obtain more near-infrared spectra of the pathogen and the pathogen identification models can be further optimized to ensure the accurate identification stability of the established models.

Near-infrared spectra contain material information that is useful to identify samples, and some spectral bands associated with material signals play important roles in qualitative analysis [14, 15, 18, 22]. To the best of our knowledge, systematic studies on material composition of wheat smut fungi have not been reported. In the future, the studies on this aspect can be conducted to explore mechanisms to interpret the near-infrared spectra of different wheat smut fungi, which is conducive to selecting the suitable modeling spectral region for pathogen identification.

To reduce noise interference and improve the signal-to-noise ratio, the number of scan processes in the measurement system was set as 8 cm^{-1} when near-infrared spectra of pathogen samples were acquired in this study. In addition, when teliospores were placed into the sample cup, the tightness of the teliospores was kept as consistent as possible to reduce the differences that may be induced by different tightness. Based on original near-infrared spectral data, satisfactory modeling results were obtained using the models built in the simply divided modeling spectral regions in this study. Therefore, no mathematical methods were applied to preprocess the acquired original near-infrared spectral data and no statistical methods were used to perform the modeling spectral region selection. At present, there are a variety of spectral preprocessing methods and modeling spectral region selection methods [15, 22]. In future studies, original near-infrared spectral data can be preprocessed using the selected suitable method to eliminate background interference that may exist, a suitable method can be chosen

Table 8: Pathogen identification results of the SVM models built with different modeling ratios in each modeling spectral region when Data treatment method 2 was used.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	Training sets		Testing sets	
		Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets	Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets
4000–12000	3 : 1	99.92	0.002280	99.78	0.006660
4000–12000	4 : 1	99.79	0.003254	99.17	0.02499
4000–12000	5 : 1	99.86	0.002720	100	0
4000–5000	3 : 1	100	0	99.78	0.006660
4000–5000	4 : 1	100	0	99.72	0.008340
4000–5000	5 : 1	100	0	99.67	0.009990
4000–6000	3 : 1	100	0	99.56	0.008880
4000–6000	4 : 1	99.86	0.004290	99.72	0.008340
4000–6000	5 : 1	100	0	99.67	0.009990
4000–7000	3 : 1	100	0	99.11	0.01776
4000–7000	4 : 1	100	0	99.72	0.008340
4000–7000	5 : 1	99.93	0.002040	100	0
4000–8000	3 : 1	100	0	99.11	0.01473
4000–8000	4 : 1	100	0	99.72	0.008340
4000–8000	5 : 1	100	0	99.67	0.009990
4000–9000	3 : 1	100	0	99.56	0.008880
4000–9000	4 : 1	100	0	98.89	0.02545
4000–9000	5 : 1	99.66	0.01026	98.67	0.03055
4000–10000	3 : 1	100	0	99.56	0.008880
4000–10000	4 : 1	99.86	0.002840	99.17	0.02499
4000–10000	5 : 1	99.73	0.006261	98.67	0.03055
5000–6000	3 : 1	99.70	0.009150	99.33	0.01421
5000–6000	4 : 1	99.86	0.004290	100	0
5000–6000	5 : 1	100	0	99.67	0.009990
5000–7000	3 : 1	100	0	99.33	0.01421
5000–7000	4 : 1	100	0	99.72	0.008340
5000–7000	5 : 1	100	0	99.67	0.009990
5000–8000	3 : 1	100	0	99.11	0.01088
5000–8000	4 : 1	100	0	98.89	0.01362
5000–8000	5 : 1	99.86	0.002720	98.67	0.02211
5000–9000	3 : 1	100	0	99.56	0.008880
5000–9000	4 : 1	100	0	99.72	0.008340
5000–9000	5 : 1	99.59	0.01233	98.67	0.03055
5000–10000	3 : 1	100	0	99.56	0.008880
5000–10000	4 : 1	99.86	0.002840	98.89	0.03333
5000–10000	5 : 1	100	0	99.67	0.009990
6000–7000	3 : 1	99.85	0.003040	99.34	0.01017
6000–7000	4 : 1	99.86	0.002840	98.33	0.03333
6000–7000	5 : 1	99.80	0.003116	99.00	0.01526
6000–8000	3 : 1	99.92	0.002280	98.89	0.01110
6000–8000	4 : 1	100	0	99.17	0.01274
6000–8000	5 : 1	99.93	0.002040	98.67	0.01631
6000–9000	3 : 1	99.92	0.002280	99.11	0.01473
6000–9000	4 : 1	99.86	0.002840	98.61	0.03345
6000–9000	5 : 1	99.86	0.002720	99.00	0.01526
6000–10000	3 : 1	99.92	0.002280	99.56	0.008880
6000–10000	4 : 1	99.86	0.002840	98.61	0.03345
6000–10000	5 : 1	99.93	0.002040	99.67	0.009990
7000–8000	3 : 1	100	0	99.11	0.01088
7000–8000	4 : 1	99.72	0.004734	98.89	0.02545
7000–8000	5 : 1	100	0	99.33	0.01332
7000–9000	3 : 1	99.77	0.006870	99.11	0.01088
7000–9000	4 : 1	99.86	0.002840	98.61	0.02241
7000–9000	5 : 1	100	0	99.33	0.01333
7000–10000	3 : 1	99.85	0.004590	99.33	0.01017

Table 8: Continued.

Spectral region (cm^{-1})	Modeling ratio of training set to testing set	Training sets		Testing sets	
		Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets	Average identification accuracy for the 10 data subsets (%)	Standard deviation of the identification accuracies for the 10 data subsets
7000–10000	4 : 1	99.86	0.002840	98.89	0.01844
7000–10000	5 : 1	99.86	0.004110	99.33	0.01332
8000–9000	3 : 1	99.77	0.004893	98.00	0.01554
8000–9000	4 : 1	99.86	0.002840	98.33	0.02224
8000–9000	5 : 1	99.73	0.006261	98.33	0.02236
8000–10000	3 : 1	99.92	0.002280	99.56	0.01332
8000–10000	4 : 1	99.79	0.004573	98.61	0.03345
8000–10000	5 : 1	99.80	0.004381	99.00	0.02135
9000–10000	3 : 1	99.70	0.005065	98.67	0.02266
9000–10000	4 : 1	99.57	0.005715	98.06	0.03298
9000–10000	5 : 1	99.52	0.005344	98.67	0.01631

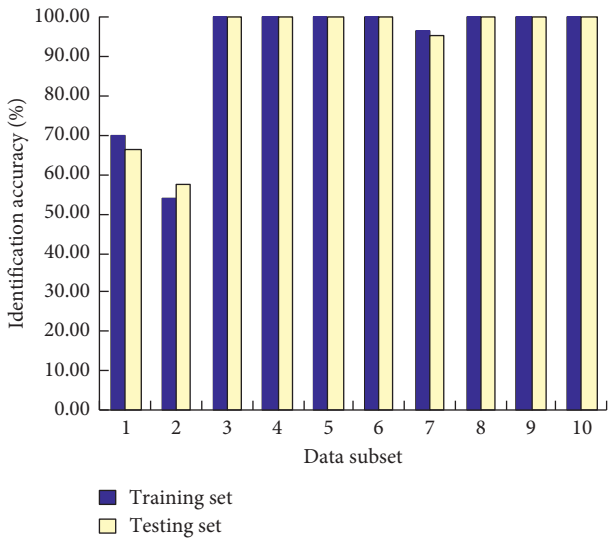

Figure 3: Pathogen identification results of the BPNN models built in the modeling spectral region of 4000–5000 cm^{-1} with a ratio of the training set to the testing set of 3 : 1 and the number of neurons in the hidden layer equal to 50 on each data subset when Data treatment method 2 was used.

to select the optimal modeling spectral region, and then the modeling results can be compared to the results obtained in this study to select the optimal modeling method.

The results obtained in this study demonstrated that the modeling effect may be affected by the spectral data used in modeling. For example, when Data treatment method 2 was used and the BPNN models were built with the modeling ratio of 3 : 1 or 5 : 1 in the spectral region of 4000–5000 cm^{-1}, with the modeling ratio of 4 : 1 in the spectral region of 5000–6000 cm^{-1}, with the modeling ratio of 3 : 1 or 5 : 1 in the spectral region of 6000–7000 cm^{-1}, or with the modeling ratio of 3 : 1 in the spectral region of 8000–10000 cm^{-1}, relatively small averages and large SD values of identification accuracies of the training and testing sets on the 10 data subsets were obtained. This was mainly due to the low identification accuracies of the training and testing sets for the model built on some data subset. In comparison, it was more likely to achieve satisfactory identification results when modeling was conducted in a relatively wide spectral region.

In this study, good identification results were obtained using the three modeling methods including DPLS, BPNN, and SVM. When Data treatment method 1 or Data treatment method 2 was used, a DPLS, BPNN, or SVM model with satisfactory identification accuracies of the training and testing sets could be obtained. In practical applications, one of the modeling methods can be used or more than one modeling methods can be used to synthetically verify identification results, according to the situation. However, in comparison, DPLS is the most stable, followed by SVM then BPNN. In this study, only three modeling methods (DPLS, BPNN, and SVM) were used to build the pathogen identification models. However, many modeling methods for qualitative identification analyses have been reported [15, 22, 24]. In further studies, other modeling methods can be used in an attempt for classification, discrimination, and identification of wheat smut fungi.

In this study, the identification of the three kinds of wheat smut fungi was investigated based on NIRS, aiming to provide a method for the rapid and nondestructive detection and identification of wheat smut fungi. According to the method proposed in this study, the teliospores of related pathogens can be obtained from wheat plants with the symptoms of wheat smuts in the field, then the near-infrared spectra can be acquired using a near-infrared spectrometer in the laboratory, and finally the samples can be identified using the built models. If possible, a portable near-infrared spectrometer can be developed for direct detection of wheat smut fungi in the field. In production practice, wheat smuts are mainly controlled by using seed treatment measures besides utilization of resistant cultivars. Pathogen identification can avoid reserving wheat seeds from the diseased fields and the further transfer of wheat seeds after reservation and can provide a basis for countermeasures. The results of this study provide a methodological reference for performing this task. Once wheat plants are infected by smut fungi, the hyphae expand inside the leaves, and during the

later stage of disease progress, typical symptoms appear and teliospores are produced. Therefore, to perform early detection of the related diseases, studies on the detection of infected wheat plants based on NIRS should be conducted as well as studies on early detection of wheat stripe rust and wheat leaf rust mentioned above [30–32].

Data Availability

The data used to support the findings of this study are included within the article and the supplementary information files.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Yaqiong Zhao and Feng Qin contributed equally to this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (2018YFD0200402), the National Key Technologies Research and Development Program of China (2012BAD19BA04), and the International Research Exchange Scheme of the Marie Curie Program of the 7th Framework Program (ref. PIRSES-GA-2013-612659).

Supplementary Materials

Table S1: pathogen identification results of the DPLS models built with different modeling ratios and numbers of principal components in each modeling spectral region when Data treatment method 1 was used. Table S2: pathogen identification results of the BPNN models built with different modeling ratios and numbers of neurons in the hidden layer in each modeling spectral region when Data treatment method 1 was used. Table S3: pathogen identification results of the DPLS models built on each data subset with different modeling ratios and numbers of principal components in each modeling spectral region when Data treatment method 2 was used. Table S4: pathogen identification results of the BPNN models built on each data subset with different modeling ratios and numbers of neurons in the hidden layer in each modeling spectral region when Data treatment method 2 was used. Table S5: pathogen identification results of the SVM models built on each data subset with different modeling ratios in each modeling spectral region when Data treatment method 2 was used. (*Supplementary Materials*)

References

[1] V. O. Stockwell and E. J. Trione, "Distinguishing teliospores of *Tilletia controversa* from those of *T. Cariesby* fluorescence microscopy," *Plant Disease*, vol. 70, no. 10, pp. 924–926, 1986.

[2] H. G. Wang, H. Y. Zhu, Z. H. Ma, L. Liu, and G. X. Zhang, "The progress and prospect in research on wheat dwarf bunt," *Review of China Agricultural Science and Technology*, vol. 7, no. 4, pp. 21–27, 2005.

[3] W. Dong, W. Q. Chen, and T. G. Liu, "Identification and detection methods of *Tilletia controversa* Kühn," *Plant Protection*, vol. 33, no. 6, pp. 128–131, 2007.

[4] L. Guo, "Taxonomy of smut fungi in China: a brief review," *Mycosystema*, vol. 34, no. 5, pp. 817–820, 2015.

[5] H. X. Yu, L. Gao, X. H. Kang, T. G. Liu, B. Liu, and W. Q. Chen, "Discrimination of teliospores in TCK and TFL with laser scanning confocal microscope," *China Plant Protection*, vol. 36, no. 2, pp. 5–8, 2016.

[6] J. G. Mcdonald, E. Wong, G. T. Kristiansson, and G. P. White, "Direct amplification by PCR of DNA from ungerminated teliospores of *Tilletia* species," *Canadian Journal of Plant Pathology*, vol. 21, no. 1, pp. 78–80, 1999.

[7] G. Bakkeren, J. W. Kronstad, and C. A. Levesque, "Comparison of AFLP fingerprints and ITS sequences as phylogenetic markers in Ustilaginomycetes," *Mycologia*, vol. 92, no. 3, pp. 510–521, 2000.

[8] R. D. Frederick, K. E. Snyder, P. W. Tooley et al., "Identification and differentiation of *Tilletia indica* and *T. walkeri* using the polymerase chain reaction," *Phytopathology*, vol. 90, no. 9, pp. 951–960, 2000.

[9] H. Liang, Y. L. Peng, G. Z. Zhang, W. Q. Chen, and T. G. Liu, "Amplification and sequence analysis of the rDNA intergenic spacer (rDNA-IGS) from three *Tilletia* species," *Acta Phytopathologica Sinica*, vol. 36, no. 5, pp. 407–412, 2006.

[10] Y. Y. Cao, L. F. Wang, L. P. Duan et al., "Development of a real-time fluorescence loop-mediated isothermal amplification assay for rapid and quantitative detection of *Ustilago maydis*," *Scientific Reports*, vol. 7, article 13394, 2017.

[11] K. G. Savchenko, L. M. Carris, J. Demers, D. S. Manamgoda, and L. A. Castlebury, "What causes flag smut of wheat?," *Plant Pathology*, vol. 66, no. 7, pp. 1139–1148, 2017.

[12] N. W. Schaad and R. D. Frederick, "Real-time PCR and its application for rapid plant disease diagnostics," *Canadian Journal of Plant Pathology*, vol. 24, no. 3, pp. 250–258, 2002.

[13] S. Sankaran, A. Mishra, R. Ehsani, and C. Davis, "A review of advanced techniques for detecting plant diseases," *Computers and Electronics in Agriculture*, vol. 72, no. 1, pp. 1–13, 2010.

[14] G. T. Xu, H. F. Yuan, and W. Z. Lu, "Development of modern near-infrared spectroscopic techniques and its applications," *Spectroscopy and Spectral Analysis*, vol. 20, no. 2, pp. 134–142, 2000.

[15] Y. L. Yan, L. L. Zhao, D. H. Han, and S. M. Yang, *Basis and Application of Near-infrared Reflectance Spectroscopy*, China Light Industry Press, Beijing, China, 2005.

[16] Q. Chen, J. Zhao, M. Liu, J. Cai, and J. Liu, "Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 46, no. 3, pp. 568–573, 2008.

[17] M. K. Ahmed and J. Levenson, "Application of near-infrared spectroscopy to the quality assurance of ethanol and butanol blended gasoline," *Petroleum Science and Technology*, vol. 30, no. 2, pp. 115–121, 2012.

[18] X. L. Chu and W. Z. Lu, "Research and application progress of near-infrared spectroscopy analytical technology in China in the past five years," *Spectroscopy and Spectral Analysis*, vol. 34, no. 10, pp. 2595–2605, 2014.

[19] H. Chen, Z. Lin, H. Wu, L. Wang, T. Wu, and C. Tan, "Diagnosis of colorectal cancer by near-infrared optical fiber

spectroscopy and random forest," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 135, pp. 185–191, 2015.

[20] N. Liu, H. A. Parra, A. Pustjens, K. Hettinga, P. Mongondry, and S. M. van Ruth, "Evaluation of portable near-infrared spectroscopy for organic milk authentication," *Talanta*, vol. 184, pp. 128–135, 2018.

[21] M. Lopo, C. A. Teixeira dos Santos, R. N. M. J. Páscoa, A. R. Graça, and J. A. Lopes, "Near-infrared spectroscopy as a tool for intensive mapping of vineyards soil," *Precision Agriculture*, vol. 19, no. 3, pp. 445–462, 2018.

[22] C. Pasquini, "Near-infrared spectroscopy: a mature analytical technique with new perspectives—a review," *Analytica Chimica Acta*, vol. 1026, pp. 8–36, 2018.

[23] P. I. Monteiro, J. S. Santos, V. R. Alvarenga Brizola et al., "Comparison between proton transfer reaction mass spectrometry and near-infrared spectroscopy for the authentication of Brazilian coffee: a preliminary chemometric study," *Food Control*, vol. 91, pp. 276–283, 2018.

[24] D. Granato, P. Putnik, D. B. Kovačević et al., "Trends in chemometrics: food authentication, microbiology, and effects of processing," *Comprehensive Reviews in Food Science and Food Safety*, vol. 17, no. 3, pp. 663–677, 2018.

[25] J. Perez-Mendoza, J. E. Throne, F. E. Dowell, and J. E. Baker, "Chronological age-grading of three species of stored-product beetles by using near-infrared spectroscopy," *Journal of Economic Entomology*, vol. 97, no. 3, pp. 1159–1167, 2004.

[26] J. Perez-Mendoza, J. E. Throne, E. B. Maghirang, F. E. Dowell, and J. E. Baker, "Insect fragments in flour: relationship to lesser grain borer (Coleoptera: bostrichidae) infestation level in wheat and rapid detection using near-infrared spectroscopy," *Journal of Economic Entomology*, vol. 98, no. 6, pp. 2282–2291, 2005.

[27] D. Wu, L. Feng, C. Q. Zhang, and Y. He, "Early detection of gray mold (Cinerea) on eggplant leaves based on vis/near-infrared spectra," *Journal of Infrared and Millimeter Waves*, vol. 26, no. 4, pp. 269–273, 2007.

[28] L. Feng, S. S. Chen, B. Feng, F. Liu, Y. He, and B. G. Lou, "Early detection of soybean pod anthracnose based on spectrum technology," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 28, no. 1, pp. 139–144, 2012.

[29] Y. D. Liu, H. C. Xiao, Q. Deng, Z. C. Zhang, X. D. Sun, and Y. S. Xiao, "Nondestructive detection of citrus greening by near-infrared spectroscopy," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 32, no. 14, pp. 202–208, 2016.

[30] X. L. Li, Z. H. Ma, L. L. Zhao, J. H. Li, and H. G. Wang, "Early diagnosis of wheat stripe rust and wheat leaf rust using near-infrared spectroscopy," *Spectroscopy and Spectral Analysis*, vol. 33, no. 10, pp. 2661–2665, 2013.

[31] X. L. Li, F. Qin, L. L. Zhao, J. H. Li, Z. H. Ma, and H. G. Wang, "Detection of *Puccinia striiformis* f. sp. *tritici* latent infections in wheat leaves using near-infrared spectroscopy technology," *Spectroscopy and Spectral Analysis*, vol. 34, no. 7, pp. 1853–1858, 2014.

[32] Y. Q. Zhao, Y. L. Gu, F. Qin et al., "Application of near-infrared spectroscopy to quantitatively determine relative content of *Puccnia striiformis* f. sp. *tritici* DNA in wheat leaves in incubation period," *Journal of Spectroscopy*, vol. 2017, Article ID 9740295, 12 pages, 2017.

[33] X. L. Li, F. Qin, L. L. Zhao, J. H. Li, Z. H. Ma, and H. G. Wang, "Identification and classification of disease severity of wheat stripe rust using near-infrared spectroscopy technology," *Spectroscopy and Spectral Analysis*, vol. 35, no. 2, pp. 367–371, 2015.

[34] X. L. Li, Z. H. Ma, L. L. Zhao, J. H. Li, and H. G. Wang, "Application of near-infrared spectroscopy to qualitative identification and quantitative determination of *Puccinia striiformis* f. sp. *tritici* and *P. recondita* f. sp. *tritici*," *Spectroscopy and Spectral Analysis*, vol. 34, no. 3, pp. 643–647, 2014.

[35] Y. Q. Zhao, F. Qin, P. Cheng et al., "Quantitative determination of germinability of *Puccinia striiformis* f. sp. *tritici* urediospores using near-infrared spectroscopy technology," *Journal of Spectroscopy*, vol. 2015, Article ID 384214, 8 pages, 2015.

[36] P. Cheng, X. L. Li, F. Qin et al., "Effects of UV-B radiation on near-infrared spectroscopy and identification of *Puccinia striiformis* f. sp. *tritici*," *Journal of Spectroscopy*, vol. 2014, Article ID 751458, 9 pages, 2014.

[37] C. C. Chang and C. J. Lin, "LIBSVM," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[38] L. H. Purdy, "Flag smut of wheat," *The Botanical Review*, vol. 31, no. 4, pp. 565–606, 1965.