agriculture

MDPI

Article

Machine Learning Identification of Saline-Alkali-Tolerant Japonica Rice Varieties Based on Raman Spectroscopy and Python Visual Analysis

Rui Liu [1], Feng Tan [2,*], Yaxuan Wang [3], Bo Ma [4], Ming Yuan [4], Lianxia Wang [4] and Xin Zhao [5]

[1] College of Agricultural Engineering, Heilongjiang Bayi Agricultural University, Daqing 163000, China; liurui@byau.edu.cn
[2] College of Electrical and Information, Heilongjiang Bayi Agricultural University, Daqing 163000, China
[3] College of Civil Engineering and Water Conservancy, Heilongjiang Bayi Agricultural University, Daqing 163000, China; wangyaxuan1980@byau.edu.cn
[4] Qiqihar Branch of Heilongjiang Academy of Agricultural Sciences, Qiqihar 161006, China; mabo8210@haas.cn (B.M.); y.m@haas.cn (M.Y.); wlx0427@haas.cn (L.W.)
[5] College of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China; zxxsnh@hrbeu.edu.cn
* Correspondence: bayitf@byau.edu.cn

Abstract: The core of saline-alkali land improvement is planting suitable plants. Planting rice in saline-alkali land can not only effectively improve saline-alkali soil, but also increase grain yield. However, traditional identification methods for saline-alkali-tolerant rice varieties require tedious and time-consuming field investigations based on growth indicators by rice breeders. In this study, the visualization method of Python data processing was used to analyze the Raman spectroscopy of japonica rice in order to study a simple and efficient identification method of saline-alkali-tolerant japonica rice varieties. Three saline-alkali-tolerant japonica varieties and three saline-alkali-sensitive japonica varieties were collected from control and saline-alkali-treated fields, respectively, and the Raman spectra of 432 samples were obtained. The data preprocessing stage used filtering-difference method to process Raman spectral data to complete interference reduction and crests extraction. In the feature selection stage, scipy.signal.find_peaks (SSFP), SelectKBest (SKB) and recursive feature elimination (RFE) were used for machine feature selection of spectral data. According to the feature dimension obtained by machine feature selection, dataset partitioning by K-fold CV, the typical linear logistic regression (LR) and typical nonlinear support vector machine (SVM) models were established for classification. Experimental results showed that the typical nonlinear SVM identification model based on both RFE machine feature selection and six-fold CV dataset partitioning had the best identification rate, which was 94%. Therefore, the SVM classification model proposed in this study could provide help in the intelligent identification of saline-alkali-tolerant japonica rice varieties.

Keywords: japonica rice; saline-alkali-tolerant; Raman spectroscopy; Python visual; RFE; typical nonlinear; SVM

Citation: Liu, R.; Tan, F.; Wang, Y.; Ma, B.; Yuan, M.; Wang, L.; Zhao, X. Machine Learning Identification of Saline-Alkali-Tolerant Japonica Rice Varieties Based on Raman Spectroscopy and Python Visual Analysis. *Agriculture* **2022**, *12*, 1048. https://doi.org/10.3390/agriculture12071048

Academic Editor: Yanbo Huang

Received: 20 May 2022
Accepted: 16 July 2022
Published: 18 July 2022

1. Introduction

Northeast China is the typical concentrated distribution area of soda saline-alkali land, with an area of 7.65 million hm^2 and an annual growth rate of 1.4% [1]. Soil saline-alkali barrier and salinization are two of the main limiting factors of current agricultural production, and the improvement and utilization of saline-alkali land is of great significance in increasing reserve resources of cultivated land and improving ecological environment [2]. Rice is a moderately saline-alkali-tolerant plant that can accelerate the desalination of saline-alkali soil and the accumulation of organic matter and realize the improved utilization of saline-alkali land by taking advantage of its unique advantages of saline-alkali tolerance

and aquatic irrigation [3,4]. Therefore, researchers regard the identification of saline-alkali-tolerant rice varieties as an important task.

At present, there are two main methods for identification of saline-alkali-tolerant rice varieties, including traditional phenotypic index method and molecular QTL labeling method. Some Chinese researchers carried out field tests to screen saline-alkali-tolerant rice varieties according to rice yield indexes [5–8], and some researchers in other countries also used the phenotypic index data of rice to conduct statistical analysis to screen saline-alkali-tolerant rice varieties [9–11]. The source of field investigation and phenotype data mainly depended on the experience of breeders. The data collection process was not only cumbersome and time-consuming, but also difficult to unify and standardize. Many researchers used molecular QTL markers to analyze rice saline-alkali tolerance based on phenotypic indicators of rice cultivation [11–14]. It was a complex process to use molecular QTL markers to locate saline-alkali-tolerant rice varieties; the high cost of gene detection limits the development of large-scale molecular detection. With the development of science and technology, the design and construction of a simple, accurate, and intelligent identification method for saline-alkali-tolerant of rice is of great significance for improving saline-alkali land.

Raman spectroscopy technology can provide rapid, simple, repeatable, and non-destructive qualitative and quantitative analysis without sample preparation, which can be directly measured by laser [15]. Raman spectroscopy qualitative analysis method includes linear and nonlinear calibration methods. Python is a computer programming language with strong operability, easy-to-use, and full-featured tools, which is widely used in data analysis [16]. Among them, logistic regression (LR) is a typical linear calibration method [17] and support vector machine (SVM) is a typical nonlinear calibration method [18]. Therefore, LR and SVM methods in Python were selected in this paper for qualitative analysis of Raman spectral data. However, when applying LR and SVM for modeling of spectral data, because the interference of instrument noise, stray light, fluorescence background, and the dimension of original spectral data are too high, this will lead to the problem of long model running time and affect the model accuracy. Some researchers [19,20] used the static tools Matplotlib and Seaborn in the Python visualization library to visualize data from different neighborhoods and analyze the results of the visualization. Matplotlib is one of the most popular data visualization libraries in Python. It can support 2D and 3D diagrams, and is an important tool for data analysis and visualization. Seaborn is based on Matplotlib for a higher level of API encapsulation, and the drawing interface is more integrated, which makes drawing easier [21]. Therefore, this experiment was based on the visual features of Python data processing, and the spectral data were subjected to interference reduction and dimensionality reduction processing before establishing a recognition model.

At present, the application of Raman spectroscopy combined with Python visual in the identification of saline-alkali-tolerant Japonica rice varieties has not been reported. In this study, Raman spectrometer was used to obtain molecular information of japonica rice varieties, and Python data analysis and visualization method (reduce interference, extract crest, reduced feature dimension, dataset partitioning, and classification model) was used to identify saline-alkali-tolerant japonica rice varieties, trying to establish a fast, convenient, economic, and accurate classification model of saline-alkali-tolerant japonica rice varieties.

2. Materials and Methods

2.1. Sample Preparation

The test materials were collected from the rice breeding experimental field of Qiqihar Branch of Heilongjiang Academy of Agricultural Sciences in September 2021. A total of 6 japonica rice varieties were tested, including 3 saline-alkali-tolerant varieties and 3 saline-alkali-sensitive varieties, respectively, which were planted in non-saline-alkali soil (control field) and saline-alkali soil (saline-alkali stress field) [22–24]. Samples collected from control fields were called control samples, and those collected from salt-alkali stress fields were

called treated samples [25,26]. Twelve holes were taken from each japonica rice variety in control field and salt-alkali stress field, for a total of 144 holes.

As shown in Table 1, the 144 holes materials obtained from the experimental field were placed in the laboratory at 25 °C for 15 days drying. Three ears of grain were taken from different positions in each hole, with 10 grains per ear of grain. The Shanghai Superstar LJJM milled rice machine was used for one-time shelling of 50 s, and 36 seeds with complete appearance after shelling were selected from each variety in each field as samples. A total of 432 samples were obtained from 6 varieties in the two experimental fields.

Table 1. Variety and quantity of test samples (1: saline-alkali-tolerant, 0: saline-alkali-sensitive, CS: control sample, TS: treated sample).

Number of Varieties	Sample	Variety of Sample	Number of CS	Number of TS	Total
1	QJ10	1	36	36	72
2	BD6	1	36	36	72
3	DF132	1	36	36	72
4	LJ12	0	36	36	72
5	KD42	0	36	36	72
6	LD107	0	36	36	72

2.2. Obtaining Spectral Information

The sample image information was obtained using Advantage 532 Raman spectrometer (excitation power was less than 5 mW, the resolution was 1.4 cm^{-1}, the measurement range was 200~3400 cm^{-1} and the scanning was four times) combined with Pro Scope HR software (Table 2), and saved in PRN format, from 1–2 November 2021, at room temperature of 25 °C. Each sample obtained 3201 Raman spectral information; four hundred and thirty-two samples obtained a total of 1,382,832 Raman spectral information.

Table 2. Pro Scope HR set parameters for obtaining sample image information.

Laser Power	Integration Time	Number of Spectrum	Display	Save Spectrum	Resolution
High	4	3	Average	ASCII	Low

2.3. Reduce Interference and Extract Crest

Extracting the original data from the database, as shown in Figure 1, four hundred and thirty-two spectral pieces of information were interlaced in disorder and difficult to distinguish. Due to the interference of instrument noise, stray light, and fluorescence background, the data accuracy was affected when collecting spectral data. Therefore, it is necessary to deal with the disturbance reduction of the original Raman spectral data.

According to the effective Raman shift of Raman spectrum 200–3400 cm^{-1}, the filtering method was used for data noise reduction and impurity removal. Constructing the filter using signal.butter, the parameter settings is as follows (1).

$$b, a = scipy.signal.butter(N, Wn) \quad (1)$$

Note: a is the denominator coefficient vector of the filter, b is the numerator coefficient vector of the filter. N is the order of the filter, Wn is the critical frequency or frequencies.

In this test, the N parameter was 2 (one-step forward and one-step backward filtering to avoid phase difference) and Wn parameter was 0.002. Compared with the original data curve (Figure 2), the filtered curve was obviously smooth. The filtering method not only filtered out various interference information of original spectral data, but also filtered out useful spectral crest information. Therefore, it is essential to extract spectral crest information. The difference method was used to extract spectral crest information, and the

intensity of wave peak can be extracted by subtracting the spectral intensity after filtering from the original data of the same sample, as shown in Formula (2):

$$y_n^{**} = y_n - y_n^* \ (1 \le n \le 3201) \quad (2)$$

Note: y is the wave intensity of original data, y^* is the wave intensity after filtering, and y^{**} is the wave intensity after filtering difference.

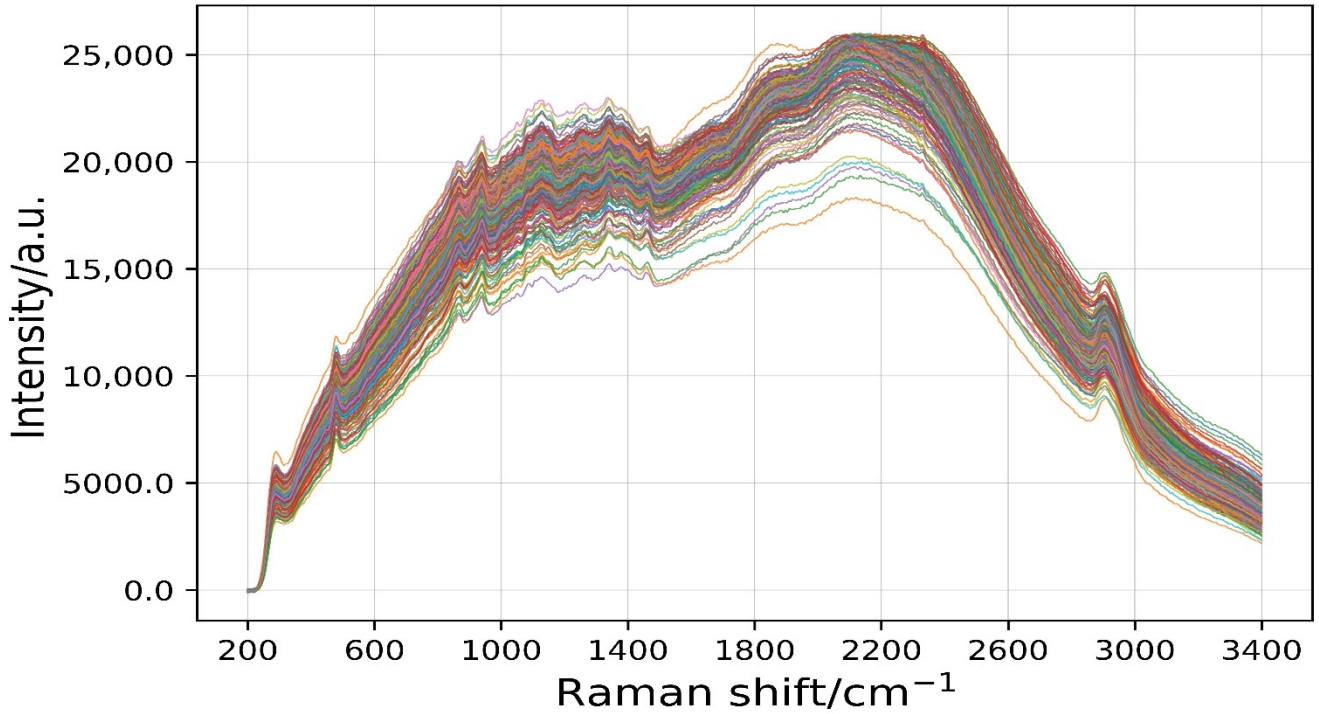

Figure 1. Four hundred and thirty-two curve of original Raman spectrum by Pro Scope HR. Raman shift is the reciprocal of wavelength, and its range is 200–3400 cm^{-1}. Intensity is the intensity of Raman scattering.

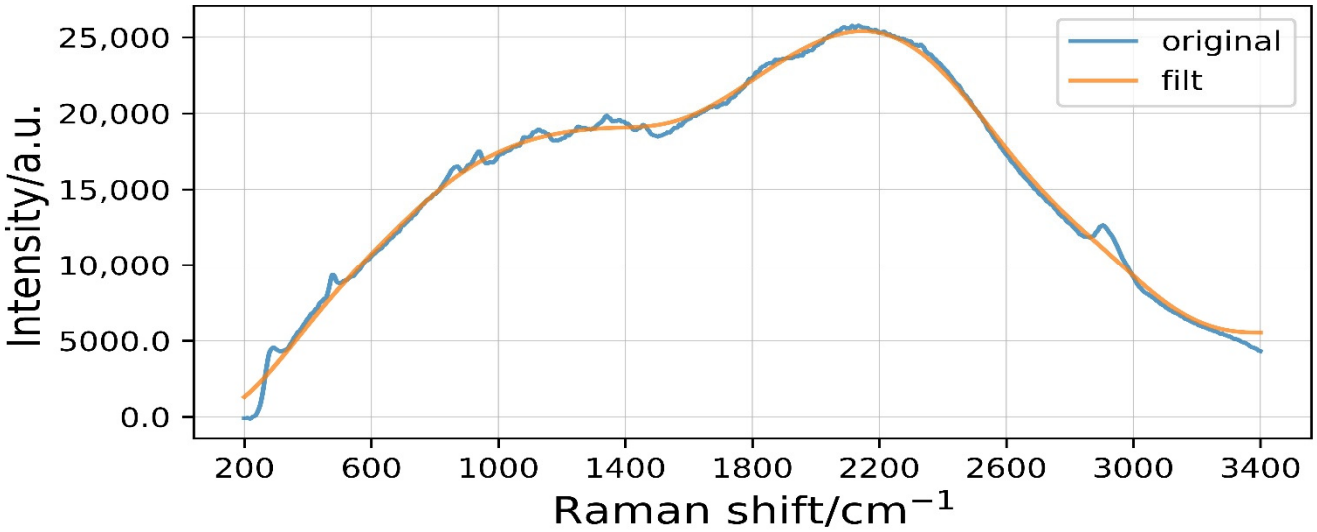

Figure 2. Comparison between filtering curve and original curve. Blue line is original curve, orange line is filtering curve.

As shown in Figure 3, the full Raman shift of the Raman spectrum is clearly visible, each japonica rice variety shows 12 significant crest points and many nonsignificant crest

points. Combined with crest extraction and rice Raman characteristics and their attribution [27,28], effective Raman shift of 200–1800 cm^{-1} and 2800–3200 cm^{-1} were selected. Seven effective crests at 480 cm^{-1}, 865 cm^{-1}, 941 cm^{-1}, 1129 cm^{-1}, 1339 cm^{-1}, 1461 cm^{-1}, and 2910 cm^{-1} were extracted, and each japonica rice variety had different peak intensities near the same Raman shift. Seven effective crests were extracted by filtering difference method, which laid a foundation for the next crest feature extraction.

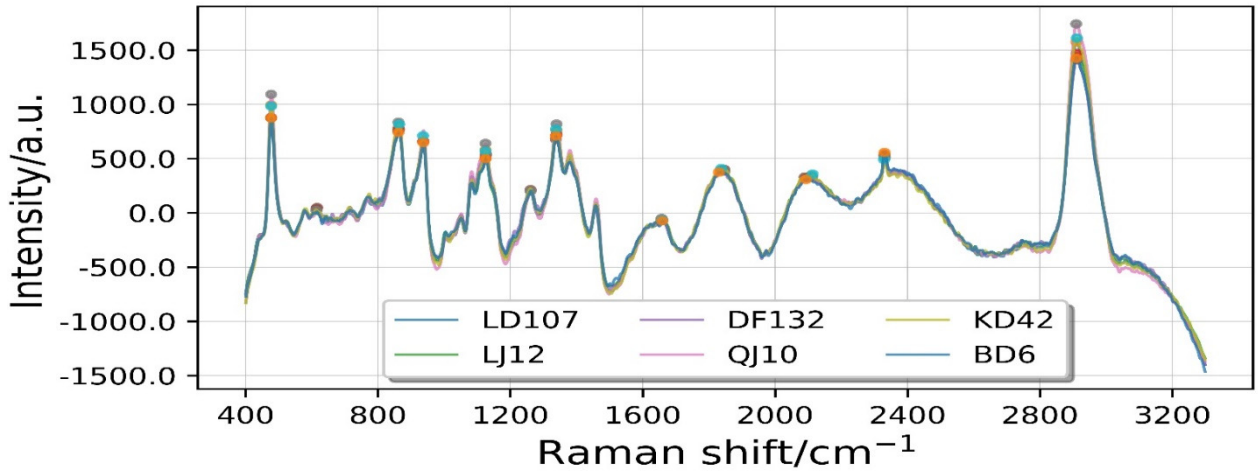

Figure 3. The peak crest of Raman spectrum was extracted by filtering difference.

2.4. Reduced Feature Dimension

2.4.1. Dimension Reduction by Scipy.Signal.Find_Peaks (SSFP)

Seven effective crests were selected based on the filtering difference method and the fingerprint identification information of rice Raman spectrum. The machine learning function of scipy.signal.find_peaks (SSFP) [29] in Python was used to automatically detect the characteristic information of crest. To further understand the information about crests, using the peak_prominences function, peak_widths function, width_height function (width_height), and peak_dif function, we calculated the prominence, width, height of width, and offset of each crest that passed the filtering difference. Four kinds of characteristic pieces of information of crests were detected, as shown in Figure 4.

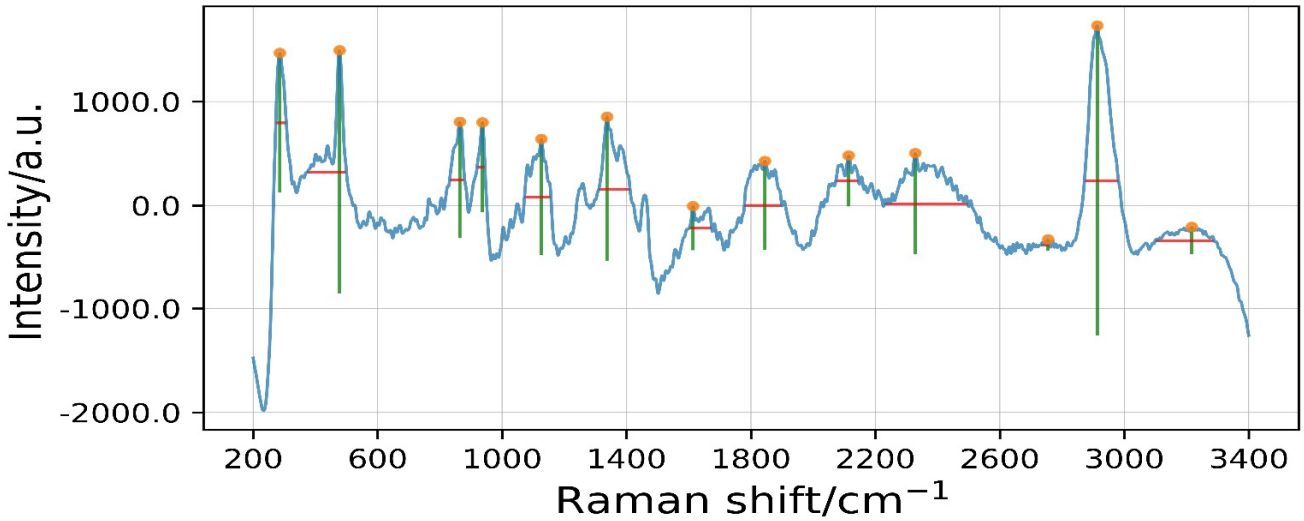

Figure 4. Four-dimensional characteristics of crest. The prominence is the height of the crest, the length of the green vertical line in the figure. The width is the width of the crest, the width of the red line in the figure. The width_height is the height of the crest width, the length from the red line to the peak. The peak_dif is the offset degree of Raman shift.

2.4.2. Dimension Reduction by SelectKBest (SKB)

SKB is one of the methods for automatically selecting feature variables in sklearn, which is famous for its large variable feature selection tool. The working principle of SKB is to use a certain parameter to score certain features and select the best k most powerful feature information. The SKB method in this test selects mutual_info_regression (MIR) parameters and uses MIR algorithm [30] to score the 28-dimensional feature information without dimensionality reduction, as shown in formula and parameter settings (3):

$$\begin{gathered} I(X;Y) \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\ selector = SelectKBest(mutual\_info\_regression, k = 10) \end{gathered} \tag{3}$$

Note: MIR algorithm is used to evaluate the correlation between category-independent variables and category-dependent variables.

2.4.3. Dimension Reduction by Recursive Feature Elimination (RFE)

The main idea of RFE is to repeatedly construct the features of the model [31], eliminate the redundancy between features, select the optimal feature combination, and reduce the feature dimension. Firstly, the original 28 features as the initial feature subset were input into the RFC classifier [32], the importance of each feature was calculated, and the classification accuracy of the initial feature subset 1 was obtained by cross-validation method. Second, from the current feature subset, characteristic features of lowest importance were removed, obtaining a new feature subset 2, which was input to the RFC model [33]. Again, the classification accuracy of the initial feature subset 2 was obtained, and recursive was repeated from the RFC classifier to the features of importance to the cross-validation method to obtain a new subset classification accuracy method, until the feature subset was empty. Finally, a total of K feature subsets with different feature numbers were obtained, and the feature subset with the highest classification accuracy was selected as the optimal feature combination. Therefore, this is a greedy algorithm to search for the optimal feature subset, as shown in parameter settings (4).

$$rfe\_select = RFE(estimator = RandomForestClassifier(), step = 1) \tag{4}$$

2.4.4. Dataset Partitioning by K-Fold Cross-Validation (K-Fold CV)

Figure 5 showed the distribution of the 432-sample dataset based on three important features. Except for the wave crest at 941 cm^{-1}, the sample distribution of the other six crests was obviously a linear non-separable datas.

The implementation of the K-fold CV algorithm is to divide the dataset into K equal sample subsets, and then traverse the K subsets in turn. The i (i = 1, 2, . . . , K) traverse will take the i subset as the test set, and all the other subsets as the training set for the training and evaluation of the model. Finally, the average value of K evaluation indexes is taken as the final evaluation index. The larger the K value is, the larger the training set used by the training model is, the less susceptible to noise, and the better the performance of the model is [34–36]. In view of the fact that each japonica rice variety sample in this study was 72, the dataset was divided into six-fold CV method, and the parameter settings were as follows (5).

$$kf = KFold(n\_splists = 6) \tag{5}$$

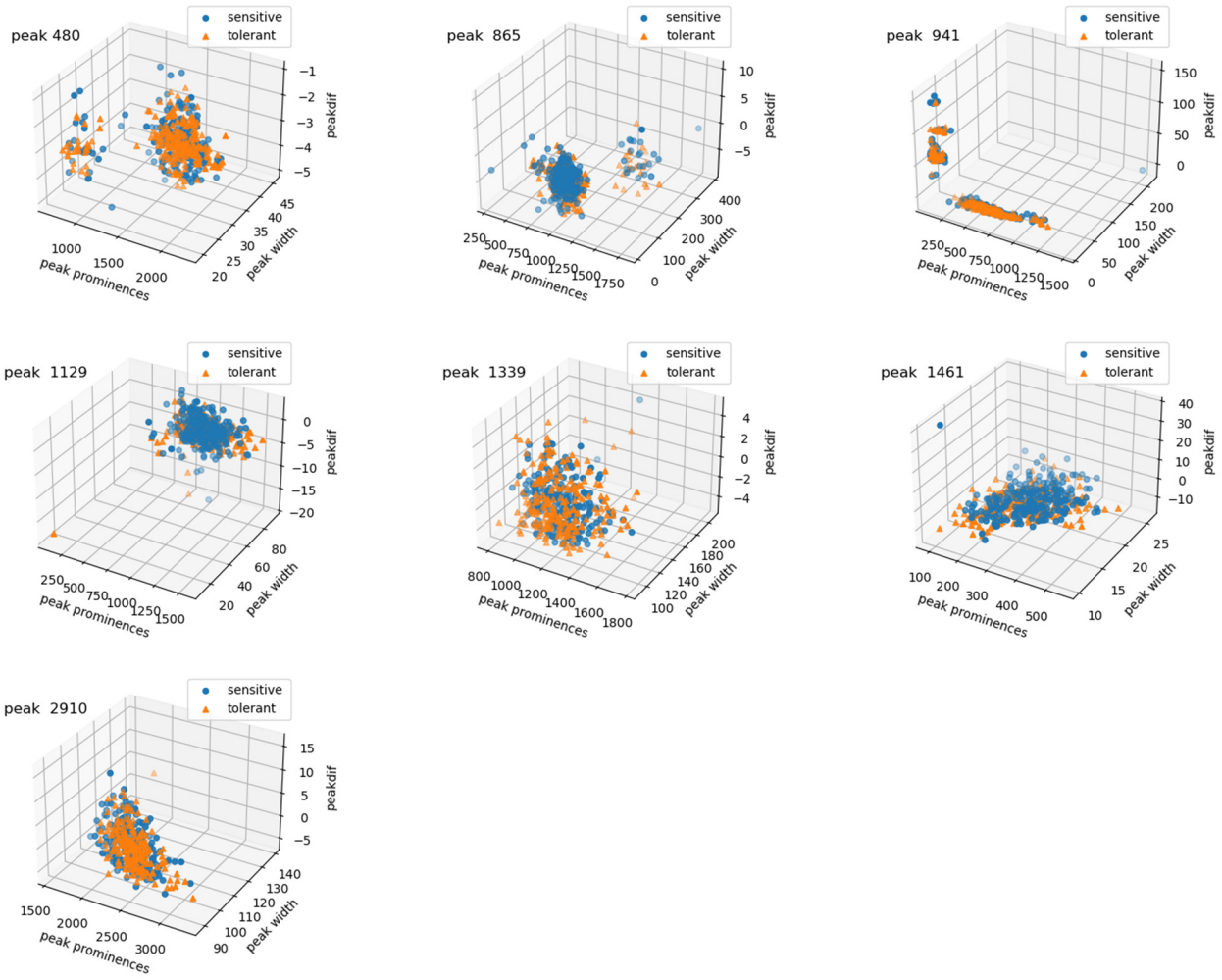

Figure 5. The sample 3D characteristic distribution of 7 wave crests at 480 cm^{-1}, 865 cm^{-1}, 941 cm^{-1}, 1129 cm^{-1}, 1339 cm^{-1}, 1461 cm^{-1}, and 2910 cm^{-1}, respectively. Yellow triangles represent samples of saline-alkali-tolerant japonica rice varieties from control and treated fields, blue circles represent samples of saline-alkali-sensitive japonica rice varieties from control and treated fields.

2.5. Identification Models Evaluate Feature Selection Methods

2.5.1. Typical Linear LR Identification Model

LR is a classification model in machine learning of a classification algorithm. Although the name has regression, it has a certain connection with regression. Due to the simplicity and efficiency of the algorithm, it is widely used in practice [37,38]. LR's input function is the result of a linear regression, as shown in Formula (6):

$$h(w) = w_1x_1 + w_2x_2 + w_3x_3 \ldots + b \tag{6}$$

Input the result of a linear regression into the sigmoid function, as shown in Formula (7):

$$Sigmoid\ Function: g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{7}$$

The sigmoid function output results are in the interval of [0, 1], in which the default machine threshold is a 0.5 function.

2.5.2. Typical Nonlinear SVM Identification Model

In the case of linear inseparability, SVM uses nonlinear mapping algorithm to transform the low-dimensional input space linearly inseparable samples into high-dimensional

feature space to make them linearly separable, thus making it possible to use linear algorithm in high-dimensional feature space to perform linear analysis on the nonlinear features of samples [39]. The solution is to map them to a higher dimensional space, but the difficulty of this approach is the increase of computational complexity, and the kernel function neatly solves this problem. Therefore, the kernel function is the key to the SVM model. The kernel list parameter of the SVM model is *rbf*, and radial basis function kernel (*RBF*) is suitable for the linear non-fractional dataset, as shown in formula and parameter settings (8):

$$\begin{gathered} RBF:\ exp\left(-\frac{1}{2\sigma^2}||X-X_i||^2\right) \\ kernel =' rbf',\ random\_state = 1,\ \max\_iter = -1, tol = 1e-4 \end{gathered} \tag{8}$$

RBF can map the samples to a higher dimensional space and process the sample when the relationship between class labels and features is nonlinear.

3. Results and Analysis

3.1. Analysis of Characteristic Information Extraction of Crest

Figure 4 shows that the SSFP function was used to extract four characteristic pieces of information of wave crest (prominence, width, width_height, and peak_dif). The shape characteristic information of wave crest were prominences and width, and the position characteristic information of wave crest were width_height and peak_dif. The SSFP automatic detection method extracts four-dimensional feature information for each wave peak, which could accurately lock the shape, position, and change of each wave peak.

Seven crests were extracted from each sample, and each crest had four-dimensional characteristic information. Therefore, each sample had 28-dimensional characteristic information. Each sample was normalized vertically for prominences, width, and width_height, while peak_dif was normalized horizontally. The normalized feature information was in the range [0, 1]. All samples were labeled based on saline-alkali-tolerant japonica variety marker 1 and saline-alkali-sensitive japonica variety marker 0, and 432 sample datasets with labeled 28-dimensional characteristic information were obtained.

3.2. Analysis of Selection of Features

Table 3 shows the filtering-difference method combined with crest extraction and rice Raman characteristics, and their attribution were used to extract seven effective wave peaks. Four-dimensional characteristic information of each crest was selected by SSFP function; each sample obtained 28-dimensional characteristic information. If 432 labeled samples with 28-dimensional feature information are directly brought into the classification model for machine learning, the large feature information matrix will lead to problems such as large amount of calculation and long recognition time. Therefore, it is essential to reduce the dimension of feature information and select the best feature information.

Table 3. Results of 3 feature information selection methods (p is prominences, w is width, wh is width_height, and pd is peak_dif).

Method	Raman Shift Position for Peak Extraction (Raman Shift/cm^{-1})							Total
	480	865	941	1129	1339	1461	2910	
SSFP	p\w\wh\pd	p\w\wh\pd	p\w\wh\pd	p\w\wh\pd	p\w\wh\pd	p\w\wh\pd	p\w\wh\pd	28
SKB	p\pd	p\pd	pd	pd	pd	wh	p\w	10
RFE	w\pd	pd	p\wh\pd	pd	p\wh\pd	p\w\pd	wh	14
Total	8	7	8	6	8	8	7	52

3.2.1. Features Selected by SKB

Through the MIR algorithm in SKB, the 10 most powerful pieces of feature information were finally selected, as shown in Table 3. The SKB method selected 10-dimensional

characteristic information of 7 crests. There were three crests of p-characteristic, one crest of w-characteristic, one crest of wh-characteristic, and five crests of pd-characteristic. There are two-dimensional characteristic pieces of information at 480 cm^{-1}, 865 cm^{-1}, and 2910 cm^{-1}, and one-dimensional characteristic information at the other four Raman shifts. Compared with the SSFP 28-dimensional feature information extraction, the feature information selection rate of the SKB method was 36%.

3.2.2. Features Selected by RFE

Table 3 shows that the RFE method selected 14-dimensional characteristic information of seven crests. There were three crests of p-characteristic, two crests of w-characteristic, three crests of wh-characteristic, and six crests of pd-characteristic. There are pieces of three-dimensional characteristic information at 941 cm^{-1}, 1339 cm^{-1}, and 1461 cm^{-1}, two-dimensional characteristic information at 480 cm^{-1}, and one-dimensional characteristic information at the other three Raman shifts. Compared with the SSFP 28-dimensional feature information extraction, the feature information selection rate of the RFE method is 50%.

Two methods of feature information selection were used to reduce the dimension of feature information extracted from SSFP, which solves the problem of reducing feature matrix and reducing computing time. Whether the method of selecting effective characteristic information can accurately identify saline-alkali-resistant japonica rice varieties requires the classification model to evaluate the validity of selecting characteristic information.

3.3. Performance Analysis of Models

Based on one feature information extraction and two feature selection methods, 28 dimensions of feature information were extracted by SSFP method, 10 dimensions of feature information were selected by SKB method, and 14 dimensions by RFE method. The data of 432 labeled samples of six japonica rice varieties were divided according to six-fold CV method of the same japonica rice variety, with five subsets as training sets and one subset as test set, and the sample sets were divided six times in total. Based on six-fold CV, the results of one feature information extraction and two feature selection methods were brought into the classification model, respectively, and then confusion matrix [40,41] was carried out to evaluate the feature information selection methods, so as to seek a fast, convenient, economic, reliable, and accurate classification model of saline-alkali-resistant japonica rice varieties.

3.3.1. Performance of Typical Linear LR Classification Model

Based on one feature information extraction by SSFP, and two feature selections by SKB and RFE method, the average accuracy rate of six-fold CV (Table 4) was 91.44%, 92.13%, and 91.67%, and the average precision rate of six-fold CV (Table 4) was 90.56%, 91.45%, and 89.36%, respectively. The results showed that there was no significant difference in both the accuracy and precision rate of the classification models selected with the three kinds of feature information. Based on the three feature selections, the average accuracy rate was lower than the average precision rate, respectively; the typical nonlinear properties of the experimental dataset greatly interfered with the classification performance of the LR model. Although SKB feature selection method had the lowest number of feature selection (10) and high accuracy rate (92.13%), the value of the average precision (91.45%) was lower than the value of the average accuracy (92.13%), which could lead to the decrease of model stability and model generalization ability.

Table 4. Confusion matrix comparison of datasets of LR classification models based on three feature selection methods. TN: true negatives are the number of predictions where a sample of a 0 is correctly classified (as 0); FP: false positives are the number of predictions where a sample of a 0 is incorrectly classified as a 1; FN: false negatives are the number of predictions where a sample of a 1 is incorrectly classified as a 0; TP: true positives are the number of predictions where a sample of a 1 is correctly classified (as a 1).

Feature Selection	Test Dataset					LR Classification Model	
	Test Data Subset	TN (0,0)	FP (0,1)	FN (1,0)	TP (1,1)	Accuracy (%)	Precision (%)
SSFP	1	32	4	2	34	0.9167	0.8947
	2	32	4	3	33	0.9028	0.8919
	3	33	3	4	32	0.9028	0.9143
	4	34	2	2	34	0.9444	0.9444
	5	30	6	4	32	0.8611	0.8421
	6	34	2	1	35	0.9583	0.9459
	The average					0.9144	0.9056
SKB	1	33	3	1	35	0.9444	0.9211
	2	31	5	2	34	0.9028	0.8718
	3	32	4	3	33	0.9028	0.8919
	4	34	2	4	32	0.9167	0.9412
	5	31	5	5	31	0.8611	0.8611
	6	36	0	0	36	1	1
	The average					0.9213	0.9145
RFE	1	33	3	3	33	0.9167	0.9167
	2	31	5	1	35	0.9167	0.8750
	3	30	6	2	34	0.8889	0.8500
	4	33	3	0	36	0.9583	0.9231
	5	29	7	3	33	0.8611	0.8250
	6	35	1	2	34	0.9583	0.9714
	The average					0.9167	0.8935

3.3.2. Performance of Typical Nonlinear SVM Classification Model

Based on one feature information extraction by SSFP, and two feature selections by SKB and RFE method, the average accuracy rate of six-fold CV (Table 5) was 93.27%, 93.06%, and 93.98%, and the average precision rate of six-fold CV was 93.53%, 93.84%, and 95.66%, respectively. The results showed that there was no significant difference in both the accuracy and precision rate of the classification models selected with the three kinds of feature information. Based on the three feature selection, the average accuracy rate of the typical nonlinear SVM classification model was higher than that of the typical linear LR classification model, respectively, and the average accuracy rate was higher than the average precision rate, respectively; the SVM classification model is obviously suitable for the linear non-fractional dataset. The RFE feature selection method had the highest average accuracy rate, and the average precision rate was higher than the average accuracy rate, which could make the model more stable with stronger generalization ability.

Table 5. Confusion matrix comparison of datasets of SVM classification models based on three feature selection methods.

Feature Selection	Test Dataset					SVM Classification Model	
	Test Data Subset	TN (0,0)	FP (0,1)	FN (1,0)	TP (1,1)	Accuracy (%)	Precision (%)
SSFP	1	34	2	2	34	0.9444	0.9444
	2	34	2	4	32	0.9167	0.9412
	3	30	6	5	31	0.8472	0.8378
	4	35	1	0	36	0.9861	0.9730

Table 5. *Cont.*

Feature Selection	Test Dataset					SVM Classification Model	
	Test Data Subset	TN (0,0)	FP (0,1)	FN (1,0)	TP (1,1)	Accuracy (%)	Precision (%)
	5	35	1	3	32	0.9437	0.9697
	6	34	2	1	35	0.9583	0.9459
			The average			0.9327	0.9353
SKB	1	33	3	3	33	0.9167	0.9167
	2	34	2	4	32	0.9167	0.9412
	3	32	4	6	30	0.8611	0.8824
	4	36	0	1	35	0.9861	1
	5	34	2	1	35	0.9583	0.9459
	6	34	2	2	34	0.9444	0.9444
			The average			0.9306	0.9384
RFE	1	35	1	5	31	0.9167	0.9688
	2	34	2	3	33	0.9306	0.9429
	3	34	2	5	31	0.9028	0.9394
	4	35	1	0	36	0.9861	0.9730
	5	35	1	2	34	0.9583	0.9714
	6	34	2	2	34	0.9444	0.9444
			The average			0.9398	0.9566

4. Discussion

The predecessors conducted many studies on rice origin tracing or variety attribute category, and the test materials included multiple varieties from the same origin and the same variety from multiple origins [42–44]. The test materials in this study were three saline-alkali-tolerant varieties and three saline-alkali-sensitive varieties, and 432 samples were collected from the control field and the saline-alkali-treated field, respectively. Saline-alkali-tolerant japonica rice variety was labeled as 1 and saline-alkali-sensitive japonica rice variety was labeled as 0 in the training set and test set. Both 0 and 1 contained samples of control field and samples of saline-alkali stress field. In view of the diversity of samples in the training set, it can avoid the underfitting situation that often occurs in machine deep learning [36,37]. In view of the extensiveness of samples in the test set, this can avoid the overfitting situation in the future application of the model [45,46]. Therefore, the diversity and extensiveness of materials selection in this study effectively avoids the fitting situation of the identification model dataset of saline-alkali-resistant japonica rice varieties.

If 432 labeled samples with 28-dimensional feature information are directly brought into the classification model for machine learning, the large feature information matrix will lead to problems such as large amount of calculation and long recognition time [47,48]. The two feature information selection methods in this experiment are machine learning methods based on Python programming software database to automatically select the best feature information; compared with the SSFP feature extraction method, SKB and RFE feature selection methods reduce the dimension of feature information by more than 50%. Compared with the SSFP feature extraction and SKB feature selection methods, the RFE feature selection method had the highest recognition rate in SVM classification models (94%). The results showed that the machine feature information selection method could effectively reduce the dimension of feature information, reduce the model identification time, and improve the model classification ability.

LR is a classification model and a typical linear classification method in machine learning whose input function is the result of a linear regression [20,37,38]. The RBF kernel function is selected by the SVM model to divide the linear non-separable dataset, which is good at mapping the linear non-separable dataset to the high-dimensional space to achieve the division of multidimensional planes [21,46]. In this test, based on the seven crests, three important features were selected for each wave peak, and the three-dimensional distribution maps of the seven wave peaks were established, respectively (Figure 5). Except

for the wave crest at 941 cm^{-1}, the sample distribution of the other six crests was obviously a linear non-separable dataset. In view of this, the dataset of this experiment was a typical nonlinear dataset, and only SVM, which is a typical nonlinear classification method, could be used for accurate identification of saline-alkali-tolerant japonica rice varieties, obtaining a 94% accuracy rate.

5. Conclusions

A Raman spectrometer can obtain data information of saline-alkali-tolerant japonica rice varieties, and Python data visualization analysis provides objective and effective choices for the classification of saline-alkali-tolerant japonica rice varieties. In this research, we propose a machine learning classification and detection method based on Raman spectroscopy and Python visual analysis that can be applied to achieve the saline-alkali-tolerant japonica rice varieties' identification gained by typical nonlinear SVM model. In order to make up for the deficiencies brought about by feature selection, two feature selection methods were selected to improve the identification efficiency. The results of the test dataset show that the use of the RFE classification was the best, the RFE classification was better than the SKB, and the accuracy of the RFE classification was nearly 94%.

After using the trained model to predict the unknown samples, the results show that the RFE–SVM analysis method was the best model and it reached the expected prediction. In addition, we discovered that the sample dataset of this experiment is typical nonlinear, which is the same that the previous researchers recognized (SVM is good at mapping the linear non-separable dataset to the high-dimensional space to achieve the division of multidimensional planes). The RFE–SVM combination was used to identify saline-alkali-tolerant japonica rice varieties; the fourth subset of six-fold CV was the best test set, with the highest accuracy value (0.9861) and the highest precision value (0.9730) at the same time.

Through the development of science and technology, image Raman spectrometer technology will be used to obtain data information of saline-alkali-tolerant rice varieties. By providing more experimental data, image recognition and data analysis of Raman spectroscopy data will be performed in the future through machine deep learning models suitable for large sample datasets. Combining the image Raman spectroscopy technique with the machine deep learning model will provide a saline-alkali-tolerant japonica rice varieties identification model with stronger generalization ability.

Author Contributions: R.L. and F.T. conceived the study and designed the project. R.L. and X.Z. performed the experiment, analyzed the data, and drafted the manuscript. B.M., Y.W., M.Y. and L.W. helped to revise the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Natural Science Fund Key Project of Heilongjiang Province (ZD2019F002), Heilongjiang Bayi Agricultural University Initiation Plan for Introducing Talents for Academic Achievement (XDB2013-18), Heilongjiang Bayi Agricultural University Support Program for San Heng San Zong (ZRCPY202120), the Scientific Research Project of Heilongjiang Provincial Scientific Research Institutes of China (2021YYYF011), and 2020 Daqing City Directive Science and Technology Project (zd-2020-68).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors hereby declare that there are no conflict of interest in the present study.

References

1. Li, M.; Zhang, Y.H. Effects of different fertilization patterns on the bacterial community dynamic in saline-alkali paddy soil. *Agric. Res. Arid Areas* **2018**, *36*, 142–148. [CrossRef]
2. Han, G.Q.; Zhou, L.R. *Improvement and Utilization of Saline Soil in Herlongjiang Province*; China Agricultural Press: Beijing, China, 2011.
3. Yang, F.; Wang, Z.C.; Ma, H.Y.; Yang, F.; An, F. Research and integrated demonstration of ecological amelioration techniques of saline-sodic land in northeast China. *Acta Ecol. Sin.* **2016**, *36*, 7054–7058. [CrossRef]

4. Zhu, J.L.; Yan, Z.Q. Screening test of saline-alkali-tolerant rice varieties in Zhoushan saline-alkali fiele. *Zhejiang Agric. Sci.* **2021**, *62*, 1913–1915. [CrossRef]
5. Ma, Z.H.; Cao, Y.; Wang, Q.L. Effects of Planting Rice on Soil Physical and Chemical Properties of Saline-alkali Land in Northern Shaanxi and Screening of Saline-alkali-tolerant Rice Varieties. *China Rice* **2022**, *28*, 80–84. [CrossRef]
6. Wang, Q.J.; Li, M.X.; Zhao, H.L.; Wang, G.S. Evaluation and Screening of Germplasm Resources with Saline-Alkali Tolerance in Heilongjiang Province. *Crops* **2012**, *4*, 116–120. [CrossRef]
7. Ding, G.H.; Liu, K.; Cao, L.Z.; Bai, L.M.; Wang, T.; Zhou, J.S.; Luo, Y.; Xia, T.S.; Yang, G.; Wang, X.Y.; et al. Breeding of a Saline-alkali Tolerant Rice Variety Longdao 124 with High Quality and Stable Yield in Cold Regions. *China Seed Ind.* **2021**, *6*, 78–81. [CrossRef]
8. Liu, B.A. Screening test report of saline-alkali tolerant rice varieties in western Jilin Province. *Jilin Agric.* **2016**, *23*, 86. [CrossRef]
9. Huang, A.Y. *Comparative Analysis of ten Rice Varieties on Salt-Endurance in Qinghua, Vietnam*; Sichuan Agricultural University: Ya'an, China, 2016.
10. Wang, Z.S.; Zhu, Y.Q.; Li, N.; Liu, H.; Zheng, H.J.; Wang, W.P.; Liu, Y. High-throughput sequencing-based analysis of the composition and diversity of endophytic bacterial community in seeds of saline-alkali tolerant rice. *Microbiol. Res.* **2021**, *250*, 126794. [CrossRef]
11. Geetha, S.; Vasuki, A.; Jagadeesh, S.P.; Saraswathi, R.; Krishnamurthy, S.L.; Palanichamy, M.; Dhasarathan, G.; Thamodharan, M.B. Development of sodicity tolerant rice varieties through marker assisted backcross breeding. *Electron. J. Plant Breed.* **2017**, *8*, 1013–1031. [CrossRef]
12. Wang, W.L. *Using Indica-Japonica Cross RIL Population to Locate QTLs Related to Salit and Alkali Tolerance in Rice*; Shenyang Agricultural University: Shenyang, China, 2020. [CrossRef]
13. Wang, H. *Screening of Saline-Alkaline Tolerant Varieties of Rice (Oryza sativa L.) and Genetic Analysis*; Northeast Forestry University: Harbin, China, 2019. [CrossRef]
14. Sun, J.; Xie, D.W.; Zhang, E.Y.; Zheng, H.; Wang, J.; Liu, H.; Yang, L.; Zhang, S.; Wang, L.; Zou, D. QTL mapping of photosynthetic-related traits in rice under salt and alkali stresses. *Euphytica* **2019**, *215*, 147. [CrossRef]
15. Hibben, J.H.; Teller, E. The Raman effect and its chemical aplications and physical research. *Ind. Eng. Chem. News Ed.* **1939**, *17*, 556.
16. Chen, H. Research on data analysis and visualization platform based on Python. *Netw. Secur. Technol. Appl.* **2022**, *2*, 57–58.
17. Zhang, T.; Fan, S.X.; Xiang, Y.; Zhang, S.J.; Wang, J.H.; Sun, Q. Non-destructive analysis of germination percentage, germination energy and simple vigour index on wheat seeds during storage by Vis/NIR and SWIR hyperspectral imaging. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *239*, 118488. [CrossRef] [PubMed]
18. Liu, J.M.; Jin, S.; Bao, C.G.; Sun, Y.; Li, W.Z. Rapid determination of lignocellulose in corn stover based on near-infrared reflectance spectroscopy and chemometrics methods. *Bioresour. Technol.* **2021**, *321*, 124449. [CrossRef]
19. He, J.; Hui, J.Z.; Wang, S.D.; Hong, X.; Wang, K. Application of Python in Visualization of CINRAD Storm Products. *Meteorol. Sci. Technol.* **2020**, *48*, 374–379. Available online: https://scjg.cnki.net/kcms/detail/detail.aspx?filename=QXKJ202003011&dbcode=CJFQ&dbname=CJFD2020&v= (accessed on 10 January 2022).
20. Gao, W.; Sun, P.P.; Li, D.Z. Visual Analysis of Film Data Based on Python Crawler. *J. Shenyang Univ. Chem. Technol.* **2020**, *34*, 73–78.
21. Pu, Y.P. Research on Data Visualization Based on Python in the Era of Big Data. *China Comput. Commun.* **2021**, *33*, 179–182.
22. Le, T.D.; Gathignol, F.; Vu, H.T.; Nguyen, K.L.; Tran, L.H.; Vu, H.T.; Dinh, T.X.; Lazennec, F.; Pham, X.H.; Véry, A.; et al. Genome-Wide Association Mapping of Salinity Tolerance at the Seedling Stage in a Panel of Vietnamese Landraces Reveals New Valuable QTLs for Salinity Stress Tolerance Breeding in Rice. *Plants* **2021**, *10*, 1088. [CrossRef]
23. Wu, N.; Tang, Y.F.; Lin, Y.J.; Zeng, Y.; Ma, J.; Wang, N. Expression of Some Genes Related to Resistance to Salt-alkali Stress in'Hitomebore'. *Mol. Plant Breed.* **2019**, *17*, 7634–7640. [CrossRef]
24. Zhu, M.X.; Gao, X.Y.; Shao, X.W.; Jin, F.; Geng, Y.Q.; Wang, S. Effect of Different Concentrations of Saline-Alkali Stress on Growth and Yield of Rice. *Jilin Agric. Sci.* **2014**, *39*, 12–16. [CrossRef]
25. Cao, Y.F.; Yuan, P.S.; Wang, H.Y.; Korohou, T.W.; Fan, J.Q.; Xu, H.L. Monitoring Index of Rice Bacterial Blight Based on Hyperspectral Fractal Dimension. *J. Agric. Mach.* **2021**, *52*, 134–140.
26. Wang, Y.N.; Fan, S.J. MSAP Analysis of Genomic DNA Methylation in Oryza sativa under Low Temperature Stress. *Anhui Agric. Sci.* **2017**, *45*, 135–137+186. [CrossRef]
27. Tian, F.M. *Identification of Rice Based on Analysis of Raman Spectrum and Organic Ingredients*; Jilin University: Jilin, China, 2018; Available online: https://www.globethesis.com/?t=1361330542982755 (accessed on 10 February 2022).
28. Almeida, M.R.; Alves, R.S.; Nascimbem, L.B.; Stephani, R.; Poppi, R.J.; Oliveira, L.F. Determination of amylose cetent in starch using Raman spectroscopy and multivariate calibration analysis. *Aanlytical Bioanal. Chem.* **2010**, *397*, 2693–2701. [CrossRef]
29. Luo, Q. *The Development of the Low Background Gamma Ray Spectrum Analysis Software*; Chengdu University of Technology: Chengdu, China, 2019. [CrossRef]
30. Noor, S.A.; Kasim, K.A.; Sameer, A. Kadhim BER Performance Improvement of Alamouti MIMO-STBC Decoder Using Mutual Information Method. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020; p. 012016.
31. Sharma, N.V.; Yadav, N.S. An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers. *Microprocess. Microsyst.* **2021**, *85*, 104293. [CrossRef]

32. Wang, C.S.; Shu, Q.Q.; Wang, X.Y.; Guo, B.; Liu, P.; Li, Q. A random forest classifier based on pixel comparison features for urban LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2019**, *148*, 75–86. [CrossRef]
33. Narasimhulu, C.V. An automatic feature selection and classification framework for analyzing ultrasound kidney images using dragonfly algorithm and random forest classifier. *IET Image Precess.* **2021**, *15*, 2080–2096. [CrossRef]
34. Mantas, L.; Arnas, U. Efficient Implementations of Echo State Network Cross-Validation. *Cogn. Comput.* **2021**, prepublish. [CrossRef]
35. Saha, A.; Pal, S.C.; Chowdhuri, I.; Islam, A.R.M. Towfiqul, Roy Paramita, Chakrabortty Rabin. Land degradation risk dynamics assessment in red and lateritic zones of eastern plateau, India: A combine approach of K-fold CV, data mining and field validation. *Ecol. Inform.* **2022**, *69*, 101653. [CrossRef]
36. Data Partitioning—Hold-Out, K-Fold CV, Bootstrap. Available online: https://blog.csdn.net/weixin_37352167/article/details/85028835 (accessed on 21 June 2022).
37. Sainani, K.L. Multinomial and Ordinal Logistic Regression. *PM&R J. Inj. Funct. Rehabil.* **2021**, *13*, 1050–1055. [CrossRef]
38. Nattino, G.; Pennell, M.L.; Lemeshow, S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics* **2020**, *76*, 549–560. [CrossRef]
39. Vladimir, N.; Michiel, K. On Stochastic Optimization and Statistical Learning in Reproducing Kernel Hilbert Spaces by Support Vector Machines(SVM). *Informatica* **2009**, *20*, 273–292.
40. Understanding the Confusion Matrix. Available online: https://blog.huati365.com/f8111c156fc686cd (accessed on 15 June 2022).
41. Wang, L.Q.; Zhang, C.; Hou, Y.C.; Tan, X.H.; Cheng, R.; Gao, X.; Bai, Y.P. Remote sensing image scene classification application based on deep learning feature fusion. *J. Nanjing Univ. Inf. Sci. Technol.* **2021**, *2021*, 6659831.
42. Sha, M.; Tang, Z.L.; Zhang, D.; Zhang, Z.Y.; Liu, J. Study on cyclic voltammetric electrochemical fingerprint method for origin traceability of rice. *J. Phys. Conf. Ser.* **2021**, *1952*, 022038. [CrossRef]
43. Violino, S.; Ortenzi, L.; Antonucci, F.; Pallottino, F.; Benincasa, C.; Figorilli, S.; Costa, C. An Artificial Intelligence Approach for Italian EVOO Origin Traceability through an Open Source IoT Spectrometer. *Foods* **2020**, *9*, 834. [CrossRef]
44. Qian, L.; Zuo, F.; Zhang, C.D.; Zhang, D.J. Geographical Origin Traceability of Rice: A Study on the Effect of Processing Precision on Index Elements. *Food Sci. Technol. Res.* **2019**, *25*, 619–624. [CrossRef]
45. Chen, X.Y.; Jin, F.; Feng, D.H.; Wang, Y.L.; Liang, Y.J. Classification of sunspot magnetic types based on two-model integration. *Astron. Res. Technol.* **2022**, *7*, 1–11. [CrossRef]
46. Kong, X.R. Overview of Machine Learning. *Electron. Manuf.* **2019**, *24*, 82–84+38. [CrossRef]
47. Zhu, L.P. Review of sparse sufficient dimension reduction: Comment. *Stat. Theory Relat. Fields* **2020**, *4*, 134. [CrossRef]
48. Flavio, E.S.; Pilar, B.; Serge, G.; Javier, M.; Elizabeth, T. A spectral envelope approach towards effective SVM-RFE on infrared data. *Pattern Recognit. Lett.* **2016**, *71*, 59–65. [CrossRef]