plants

MDPI

Article

Intelligent Identification and Features Attribution of Saline–Alkali-Tolerant Rice Varieties Based on Raman Spectroscopy

Bo Ma [1,2], Chuanzeng Liu [1,2], Jifang Hu [1,2], Kai Liu [2,3], Fuyang Zhao [1,2], Junqiang Wang [1], Xin Zhao [4], Zhenhua Guo [3], Lijuan Song [3], Yongcai Lai [2,3,*] and Kefei Tan [1,2,*]

[1] Qiqihar Branch of Heilongjiang Academy of Agricultural Sciences, Qiqihar 161006, China; mabo8210@haas.cn (B.M.); cjf69@163.com (C.L.); hujifang7@haas.cn (J.H.); zfyhhz@haas.cn (F.Z.); august-wjq@haas.cn (J.W.)
[2] Northeast Branch of National Saline–Alkali-Tolerant Rice Technology Innovation Center, Harbin 150000, China; liukailouis@163.com
[3] Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China; hljsdsgzh@haas.cn (Z.G.); songlijuan@haas.cn (L.S.)
[4] College of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China; zxxsnh@hrbeu.edu.cn
* Correspondence: yame0451@163.com (Y.L.); tkfhlj@haas.cn (K.T.)

check for updates

Citation: Ma, B.; Liu, C.; Hu, J.; Liu, K.; Zhao, F.; Wang, J.; Zhao, X.; Guo, Z.; Song, L.; Lai, Y.; et al. Intelligent Identification and Features Attribution of Saline–Alkali-Tolerant Rice Varieties Based on Raman Spectroscopy. *Plants* **2022**, *11*, 1210. https://doi.org/10.3390/plants11091210

Academic Editors: Laura-Jayne Gardiner and Ritesh Krishna

Received: 10 April 2022
Accepted: 27 April 2022
Published: 29 April 2022

Abstract: Planting rice in saline–alkali land can effectively improve saline–alkali soil and increase grain yield, but 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 Python machine deep learning method was used to analyze the Raman molecular spectroscopy of rice and assist in feature attribution, in order to study a fast and efficient identification method of saline–alkali-tolerant rice varieties. A total of 156 Raman spectra of four rice varieties (two saline–alkali-tolerant rice varieties and two saline–alkali-sensitive rice varieties) were analyzed, and the wave crests were extracted by an improved signal filtering difference method and the feature information of the wave crest was automatically extracted by scipy.signal.find_peaks. Select K Best (SKB), Recursive Feature Elimination (RFE) and Select F Model (SFM) were used to select useful molecular features. Based on these feature selection methods, a Logistic Regression Model (LRM) and Random Forests Model (RFM) were established for discriminant analysis. The experimental results showed that the RFM identification model based on the RFE method reached a higher recognition rate of 89.36%. According to the identification results of RFM and the identification of feature attribution materials, amylum was the most significant substance in the identification of saline–alkali-tolerant rice varieties. Therefore, an intelligent method for the identification of saline–alkali-tolerant rice varieties based on Raman molecular spectroscopy is proposed.

Keywords: saline–alkali-tolerant rice; Raman spectroscopy; Python; scipy.signal.filtfilt difference; identification feature information

1. Introduction

Saline–alkali fields are a widely distributed type of land presenting an agricultural barrier. They are often in uncultivated or semi-uncultivated land due to the poor suitability of saline–alkali soil for cultivation and the harm of saline–alkali to crops [1]. Rice is a moderately saline–alkali-tolerant plant and is considered the "pioneer" crop for soil improvement in saline–alkali land, which can accelerate the desalination and accumulation of organic matter by taking advantage of its unique advantages of saline–alkali resistance and growth in water [2,3]. Thus, planting rice in saline–alkali land can increase grain yield, improve saline–alkali soil and improve the ecological environment [4].

However, there are great differences in saline–alkali tolerance among different rice varieties [5], so rice breeding experts pay much attention to the identification and evaluation of saline–alkali-tolerant rice varieties. Wang et al. conducted a screening and identification of saline–alkali tolerance of japonica rice varieties in cold areas through an investigation and analysis of phenotypic morphological indicators [6–8]. However, this traditional identification method requires a large number of field tests, which are not only subject to the limitations of growth cycle and the natural environment, but also have the problems of a heavy workload and time consumption. Some researchers used genomics and proteomics methods to identify QTLs, analyzed the role of specific protein-linked genes, then identified and cultivated saline–alkali rice varieties by marker-assisted selection [9–11]. Genetic identification methods still require professionals to operate advanced equipment skillfully, and then analyze complex data, which is not only complicated in the detection process, but also high in cost.

Raman spectroscopy was first used in physical and chemical studies by Hibben J H and Teller E [12], which was mainly used to study the vibration and structure of molecular groups [13]. Since then, Raman spectroscopy technology has been widely applied in many other fields: food field [14], geological field [15], medical field [16] and agricultural field [17], etc. Giang Le Truong obtained sample information from 32 rice varieties in Vietnam through Raman spectroscopy, and conducted a multivariate analysis, such as PCA, KNN and HCA, to evaluate and identify rice varieties [18]. Japanese researchers established spectral information at the molecular level by Raman spectrometer on amylose, amylopectin, protein content and chemical residues of six rice varieties in Japan and proposed to use bar codes specially formulated for Raman spectroscopy to show the nutritional quality of labeled rice products [19,20]. In a word, a Raman spectrometer measurement has the advantages of fast, efficient and accurate, and Raman spectroscopy as a molecular spectrum can achieve rapid, accurate and nondestructive analysis of samples.

Python not only has the matrix manipulation functionality supported by commercial high-level languages such as Matlab, but it is also cleaner and more concise than languages such as Java and C [21]. Python has become the most popular programming language for big data analysts and has a powerful deep learning method that visualizes data from different neighborhoods to further analyze the visualizations [22–24]. Python can effectively deal with various interferences of spectral data based on its data visualization features and chooses the best algorithm for spectral data disturbance reduction according to the visualization effect. Python can extract spectral crests and features of the crest based on its data visualization features. Python can automatically reduce the dimensions of spectral features based on its powerful deep learning method function, which not only reduces the calculation time of the identification model, but also effectively improves the identification effect of the optimized model. The fingerprint identification of selected feature information is carried out according to the discriminant result of the optimized model.

At present, the application of Raman spectroscopy combined with Python in the identification of saline–alkali-tolerant rice varieties have not been reported. In view of this, this study used Raman spectrometer to acquire the molecular information of rice varieties, combined with Python deep learning and visual analysis method of data, was used to establish an identification model for saline–alkali-tolerant rice varieties [25]. The main purpose of this study is to obtain the identification model of saline–alkali-tolerant rice and provide an intelligent method for rice breeders to identify saline–alkali-tolerant rice varieties.

2. Materials and Methods

2.1. Test Material

The test materials were two saline–alkali-tolerant rice varieties and two saline–alkali-sensitive rice varieties [26,27], which were, respectively, taken from control soil and saline–alkali soil, and identified, screened and provided by Qiqihar Branch of Heilongjiang

Academy of Agricultural Sciences. Soil basic nutrients and saline–alkali status was shown in Table 1.

Table 1. Basic nutrient status of soil (AHN: alkali-hydro nitrogen; AP: available phosphorus; RAP: rapidly available potassium; OM: organic matter).

Soil Types	AHN (mg/kg)	AP (mg/kg)	RAP (mg/kg)	OM (g/kg)	pH	Salt Content (%)
Saline–alkali soil	146.7	38.6	138.4	30.5	9.2	0.46
Control soil	144.7	39.2	138.9	31.8	6.9	0.23

Saline–alkali-tolerant coefficients (STC), namely, the relative values of each saline–alkali tolerance index, were used to evaluate the saline–alkali tolerance of rice, so as to eliminate the differences in basic traits among the tested materials [28]. The saline–alkali tolerance indexes in this experiment mainly investigated the four main agronomic traits of rice, namely plant height, tiller number, grain number per spike and 1000-grain weight.

The results showed that the STC of two saline–alkali-tolerant varieties BD6 and DF132 were significantly higher than those of the saline–alkali-sensitive varieties KD42 and LD12 in plant height, tiller number, grain number per spike and 1000-grain weight (p < 0.05; Figure 1).

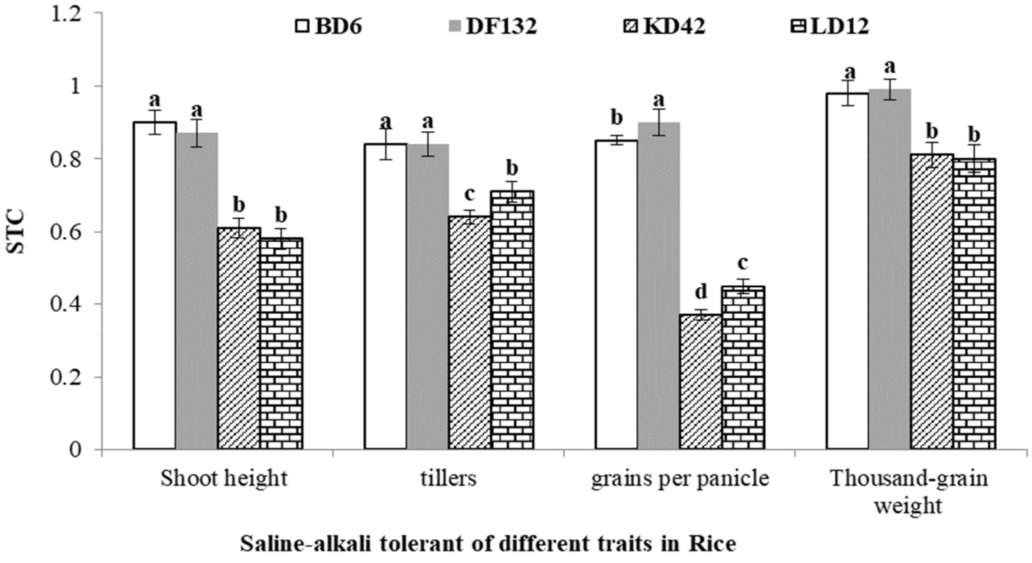

Figure 1. The saline–alkali tolerance coefficient of plant height, tiller number, grain number per ear and 1000-grain weight in rice. Values with different superscript letters are significantly different at p < 0.05. STC = index under saline–alkali stress/control index.

2.2. Sample Processing

In October 2021, four rice varieties (10 caves for each variety, a total of 40 caves) obtained from the control field were dried naturally for 20 days in a laboratory of 22–25 °C, and the water content was reduced to about 15%. A total of 10 ears were taken from different positions in each cave, and 10 grains were taken from different positions of each ear, a total of 1000 grains. These grains were shelled by Shanghai Chao Xing LJJM for 40 s. 39 rice grains (Table 2) with complete appearance after shelling for each variety were selected as samples, and a total of 156 samples were obtained.

Table 2. Variety and quantity of test samples (1: saline–alkali-tolerant; 0: saline–alkali-sensitive).

Numeral	Name of Sample	Variety of Sample	Number of Samples
1	BD6	1	39
2	DF132	1	39
3	KD42	0	39
4	LD12	0	39

2.3. Obtaining Spectral Information

The sample information was collected by Advantage 532 Desktop Raman spectrometer. The excitation wavelength was 532 nm, which was an ideal light source for resonance Raman research, the measurement range was 200–3400 cm^{-1}, the excitation power was less than 5 mw, the resolution was 1.4 cm^{-1}, and the spectral information of 156 samples was obtained with 4 scanning times. Pro Scope HR software was used to obtain sample image information and sample data, which were saved in PRN format. The laser power was high, integration time was 4, number of spectrums was 3, display was average, save spectrum was ASCII and resolution was low. The data processing software was Python.

3. Results and Analysis

3.1. Disturbance Reduction and Crest Extraction

3.1.1. Extraction of Original Raman Spectral Information

The raw data extracted from Python are shown in Figure 2. When using Raman spectrometer to collect rice spectral data, due to the interference of the instrument in terms of noise, stray light and fluorescence background, the data accuracy was affected. The spectral information of the four rice varieties was intertwined in a disorderly manner, making it difficult to distinguish. Therefore, it is essential to denoise and remove impurities from the original Raman spectral data.

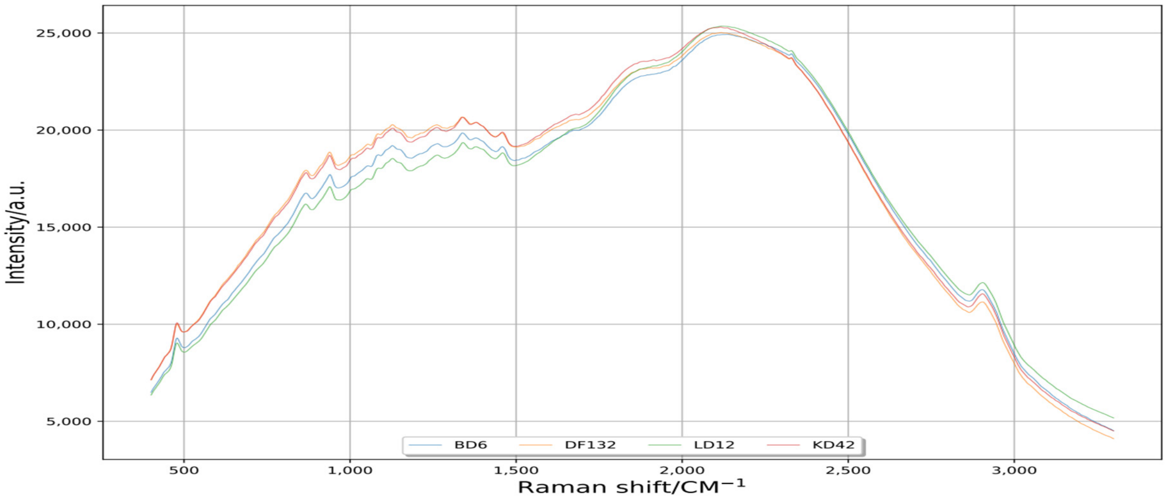

Figure 2. Original spectral curve. Raman shift is the reciprocal of wavelength, unrelated to the frequency of incident light and only related to the vibration frequency of sample molecules, and its range is 200–3400 cm^{-1}. Intensity is the intensity of Raman scattering, which is the anti-Stokes line; the anti-Stokes line is the scattering light of frequency shifted light from monochromatic incident light in a molecule.

3.1.2. Reduction in Disturbance of Scipy.Signal.Lfilter Method

The scipy.signal.lfilter works with many basic data types when dealing with data dis-turbance reduction by Python. This filter is an implementation of the transpose of the di-rect form of the standard difference equation [29]. In this experiment, scipy.signal.lfilter was used, and the results are shown in Figure 3. The curve information of disturbance reduction by the scipy.signal.lfilter method was obviously smooth compared to that of the original data. This scipy.signal.lfilter method obtained a good effect of data disturbance reduction.

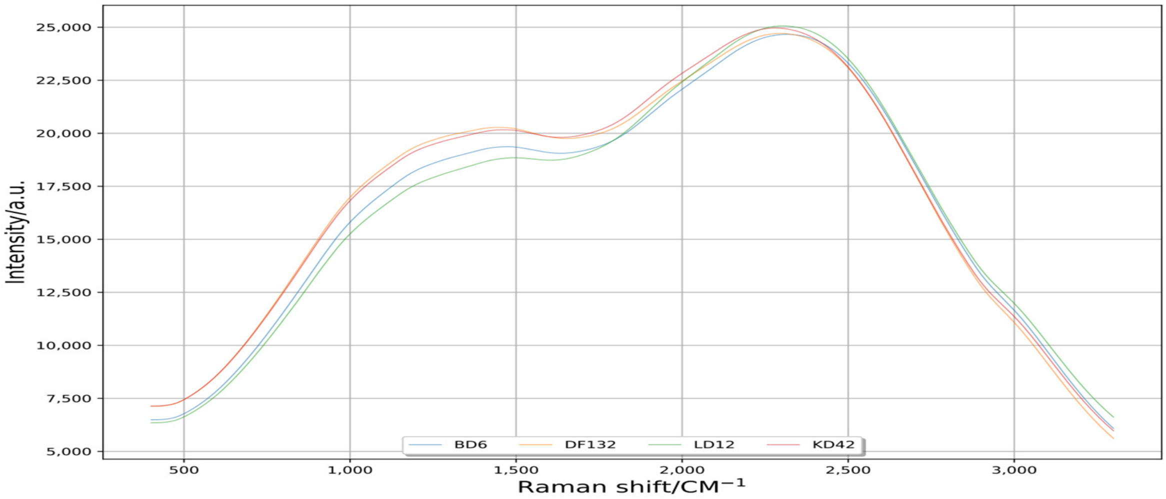

Figure 3. Spectral curves of disturbance reduction by filtering method. Raman shift is the reciprocal of wavelength, unrelated to the frequency of incident light and only related to the vibration frequency of sample molecules, and its range is 200–3400 cm^{-1}. Intensity is the intensity of Raman scattering, which is the anti-Stokes line; the anti-Stokes line is the scattering light of frequency shifted light from monochromatic incident light in a molecule.

3.1.3. Reduction in Disturbance of Scipy.Signal.Filtfilt Method

The scipy.signal.filtfilt method could quickly help to achieve a reduction in the interference of data. This method used a linear digital filter, which had twice the order of other filters, once forward and once back [30]. As shown in Figure 4, the curve processed by the scipy.signal.filtfilt method was smoother than that processed by scipy.signal.lfilter method. So, the scipy.signal.filtfilt method obtained better disturbance reduction effect on the original data of the Raman spectrum.

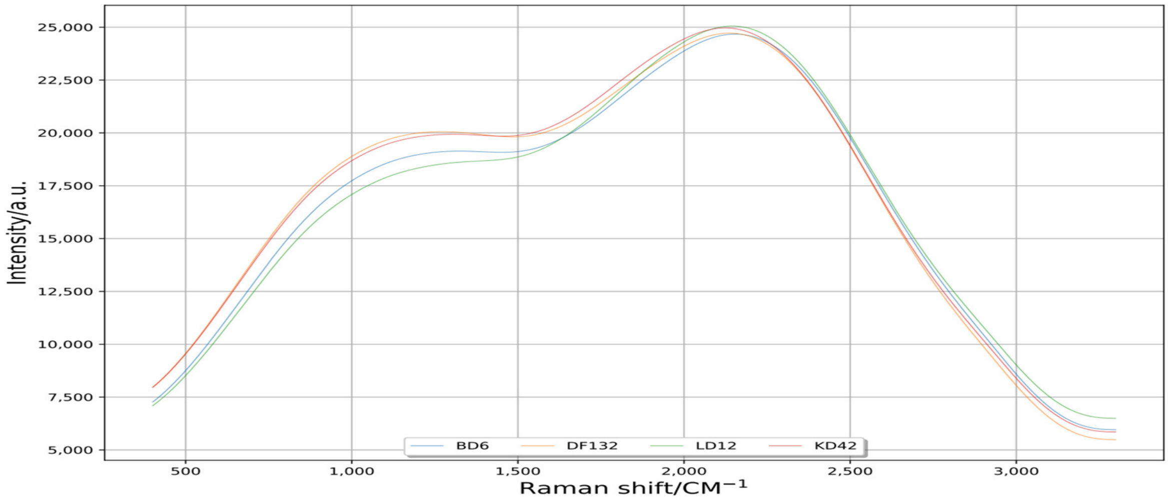

Figure 4. Spectral curves of disturbance reduction by signal filtering method. Raman shift is the reciprocal of wavelength, unrelated to the frequency of incident light, and only related to the vibration frequency of sample molecules, and its range is 200–3400 cm^{-1}. Intensity is the intensity of Raman scattering, which is the anti-Stokes line; the anti-Stokes line is the scattering light of frequency shifted light from monochromatic incident light in a molecule.

3.1.4. Extraction of Wave Crest by Difference Method of Scipy.Signal.Filtfilt Method

Although the curve processed by scipy.signal.filtfilt had a good effect of removing disturbance, it filtered the useful spectral crest information. Therefore, the extraction of spectral crest information is essential. The improved signal filtering method, namely the scipy.signal.filtfilt method difference method, was used to extract the spectral crest information, as shown in Formula (1):

Differential equation:

$$X = x - y \tag{1}$$

Note: The x is the axis of data to be filtered; the y is the filtered data axis.

The full band of Raman spectrum was clearly visible, as shown in Figure 5. The eighteen significant crest information appeared in each rice variety. Referring to wave crest extraction, Raman spectral characteristics and the attribution of rice [31], seven effective crests were extracted at 480, 865, 941, 1129, 1339, 1461 and 2910 cm^{-1}. Each rice variety had different wave intensities near the same wave value. Seven effective crests extracted by the signal filtering difference method laid the foundation for the extraction of crest feature in the next step.

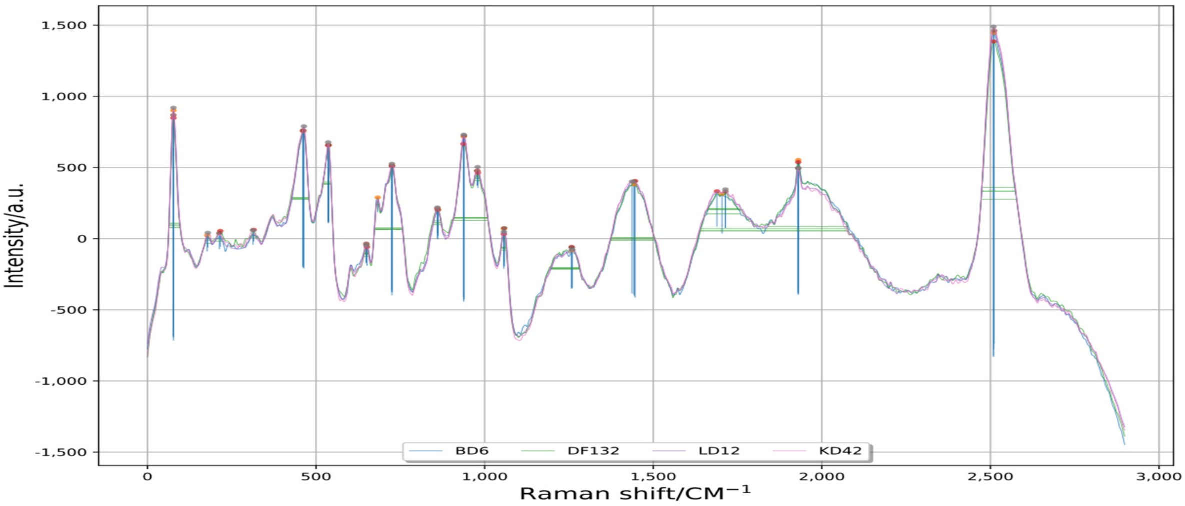

Figure 5. Crest and its four-dimensional feature information. Prominence is the length of the blue vertical line, width is the width of the green horizontal line, width_height is the length from the green line to the peak, and peak_dif is the peak offset. Raman shift is the reciprocal of wavelength, unrelated to the frequency of incident light and only related to the vibration frequency of sample molecules, and its range is 200–3400 cm^{-1}. Intensity is the intensity of Raman scattering, which is the anti-Stokes line; the anti-Stokes line is the scattering light of frequency shifted light from monochromatic incident light in a molecule.

3.2. Extraction of Wave Crest Feature

Four feature data for each wave crest were extracted by scipy.signal.find_peaks method in Python. As shown in Figure 5, four feature information (prominence, width, width_height and peak_dif) could accurately lock the shape and position of each wave crest.

The experiment consisted of 156 samples; each sample had seven crests and contained 4-dimensional feature information, so each sample owned 28-dimensional feature information. In addition, each sample owned its own three-dimensional feature information, i.e., name of sample, variety of sample and number of samples. Each sample owned 31-dimensional feature information based on the initial feature extraction. With the above machine learning approach, a 156 × 31-dimensional matrix feature information was established.

3.3. Reduction in Dimensionality of Features Information

If the feature information of 156 × 31-dimension matrix is directly brought into the classification model for machine deep learning, the large feature information matrix will lead to a large amount of computation and a long training time. Therefore, it is essential to reduce the dimensions of feature information.

3.3.1. SKB Method for Dimensionality Reduction

The working principle of SKB is to use a certain parameter to score features and select the best several feature information. It is known as a large variable-characteristics selection tool. In this experiment, the SKB method selected mutual_info_regression parameters, and the mutual information method scored the 28-dimensional feature information of dimensionality unreduced, as shown in Formula (2) [32]. The 10 most powerful feature information were finally selected by the mutual trust method, as shown in Table 3.

$$\mathrm{I}(X;Y) = \sum_{x\in X}\sum_{y\in Y} \mathrm{p}(x,y)\log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)} \tag{2}$$

Table 3. Results of three methods for dimensionality reduction (p: prominences; w: width; wh: width_height; pd: peak_dif).

Number	Raman Shift/cm^{-1}	Initial Feature Extraction	SKB	RFE	SFM
1	480	p\w\wh\pd	p\pd	w\pd	pd
2	865	p\w\wh\pd	p\pd	pd	pd
3	941	p\w\wh\pd	pd	p\wh\pd	pd
4	1129	p\w\wh\pd	pd	pd	pd
5	1339	p\w\wh\pd	pd	p\wh\pd	p\wh\pd
6	1461	p\w\wh\pd	wh	p\w\pd	p\w\pd
7	2910	p\w\wh\pd	w\p	wh	pd
Total		28	10	14	11

Note: The mutual information method is used to evaluate the correlation between category independent variables and category dependent variables.

The SKB method selected seven crests and 10-dimensional feature information. Compared with the initial method, the select rate of effective crest was 100%, and the selection rate of effective feature information was 35.71%. A 156 × 13-dimensional matrix feature information was established.

3.3.2. RFE Method for Dimensionality Reduction

The main idea of RFE is to iteratively build the feature of the model, eliminate the redundancy between features, select the optimal feature combination, and reduce the feature dimension. This experimental method used the Random Forest Classifier (RFC) module [33]. Firstly, the initial subset of 28-dimensional features was input into the RFC module, the importance of each feature was calculated, and the classification accuracy of the initial feature subset 1 was obtained by a cross-validation method. Second, the feature with the lowest significance was removed from feature subset 1, and feature subset 2 was obtained, which was input into the RFC module again to obtain the classification accuracy of subset 2. The above process was repeated until the feature subset is empty. Finally, several feature subsets with different number of features were obtained, and the feature subset with the highest classification accuracy was selected as the optimal feature combination. Therefore, this was a greedy algorithm to find the optimal feature subset, and finally selected 14 optimal feature data, as shown in Table 3.

The RFE method selected seven crests and 14-dimensional feature information. Compared with the initial method, the select rate of effective crest was 100%, and the selection rate of effective feature information was 50%. A 156 × 17-dimensional matrix feature information was established.

3.3.3. SFM Method for Dimensionality Reduction

SFM is a built-in general transformer model in the feature selection method, which can perform feature selection through the indicators given by the model itself. In this experiment, the free feature selection method in SFM was selected, and the importance degree of different features was obtained after training with RFC model [34]. Features were selected according to the weight of importance, and 11 optimal feature data were finally selected, as shown in Table 3.

The SFM method selected seven peaks and 11-dimensional feature information. Compared with the initial method, the select rate of effective crest was 100%, and the selection rate of effective feature information was 39.29%. A 156 × 14-dimensional matrix feature information was established.

Three feature information selection methods were used to reduce the dimension of the initial feature information, which reduced the feature matrix and computing time. Whether the dimensionality reduction method of feature information could accurately identify saline–alkali-tolerant rice varieties required the classification model to evaluate the effectiveness of the feature information selection.

3.4. Establishment of Recognition Model

Based on one feature extraction and three feature selection methods, the feature information of a 156 × 31-dimensional matrix was established by the initial method, the feature information of a 156 × 13-dimensional matrix was established by the SKB method, the feature information of a 156 × 17-dimensional matrix was established by the RFE method and the feature information of a 156 × 14-dimensional matrix was established by the SFM method. The sample data of each rice variety were divided according to 7:3. A total 109 samples were divided into training sets, and 47 samples were divided into test sets. In order to find a fast, convenient, economical, reliable and accurate recognition model for saline–alkali-tolerant rice varieties, four matrix feature data of different dimensions were brought into the classification model, respectively, to conduct machine learning and evaluate the selection method of feature information.

3.4.1. Establishment of Logistic Regression Model (LRM)

LRM is a classification model in machine learning [35], whose input function is the result of a linear regression, as shown in Formula (3):

$$\mathrm{h(w)} = \mathrm{w_1x_1} + \mathrm{w_2x_2} + \mathrm{w_3x_3} \ldots + \mathrm{b} \tag{3}$$

Input the LRM output results into the sigmoid function, as shown in Formula (4):

$$\text{Sigmoid function}: \ \mathrm{g}\left(\theta^{\mathrm{T}}\mathrm{x}\right) = \frac{1}{1 + \mathrm{e}^{-\theta^{\mathrm{T}}\mathrm{x}}} \tag{4}$$

The sigmoid function output results were in the interval of [0,1], in which the default machine threshold was 0.5, and the classification results of the unoptimized LRM were obtained, as shown in Table 4. The matrix feature information of four different dimensions was brought into the unoptimized LRM for the recognition of saline–alkali-tolerant rice varieties, and the recognition rates were 80.85%, 74.47%, 76.60% and 80.85%, respectively.

Table 4. Modeling results of four feature selection methods for saline–alkali-tolerant rice varieties.

Feature Selection	Matrix Dimension	Accuracy		Accuracy Improvement
		LRM	RFM	
Initial	156 × 31	80.85%	80.85%	0
SKB	156 × 13	74.47%	82.98%	8.51
RFE	156 × 17	76.60%	89.36%	12.76
SFM	156 × 14	80.85%	85.11%	4.26

3.4.2. Establishment of Random Forests Model (RFM)

RFM establishes a forest in a random way, and there are many decision trees in the forest. There is no correlation between each decision tree in RFM. For the classification algorithm, when a new input sample enters, each decision tree in RFM is asked to make a judgment, respectively, to see which category this sample should belong to, and then see which category is selected the most, and predict this sample to be that category. The criterion parameter in this test was the gini coefficient, namely the decision tree of the CART category, as shown in Formulas (5) and (6).

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{5}$$

Note: Two samples were randomly drawn from dataset D with the probability that their class labels were inconsistent. Therefore, the smaller the *Gini* (D) value, the higher the purity of the dataset D.

$$Gini\_index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{6}$$

Note: The attribute that minimizes *Gini_index* (*D*,a) after division was selected as the optimal score attribute.

The CART could not only classify and regression, but also handled discrete and continuous attributes, and the classification results of the optimized RFM were obtained, as shown in Table 4. The matrix characteristic information of four different dimensions was brought into the optimized RFM for the recognition of saline–alkali-tolerant rice varieties, and the recognition rates were 80.85%, 82.98%, 89.36% and 85.11%, respectively. The results showed that the overall recognition rate of RFM was higher than those of LRM.

3.5. Attribution of Spectral Features Information

The main components of rice include starch, protein, fat, etc. [31]. Due to different content and structure, different spectral vibration information can be obtained (Table 5). In this experiment, the Raman spectrum of rice reached obvious Raman feature crests at 480, 865, 941, 1129, 1339, 1461 and 2910 cm^{-1}. The spectral attribute of feature crest at 480 cm^{-1} was amylum, and the pattern of manifestation was skeleton vibration. The spectral attribute of the feature crest at 865 cm^{-1} was amylopectin and sugar ring, and the pattern of manifestation was the vibration of C-H deformation and C-O ring. The spectral attribute of the feature crest at 941 cm^{-1} was amylopectin, and the pattern of manifestation was symmetric stretching vibration of C-O-C. The spectral attribute of feature crest at 1129 cm^{-1} was sugar, and the pattern of manifestation was the vibration of C-O stretching and C-O-H bending deformation. The spectral attribute of feature crest at 1339 cm^{-1} was sugar, and the pattern of manifestation was C-O-H bending and the vibration of C-C stretching. The spectral attribute of feature crest at 1461 cm^{-1} was sugar, and the pattern of manifestation was C-H bending vibration in-plane. The spectral attribute of feature crest at 2910 cm^{-1} was amylum, and the pattern of manifestation was stretching vibration of CH_2 and NH_2.

Table 5. Features and attribution of rice Raman spectra (s: strong; p: prominences; w: width; wh: width_height; pd: peak_dif).

Number	Raman Shift/cm^{-1}	Pattern of Manifestation	Spectral Attribution	Methods	Feature Information			
1	480 s	Skeleton vibration	amylum	SKB	p			pd
				RFE		w		pd
				SFM				pd
2	865 s	The vibration of C-H deformation and C-O ring	amylopectn sugar ring	SKB	p			pd
				RFE				pd
				SFM				pd
3	941 s	Symmetric stretching vibration of C-O-C	amylopectn	SKB				pd
				RFE	p		wh	pd
				SFM				pd
4	1129 s	The vibration of C-O stretching and C-O-H bending deformation	sugar	SKB				pd
				RFE				pd
				SFM				pd
5	1339 s	C-O-H bending and the vibration of C-C stretching	sugar	SKB				pd
				RFE	p		wh	pd
				SFM	p		wh	pd
6	1461 s	C-H bending vibration in-plane	sugar	SKB			wh	
				RFE	p	w		pd
				SFM	p	w		pd
7	2910 s	Stretching vibration of CH_2 and NH_2	amylum	SKB	p	w		
				RFE			wh	
				SFM				pd
		Total			8	4	5	18

4. Discussion

4.1. Interpretation of the Result of Reduction in Disturbance of Original Raman Spectral Information

With the improvements in the acquisition precision of Raman spectroscopy instruments, the collected Raman spectroscopy data contain terms of noise, stray light and fluorescence background [36]. These disturbances seriously affect the prediction accuracy of the model [37]. The effect of various interferences on model accuracy can be effectively eliminated by filtering disturbance reduction technology [38–40]. Python can visualize data from different neighborhoods to further analyze the visualization results [22–24]. In the Python language environment, scipy.signal.lfilter is a one-dimensional digital filter, and scipy.signal.filtfilt is called a zero-phase digital filter. In the basic implementation process of scipy.signal.filtfilt method, the signal is firstly filtered by scipy.signal.lfilter method, and then the signal is time-domain reversed and filtered by scipy.signal.lfilter method again. In this way, the phase is zero after two filters [41,42]. In this study, the curve processed by scipy.signal.filtfilt was smoother than that processed by scipy.signal.lfilter (Figures 2–4); the waveform processed by scipy.signal.lfilter had offset, but the one processed by scipy.signal.filtfilt did not, so the scipy.signal.filtfilt could be used for reducing the disturbance of original Raman spectral information well [43–45].

4.2. Interpretation of the Result of Dimensionality Reduction

In recent years, with the improvement of data collection and storage capacity, data dimension increases exponentially along with samples and frequently appears in many scientific neighborhoods [46]. For problems of large scale or dimensionality, accuracy of estimation and computational cost are two top concerns [47]. SKB, RFE and SFM can achieve effective reduction in the dimensions of data through the powerful deep learning methods of Python [21,48–50]. In this test, the SKB method used the mutual_info_regression parameter for feature dimensionality reduction, and the RFE and SFM methods used the RFC module for feature dimensionality reduction. Compared with the 28-dimensional features extracted by the original method (Table 3), these three dimensionality reduction methods finally obtained 10-dimensional features, 14-dimensional features and 11-dimensional features, respectively. Therefore, three methods of feature dimension reduction were superior in reducing data dimensions and increasing efficiency [51].

4.3. Interpretation of the Establishment of Recognition Model

For one feature extraction and three feature selection methods (Table 4), modeling performance was ranked as RFE-RFM > SFM-RFM > SKB-RFM > Initial-RFM = Initial-LRM > SFM-LRM > RFE-LRM > SKB-LRM. The reasons for this are as follows: firstly, in this study, two models were used, one of which was a probabilistic model (LRM), and the other an important machine learning model (RFM) [52]. RFM are able to handle multi-dimensional and multi-variety data [53], and the results of this study showed that RFM owned greater ability to identify saline–alkali-tolerant rice varieties than LRM. Secondly, when the 28-dimensional features extracted in this experiment were put into LRM and RFM, respectively, RFM did not improve the recognition rate compared to LRM. After dimensionality reduction, the feature dimension for SKB, RFE and SFM were 10, 14 and 11, respectively. When they were brought into LRM and SFM, compared with LRM, the recognition rate of RFM increased by 8.51, 12.76 and 4.26, respectively. Feature extraction and feature selection (feature dimension reduction) are not completely separated in practical application; there are few studies on the combination of the two and further research is needed [54]. Finally, the purpose of this research was to build a saline–alkali rice variety recognition model using stable and efficient machine learning algorithms, the literature has confirmed the superior performance of RFM in classification evaluation [53,55]; the RFE method can eliminate the features of a low contribution rate without increasing model errors through repeated iterations to select the best features [33,56]. Therefore, RFE dimension reduction method was combined with the RFM in this study to improve the recognition rate of the identification model. The results showed that the RFE-RFM model of saline–alkali-tolerant rice varieties had the best recognition rate (Table 4).

4.4. Interpretation of Attribution of Spectral Features Information

This study found (Table 5) that the attributions of the Raman spectral crest in this experiment were amylum, amylopectin, sugar and sugar ring, and the pattern of manifestation were skeleton vibration, the vibration of C-H deformation, C-O ring, symmetric stretching vibration of C-O-C, the vibration of C-O stretching, C-O-H bending deformation, C-O-H bending and the vibration of C-C stretching, the stretching vibration of CH2, and NH2 and C-H in-plane bending vibrations. Amylum contained amylopectin, which was the aggregation of sugar molecules and the most common storage form of carbohydrates in cells [57]. In saline–alkali-tolerant and saline–alkali-sensitive rice varieties, the carbohydrate transportation to grain was closely related to the filling stage, and the dynamic accumulation of amylum played a decisive role [58]. Under the same planting environment, the dynamic accumulation capacity of amylum of saline–alkali-tolerant rice varieties ware higher than that of saline–alkali-sensitive rice varieties at the filling stage [59]. Therefore, the attributions of Raman spectral crest and its pattern of manifestation provided a biological theoretical basis for the identification of saline–alkali-tolerant rice varieties.

5. Conclusions

The present study exhibits the feasibility of Raman spectroscopy combined with Python machine deep learning method for the rapid and accurate identification of saline–alkali-tolerant rice. To improve detection accuracy and efficiency, the spectral data were preprocessed by the combination of scipy.signal.filtfilt and scipy.signal.find_peaks. Using the feature extraction combined with feature dimension reduction and classification modeling, the RFE-RFM models outperformed the other seven combined models. Therefore, it can quickly and accurately identify saline–alkali-tolerant rice varieties and provide variety guarantee for rice planting in saline–alkali fields.

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

Funding: This research was funded by the Scientific Research Project of Heilongjiang Provincial Scientific Research Institutes of China (CZKYF2021-2-A002) and Scientific and Technological Project of Qiqihar City of China (CNYGG-2021031).

Data Availability Statement: The data and code presented in this study are openly available at github.com.

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

References

1. Xu, L.; Wang, Z.C.; Zhao, C.W.; Wang, M.; Ma, H. A Review of Saline-sodic Soil and Tillage Amelioration in Northeast of China. *Chin. Agric. Sci. Bull.* **2011**, *27*, 23–31.
2. Yu, B.L. Remediation Measures of Saline-alkali Land: A Review. *Chin. Agric. Sci. Bull.* **2021**, *37*, 81–87.
3. Zhang, Q.; Chen, F.D.; Feng, G.Y.; Hong, Q.I.; Wang, S.; Liang, Q.; Lei, X.; Wang, Y.; Lin, Y. Literature Review of Saline Soil Improvement and Utilization. *Eng. Technol. Res.* **2016**, *22*, 35–39. [CrossRef]
4. Yuan, B.F.; Ma, Y.T.; Bao, Y.; Zhang, J.; Sun, Q.; Wang, L. Effect of Rice Cultivation on Ameliorating Soil Fertility of Soda Saline-Alkali Soil in Western Jilin Province. *J. Soil Water Conserv.* **2019**, *33*, 320–326. [CrossRef]
5. Teng, F.K.; Zhu, W.; Tang, Z.B.; Niu, T.X.; Han, M.; Lan, P. Comparison of Salt and Alkaline Tolerance of Several Rice Varieties in Seeding Stage. *Agric. Technol.* **2021**, *41*. [CrossRef]
6. Wang, Z.X. *Saline-Alkaline Tolerance Evaluation and QTL Mapping of Japonica in The Northeast of China*; Northeast Agricultural University: Harbin, China, 2012.
7. Li, H.Y.; Si, Y.; Li, Y.; Du, C.; Zhou, X.; Liu, M.; Ning, H.; Ye, P. Principal Component Analysis and Comprehensive Evaluation of Saline-Alkaline Tolerance Related Traits of Northern Japonica Rice. *J. Nucl. Agric. Sci.* **2020**, *34*, 1862–1971. [CrossRef]
8. Ma, B.; Liu, C.Z.; Hu, J.F.; Chai, L.; Wang, L. Screening and Selecting of Japonica Germplasm with Saline-Alkali Tolerance in Cold Region. *Heilongjiang Agric. Sci.* **2011**, 6–8.
9. Liu, J.Y.; Shao, X.Y.; Zhou, D.; Shan, Z.; Mi, T.Z.; Yin, H.D.; Li, J.M. Research Progress on Identification Methods and Evaluation Indexes for Salt-alkali Tolerance in Rice. *Hybrid Rice* **2019**, *34*, 1–6. [CrossRef]
10. Inja, N.B.L.; KIM, B.K.; Yoon, I.S.; Kim, K.H.; Kwon, T.R. Salt Tolerance in Rice: Focus on Mechanisms and Approaches. *Rice Sci.* **2017**, *24*, 123–144.
11. Pires, I.S.; Negrão, S.; Oliveira, M.M.; Purugganan, M.D. Comprehensive phenotypic analysis of rice (*Oryza sativa*) response to salinity stress. *Physiol. Plant* **2015**, *155*, 43–54. [CrossRef]
12. Hibben, J.H.; Teller, E. The Raman effect and its chemical aplications and physical research. Industrial and Engineering Chemistry. *News Ed.* **1939**, *17*, 556.
13. Herzberg, G. Molecular spectra and molecular structure. In *Infrared and Raman Spectra of Polyatomid Molecules*; Van Nostrand, R., Ed.; American Journal of Physics: New York, NY, USA, 1945; Volume 2.
14. Shipp, D.W.; Sinjab, F.; Nothingher, I. Raman spectroscopy: Techniques and applications in the life sciences. *Adv. Opt. Photonics* **2017**, *9*, 315–428. [CrossRef]
15. Alian, W.; Randy, L.K.; Bradley, L.; Jolliff, Z.L. Raman imaging of extraterrestrial materials. *Planet. Space Sci.* **2015**, *112*, 23–34. [CrossRef]
16. Vankeirsbilck, T.; Vercauteren, A.; Baeyens, W.; Van, D.W.D. Applications of Raman spectroscopy in pharmaceutical analysis. *Trends Aanlytical. Chem.* **2002**, *21*, 869–877. [CrossRef]
17. Weng, S.Z.; Zhu, W.X.; Zhang, X.Y.; Yuan, H.C.; Zheng, L.; Zhao, J.L.; Huang, L.S.; Han, P. Recent advances in Raman technology with applications in agriculture, food and biosystems: A review. *Artif. Intell. Agric.* **2019**, *3*, 1–10. [CrossRef]
18. Giang, L.T.; Trung, P.Q.; Yen, D.H. Identification of rice varieties speciaties in Vietnam using Raman spectroscopy. *Vietnam J. Chem.* **2020**, *58*, 711–718. [CrossRef]
19. Pezzotti, G.; Zhu, W.L.; Chikaguchi, H.; Marin, E.; Masumura, T.; Sato, Y.; Nakazaki, T. Raman spectroscopic analysis of polysaccharides in popular Japanese rice cultivars. *Food Chem.* **2021**, *354*, 129434. [CrossRef]
20. Pezzotti, G.; Zhu, W.; Chikaguchi, H.; Marin, E.; Boschetto, F.; Masumura, T.; Sato, Y.; Nakazaki, T. Raman Molecular Fingerprints of Rice Nutritional Quality and the Concept of Raman Barcode. *Front. Nutr.* **2021**, *8*, 663569. [CrossRef]
21. Kong, H.P.; Liu, Z.Q. Implementation of Dimension Reduction Method of On-board Data of High-speed EMU Based on Python. *Comput. Eng. Softw.* **2020**, *41*, 114–117. [CrossRef]

22. Perkel, J.M. Python power-up: New image tool visualizes complex data. *Nature* **2021**, *600*, 347–348. [CrossRef]
23. Cao, S.; Zeng, Y.; Yang, S.; Cao, S. Research on Python Data Visualization Technology. *J. Phys. Conf. Ser.* **2021**, *1757*, 012122. [CrossRef]
24. Martin, L.; Bramkamp, Y.; Köster, T.; Staiger, D. SEQing: Web-based visualization of iCLIP and RNA-seq data in an interactive python framework. *BMC Bioinform.* **2020**, *21*, 113. [CrossRef]
25. Wagle, S.A.; Harkrishnan, R.; Ali, S.H.M. Faseehuddin Mohammad. Classification of Plant Leaves Using New Compact Convolutional Neural Network Models. *Plants* **2021**, *11*, 24. [CrossRef]
26. Song, F.R.; Song, L.Q.; Cao, Z.K.; Song, Z.; Nie Ho Shi, Y. Organosilicon Soil Conditioner: Effect on Soda-type Saline-alkali Soil Improvement and Rice Yield. *J. Agric.* **2021**, *11*, 58–63.
27. Gu, X.; Ren, C.M.; Wang., L.N.; Yang, L.; Zhang, H.Y.; Liu, B.; Sun, X.R.; Xu, H.C.; Zhao, J.R. Effects of humic acid application on soda Saline-alkali soil in Daqing. *China Soil Fertil.* **2021**, 77–82. [CrossRef]
28. Yi, Y.; Peng, Y.; Song, T.; Lu, S.; Teng, Z.; Zheng, Q.; Zhao, F.; Meng, S.; Liu, B.; Peng, Y.; et al. NLP2-NR Module Associated NO Is Involved in Regulating Seed Germination in Rice under Salt Stress. *Plants* **2022**, *11*, 795. [CrossRef] [PubMed]
29. Voskoboinikov, Y.E. Optimal Parameter Estimation of Spatial-Local Signal Filtering Algorithms. *Optoelectron. Instrum. Data Processing* **2019**, *55*, 222–229. [CrossRef]
30. Zhang, F.; Wang, D. Vibration Signal Filtering Algorithm Based on Singular Value Subspace Decomposition. *Rev. Téc. Ing. Univ. Zulia* **2017**, *39*, 106–113. [CrossRef]
31. Tian, F.M. *Identification of Rice Based on Analysis of Raman spectrum and Organic Ingredients*; Jilin University: Jilin, China, 2018.
32. Liu, X.H.; Huo, H. Automatic summarization for text based on mutual information. *J. Hefei Univ. Technol.* **2014**, *37*, 1198–1203. [CrossRef]
33. Pradier, L.; Tissot, T.; Fiston-Lavier, A.-S.; Bedhomme, S. PlasForest: A homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinform.* **2021**, *22*, 1–17. [CrossRef]
34. Mirmohammadi, P.; Ameri, A.; Shalbaf, A. Recognition of acute lymphoblastic leukemia and lymphocytes cell subtypes in microscopic images using random forest classifier. *Phys. Eng. Sci. Med.* **2021**, *44*, 433–441. [CrossRef] [PubMed]
35. Schober, P.; Vetter, T.R. Logistic Regression in Medical Research. *Anesth. Analg.* **2021**, *132*, 365–366. [CrossRef] [PubMed]
36. Yun, Y.-H.; Bin, J.; Liu, D.-L.; Xu, L.; Yan, T.-L.; Cao, D.-S.; Xu, Q.-S. A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration. *Anal. Chim. Acta* **2019**, *1058*, 58–69. [CrossRef] [PubMed]
37. Liu, J.; Chu, X.; Wang, Z.; Xu, Y.; Li, W.; Sun, Y. Optimization of Characteristic Wavelength Variables of Near Infrared Spectroscopy for Detecting Contents of Cellulose and Hemicellulose in Corn Stover. *Spectrosc. Spect. Anal.* **2019**, *39*, 743–750.
38. Huang, X.S.; Xu, G.B. Design of Audio Denoising IIR Filter Based on MATLAB. *Mod. Comput.* **2016**, *22*, 48–52. [CrossRef]
39. Wang, T. Application of digital filter in real speech noise reduction. *Inf. Commun.* **2018**, *3*, 30–31.
40. Wang, C.H.; Luo, X.; Ren, G.P.; Lu, M. On the Management of State-owned Group Enterprises. *Mod. Ind. Econ. Inf.* **2019**, *9*, 80–81. [CrossRef]
41. Available online: https://blog.csdn.net/Galaxy_Robot/article/details/106976165 (accessed on 4 April 2022).
42. Available online: https://blog.csdn.net/weixin_45366564/article/details/115122651 (accessed on 4 April 2022).
43. Yuan, F. Research on the influence of filter length on filtering results. *Internet Things Technol.* **2015**, *5*, 44–46. [CrossRef]
44. Shang, X.H.; Guo, A.H.; Li, G.Y. Analysis of boundary problem based on zero phase digital filter. *Electron. Meas. Technol.* **2010**, *33*, 25–27. [CrossRef]
45. Yang, S.K.; Liu, H.; Han, G.H.; Yang, L.; Yang, J.X.; Chu, S.B. Application of zero phase filter in power grid traveling wave data processing. *Inf. Technol. Informatiz.* **2009**, *5*, 22–23.
46. Fan, J.Q.; Lv, J.C. Sure Independence Screening for Ultrahigh Dimensional Feature Space. *J. R. Stat. Society. Ser. B* **2008**, *70*, 849–911. [CrossRef] [PubMed]
47. Niu, Y.; Li, H.P.; Liu, Y.H.; Xiong, S.; Yu, Z.; Zhang, R. Overview of Feature Screening Methods for Ultra-high Dimensional Data. *Chin. J. Appl. Probab. Stat.* **2021**, *37*, 69–110. [CrossRef]
48. Yadav, S.; Ankur, S.B. Genetic Algorithm Based Feature Selection for Extreme Learning Machines. *Asian J. Math. Comput. Res.* **2016**, *13*, 34–39.
49. Artur, M. Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features. *Procedia Comput. Sci.* **2021**, *190*, 564–570. [CrossRef]
50. Zhang, Z.; Qiu, H.; Li, W.; Chen, Y. A stacking-based model for predicting 30-day all-cause hospital readmissions of patients with acute myocardial infarction. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 335. [CrossRef]
51. Chen, Q.; Meng, Z.; Liu, X.; Jin, Q.; Su, R. Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE. *Genes* **2018**, *9*, 301. [CrossRef]
52. Mayfield, H.; Smith, C.; Gallagher, M.; Hockings, M. Use of freely available datasets and machine learning methods in predicting deforestation. *Environ. Model Softw.* **2017**, *87*, 17–28. [CrossRef]
53. Zhou, X.Z.; Wen, H.; Zhang, Y.; Xu, J.; Zhang, W. Landslide susceptibility mapping using hybrid random forest with Geo Detector and RFE for factor optimization. *Geosci. Front.* **2021**, *12*, 101211. [CrossRef]
54. Jiang, Y.H.; Wang, T.; Chang, H.W. An overview of hyperspectral image feature extraction. *Electron. Opt. Control.* **2020**, *27*, 73–77. [CrossRef]

55. Wei, X.; Johnson, M.A.; Langston, D.B., Jr.; Mehl, H.L.; Li, S. Identifying Optimal Wavelengths as Disease Signatures Using Hyperspectral Sensor and Machine Learning. *Remote Sens.* **2021**, *13*, 2833. [CrossRef]
56. Lee, M.; Lee, J.H.; Kim, D.H. Gender recognition using optimal gait feature based on recursive feature elimination in normal walking. *Expert Syst. Appl.* **2022**, *189*, 116040. [CrossRef]
57. Li, Q.Q. *Effects of Irrigation Practices on Carbohydrate Accumulation and Translocation and Grain Filling in Rice*; Yang Zhou University: Yangzhou, China, 2020. [CrossRef]
58. He, Q.; Zhang, J.Y.; Han, S.Y.; Wang, X.; Ma, H.W.; Yin, Y.B. Study on Saline-alkali Resistance Rice Varieties Screened by Amylase and Amylopectin Accumulation Dynamics. *Ningxia Acad. Agric. For. Sci.* **2017**, *58*, 4–6.
59. He, Q.; Wang, X.; Ma, H.W.; Feng, W.D.; Zhang, Y.M. Effects of Saline-Alkali Stress on Grain Filling and Panicle Traits of Ningxia Rice. *J. Northeast Agric. Sci.* **2021**, *46*, 11–16, 69. [CrossRef]