agriculture

MDPI

Article

Soil Organic Matter Detection Based on Pyrolysis and Electronic Nose Combined with Multi-Feature Data Fusion Optimization

Xiaomeng Xia [1,2], Mingwei Li [1,2], He Liu [1,2], Qinghui Zhu [1,2] and Dongyan Huang [1,2,*]

[1] Key Laboratory of Bionics Engineering, Ministry of Education, Jilin University, Changchun 130025, China
[2] College of Biological and Agricultural Engineering, Jilin University, Changchun 130025, China
* Correspondence: huangdy@jlu.edu.cn; Tel.: +86-136-107-12601

Abstract: Soil organic matter (SOM) is one of the main sources of plant nutrition and promotes plant growth and development. The content of SOM varies in different areas of the field. In this study, a method based on pyrolysis and electronic nose combined with multi-feature data fusion optimization was proposed to realize rapid, accurate and low-cost measurement of SOM content. Firstly, an electronic nose was used to collect response data from the soil pyrolysis gas, and the sensor features (10 × 6) were extracted to form the original feature space. Secondly, Pearson correlation coefficient (PCC), one-way analysis of variance (One-Way ANOVA), principal component analysis algorithm (PCA), linear discriminant analysis algorithm (LDA), and genetic algorithm-backpropagation neural network algorithm (GA-BP) were used to realize multi-feature data fusion optimization. Thirdly, the optimized feature space was used to train the PLSR models, and the predictive performance of the models were used as an indicator to evaluate different feature optimization algorithms. The results showed that the PLSR model with GA-BP for feature optimization had the best predictive performance (R^2 = 0.90) and could achieve accurate quantitative prediction of SOM content. The dimensionality of the optimized feature space was reduced to 30 and there was no redundancy in the sensor array.

Keywords: pyrolysis; electronic nose; soil organic matter; feature optimization; prediction model

Citation: Xia, X.; Li, M.; Liu, H.; Zhu, Q.; Huang, D. Soil Organic Matter Detection Based on Pyrolysis and Electronic Nose Combined with Multi-Feature Data Fusion Optimization. *Agriculture* **2022**, *12*, 1540. https://doi.org/10.3390/agriculture12101540

Academic Editors: Xiuliang Jin, Hao Yang, Zhenhai Li, Changping Huang and Dameng Yin

Received: 9 August 2022
Accepted: 20 September 2022
Published: 24 September 2022

1. Introduction

Soil organic matter (SOM) is all organic matter containing carbon present in the soil, including plant and animal residues, soil microorganisms and the various organic substances they decompose and synthesize [1,2]. SOM is one of the important indicators of soil nutrient supply capacity and fertility [3,4]. Understanding soil fertility information based on SOM content is one of the most important elements in achieving precision agriculture and promoting sustainable agricultural development. Therefore, there is an urgent need for an accurate, fast, and low-cost determination method to detect the SOM content.

Traditional methods of measuring SOM include the dry burning method, wet burning method and potassium dichromate volumetric method, etc. [5–7]. The potassium dichromate volumetric method is widely used for SOM content measurement because of its accurate measurement results [8]. However, this method needs to be handled by professionals in a chemical laboratory, and there are problems such as time-consuming and laborious, complex operation and pollution. In recent years, near infrared spectroscopy (NIRS) and pyrolysis gas chromatography-mass spectrometry (Py-GC/MS) have been applied to the measurement of SOM content. NIRS allows for non-destructive, efficient, real-time measurement of parameters in large numbers of soil samples [9–12]. However, this method is affected by soil moisture, iron oxide and soil texture. Py-GC/MS is widely used in the analysis of soil constituents because of its rapidity, sensitivity, and low sample

requirement [13,14]. However, this method requires expensive equipment, complex operational procedures and many measurement indicators, which are not conducive to rapid testing in agriculture [15,16].

Electronic nose technology is an integrated detection technology combining sensor technology, signal processing, computer science and pattern recognition, which simulates the process of perception, analysis and recognition of gases by the human olfactory system [17–19]. Electronic nose technology is widely used in food safety, medical analysis, and environmental testing [20–23]. Currently, this technique is used to detect soil characterization and SOM [24–27]. In the practical application of the electronic nose, feature selection and feature dimensionality reduction in the response data are required to better represent the electronic nose response data. Feature optimization reduces non-linearity and correlation between features and removes features that contribute little to modelling, resulting in a feature space with strong recognition ability and less dimensionality [28,29]. Xv et al. extracted the mean differential value, stability value, and area value of sensor response curves to form a feature space and applied it to the pattern recognition of pecan aging time [30]. Wei et al. used the "maximum value", "area value" and "70th s value" methods to extract feature data from the electronic nose response and applied the feature data to peanut quality detection [31]. Cevoli et al. used PCA and four classical feature extraction algorithms to optimize the e-nose response data and tested the classification ability of these algorithms with artificial neural networks (ANN) [32]. Sun et al. used Wilks' lambda statistic, Mahalanobis distance, PCA, linear discriminant analysis and the genetic algorithm to simplify the sensor array and improve the recognition rate for bacteria samples [33]. Zhang et al. used partial least squares regression (PLSR), principal component regression (PCR) and support vector machine regression (SVR) methods to develop a regression model for predicting the population density of Herbst in wheat [34]. Qiu et al. used multivariate statistical methods (LDA and PLSR) and neural networks (random forests and support vector machines) for qualitative classification and quantitative regression, which effectively improved the prediction accuracy of e-nose (E-nose) for fruit juice [35]. Jiang et al. used PCA to visualize the discrimination between pecans based on E-nose data and used the BPNN model and PLSR model to predict pecan storage time and fatty acid content [36].

The authors' group developed a method for measuring the content of soil total nitrogen based on thermal cracking and an artificial olfactory system [37]. Sample rejection methods (MCCV and K-means LOOCV) and feature reduction algorithms (PCA and GA-BP) were applied to solve the problem of prediction accuracy of STN based on manual olfaction, and the results showed that the PLSR prediction model showed the best predictive performance, after optimizing the treatment of the feature space using MCCV and GA-BP methods. However, the method was unable to achieve quantitative prediction of SOM content. In this work, a method based on pyrolysis and electronic nose combined with multi-feature data fusion optimization was proposed to realize the measurement of SOM content. Firstly, the pyrolysis chamber was used to rapidly pyrolyze soil samples. Electronic nose was used to collect response curves of pyrolysis gas from different soil samples, and six features of the sensor response data were extracted to form the original feature space. Secondly, two single feature selection algorithms (PCC, One-Way ANOVA) and three feature dimensionality reduction algorithms (PCA, LDA and GA-BP) were used to optimize the feature space. Thirdly, optimized feature space was used to train the PLSR model and the effectiveness of different feature optimization algorithms were evaluated using the predictive performance of the model as an evaluation metric.

2. Materials and Methods

2.1. Study Area and Soil Samples

The 121 soil samples were collected from different areas of Jilin Province. Jilin province belongs to the central part of Northeast China, bordering Russia and North Korea, in the geographic center of Northeast Asia on the east side of mid-latitude Eurasia. Figure 1 shows the location of Jilin Province and the soil sample sampling locations.

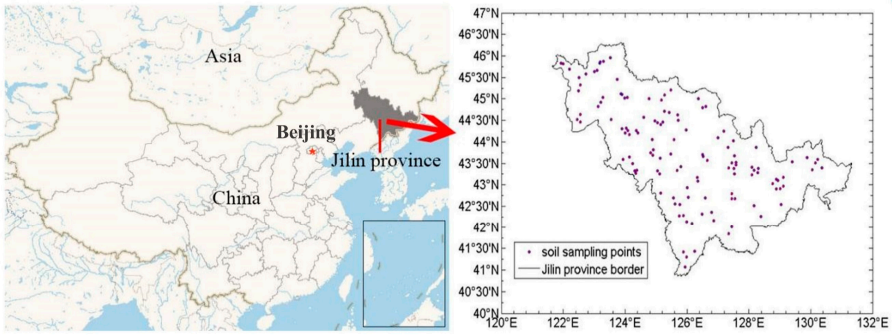

Figure 1. Location of study area and soil sample sampling locations.

To ensure the validity of the soil samples, the sampling sites avoided roadsides, ditch edges and areas where decaying material and manure had accumulated. Sixteen soil samples were collected at a depth of 0–20 cm using an S-shaped spread sampling method and then mixed well. According to the quartering method, 1 kg of soil sample was retained for each sample. The soil samples, stripped of debris such as animal and plant remains and stones, were left to dry indoors in a ventilated place for 3–5 days. The air-dried soil samples were ground and passed through a 1 mm aperture sieve. Depending on the needs of the experiment, each soil sample was split into two, one measured by the potassium dichromate volumetric method and the other by pyrolysis and the electronic nose method.

2.2. Chemical Testing of Soil Samples

The potassium dichromate volumetric method is the standard method for measuring the SOM content. In this study, the organic matter content of 121 soil samples was first measured using the potassium dichromate volumetric method. Each sample was measured three times, and the results were averaged. The SOM content was based on the results of the potassium dichromate method, which was used for subsequent predictive modelling. The measurement results of soil samples were statistically described by SPPSS 24 software (Figure 2). The organic matter of arable soils can be divided into six classes according to their content, i.e., greater than 40 g/kg, 30–40 g/kg, 20–30 g/kg, 10–20 g/kg, 6–10 g/kg and less than (equal to) 6 g/kg. Generally, the best standard for organic matter content in the field is 30–50 g/kg. As shown in Figure 2, the SOM content of 121 soil samples ranged from 6.32 to 78.81 g/kg, the mean value was 31.40 g/kg, the standard deviation was 14.95 g/kg, and the coefficient of variation was 47.61%. The organic matter content of the soil samples approximately obeyed a normal distribution and showed a large variability, which provided a solid basis for establishing a robust model.

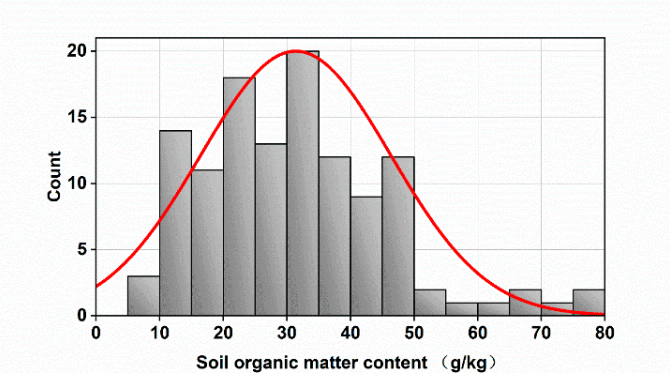

Figure 2. Frequency histogram of SOM content.

2.3. Pyrolysis and Electronic Nose Detection System

2.3.1. Structure of the System

The device used in the paper was a SOM detection system based on pyrolysis and electronic nose, as shown in Figure 3. The system consists of a pyrolysis chamber and electronic nose system. The pyrolysis chamber consists of a pyrolysis furnace, vacuum flange, quartz boat and quartz tube, etc. The electronic nose system is mainly composed a of gas reaction chamber (internally mounted gas sensor array), signal processing circuit, NI data acquisition card and a computer. The gas reaction chamber is rectangular in shape, and it has a volume of 400 cm^3. The gas reaction chamber is made of polypropylene resin, which has good chemical stability and corrosion resistance. The pyrolysis furnace was manufactured by Thermo Scientific Lindberg, USA. The products after soil pyrolysis mainly include alkanes, olefins, aromatics, nitrogen-containing compounds, fatty acids, lignin, phenolic substances, polysaccharides, chitin and other substances [38]. Therefore, 10 kinds of oxide semiconductor gas sensors produced by Figaro were selected to form the gas sensor array. Table 1 shows the names and main parameters of e-nose sensors.

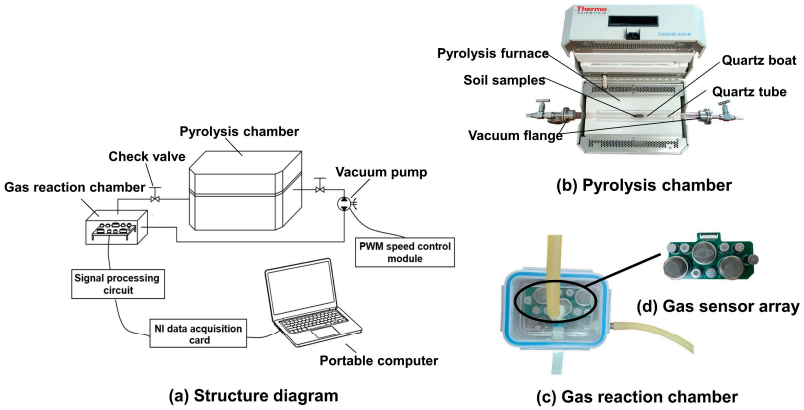

Figure 3. SOM content detection system based on pyrolysis and electronic nose.

Table 1. Gas sensor model and parameters.

Senor Number	Model	Detection of Gas Types	Measuring Range (ppm)
S1	TGS826	Ammonia	30–300
S2	TGS2602	Ammonia, VOC, hydrogen sulfide, etc.	1–30
S3	TGS2610	Butane, LP gas	500–10,000
S4	TGS2620	Ethanol, organic solvent	50–5000
S5	TGS821	Hydrogen	100–1000
S6	TGS2603	Trimethylamine, methyl mercaptan, etc.	1–10
S7	TGS2611	Methane, natural gas	500–10,000
S8	TGS823	Ethanol	50–300
S9	TGS2600	Hydrogen, alcohol, etc.	1–30
S10	TGS2612	Methane, propane, isobutane	3000–9000

2.3.2. Detection Method

When the detection system was working, a quartz boat was used to hold 1.68 g of the soil sample and placed it in the center of the quartz tube. The vacuum flange was closed to keep the pyrolysis chamber in a sealed state. The pyrolysis temperature was set to 384 °C, and the pyrolysis time was 2 min 41 s. After completion of pyrolysis, the LabVIEW detection program was started, the vacuum flange was opened, and the flow rate of the vacuum pump was set at 1 L/min, so that the cracking gas in the cracking room entered the gas reaction room, and the response data of the sensor array to the cracking gas were collected. The acquisition time was 60 s, and the acquisition frequency was 10 Hz.

When data acquisition was complete, the reaction chamber and interconnecting pipes were cleaned with 3 L/min clean air for 2 min to complete the sensor reset, and the quartz boat and quartz tube were washed with water. The measurement was repeated three times for each sample. This system takes 10 min to test a soil sample, half the time of the potassium dichromate volumetric method.

2.3.3. The Response of Sensor Arrays to SOM

In order to verify whether the sensor array composition was reasonable, pyrolysis gas data were selected with a SOM content of 6.32 g/kg and 78.71 g/kg, respectively, and Figure 4 was obtained. As shown in Figure 4, the sensor stabilization time was 40 s. In addition, each sensor showed a large difference in response to different soil gases simultaneously, indicating that the array had good sensitivity to the difference in pyrolysis gases and the sensor array was reasonable.

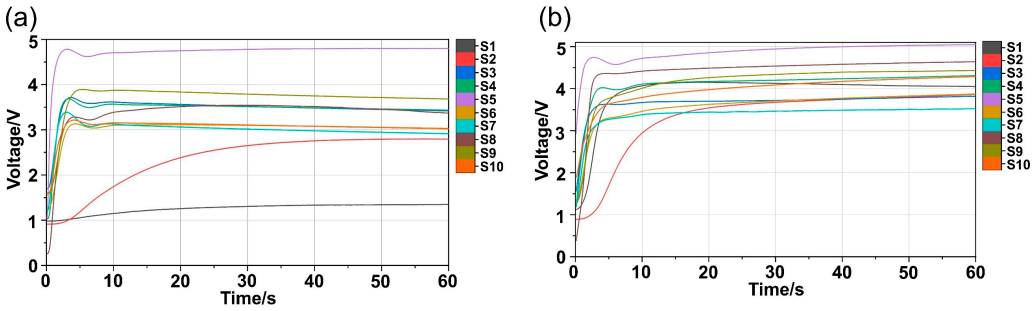

Figure 4. Sensor response curves: (**a**) 6.32 g/kg soil sample; (**b**) 78.71 g/kg soil sample.

2.4. Feature Selection

The appropriate features extracted from the response curves of the sensor array are beneficial for building predictive models with strong generalization ability and high coefficient of determination. The paper extracted the mean value (V_{mean}), the maximum gradient value (V_{mgv}), the maximum value (V_{max}), the response area value (V_{rav}), the 3rd-second transient value ($V_{3\mathrm{s}}$), and the mean differential coefficient value (V_{mdc}) of the sensor response data to construct a feature space. The feature space contained the transient value, the stable value, the dispersion, the overall strength and the change rate of sensor array response curves, which could characterize the response curves of gas sensors. The formula of V_{mean}, V_{mgv}, V_{rav} and V_{mdc} are as follows:

$$V_{mean} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

$$V_{mgv} = \frac{X_{i\max} - X_0}{i} \tag{2}$$

$$V_{rav} = \sum_{i=1}^{N} X_i \triangle t \tag{3}$$

$$V_{mdc} = \frac{1}{N-1} \sum_{i}^{n-1} \frac{X_{i+1} - X_i}{\triangle t} \tag{4}$$

where X_i is the i-th data of the sensor data, $X_{i\max}$ is the maximum value in the sensor data, and X_0 is the initial value of the sensor data, Δt is the interval time between two adjacent collection points, taking 0.1 s, N is the total number of sensor data.

The sensor array was composed of 10 gas sensors and 6 features of each sensor response curve were extracted, combined into a 121 × 10 × 6 original feature space. Table 2 shows the correspondence between each feature and feature number. In order to eliminate the influence of order of magnitude and dimension on modeling, the z-score

standardization method was used to standardize the extracted features. This method allowed the feature space to satisfy the standard normal distribution and is suitable for situations where there are outliers in the sequence that are outside the range of values. The conversion formula is as follows:

$$z_i = \frac{V_i - \mu}{\sigma} \quad (5)$$

where V_i is the feature value (V_{mean}, V_{mgv}, V_{max}, V_{rav}, V_{3s} and V_{mdc}), μ is the mean value of a feature value of a sensor, and σ is the standard deviation of a feature value of a sensor.

Table 2. The correspondence between each feature and feature number.

Feature Parameters	Sensor Number	Feature Number
V_{mean}	S1, S2, S3, ... S9, S10	$E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}$
V_{mgv}	S1, S2, S3, ... S9, S10	$G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8, G_9, G_{10}$
V_{max}	S1, S2, S3, ... S9, S10	$M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9, M_{10}$
V_{rav}	S1, S2, S3, ... S9, S10	$R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}$
V_{3s}	S1, S2, S3, ... S9, S10	$V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9, V_{10}$
V_{mdc}	S1, S2, S3, ... S9, S10	$D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}$

2.5. Training Set and Testing Set Division

In order to build a model that can accurately predict the content of the SOM, the sample set needed to be divided into a training set and a testing set. The training set was used to construct the prediction model and the testing set was used to verify the prediction performance of the model. The Kennard Stone algorithm, which divides the training and testing sets based on Euclidean distance, is an excellent method for dividing sample sets [39,40]. A total of 121 soil samples were collected in this study, and the Kennard–Stone method was used to divide the 85 samples into a training set and the 36 samples into a testing set.

2.6. Feature Optimization Algorithms

2.6.1. Single Feature Ranking Algorithms

The single feature ranking algorithms rank the features according to their importance, remove the unimportant features in turn, and finally obtain the best feature subset. In this study, Pearson correlation coefficients and analysis of variance were used to rank the features of the feature space according to the importance of features.

Pearson correlation coefficient (PCC) is used to measure the correlation between variable X and variable Y, and its value range is −1–1 [41]. When PCC is 1, it indicates that X and Y are positive linear correlation, and all data points fall on a straight line. When PCC is −1, it indicates that X and Y are negative linear correlation. When PCC is 0, it indicates that there is no linear relationship between X and Y.

Analysis of variance (ANOVA), also known as "analysis of variance", was used to test the significance of the difference between the mean of two or more samples [42]. It determines the influence of controllable factors on the research results by analyzing the contribution of variation from different sources to the total variation. In this work, one-way analysis of variance (One-Way ANOVA) was used to analyze the relationship between feature and SOM content.

2.6.2. Feature Dimensionality Reduction Algorithms

In this study, three feature dimensionality reduction algorithms were used to reduce the dimensionality of the feature space in order to obtain a set of features with high recognition ability and low redundant features.

Principal component analysis algorithm (PCA) is one of the most widely used unsupervised dimensionality reduction algorithms. The main idea of PCA algorithm is to map n-dimensional features onto k-dimensional. The reconstructed k-dimensional orthogonal feature based on the original n-dimensional feature is called the principal component [43].

PCA reduces the dimensionality of the feature space by retaining only those features that contain most of the variance and ignoring those that contain almost zero variance.

Linear discriminant analysis algorithm (LDA) is a classical linear learning method and a supervised learning technique for dimensionality reduction [44]. LDA projects the data in a low dimension. After projection, it is hoped that the projection points of each category of data are as close as possible, while the distance between the category centers of different categories of data is as large as possible.

The GA-BP is a combination algorithm of genetic algorithm (GA) and back propagation neural network algorithm (BPNN). GA is a method for searching for optimal solutions by simulating the natural evolutionary process, and it is very suitable for dealing with complex and non-linear optimization problems that are difficult to solve with traditional search algorithms. BPNN has nonlinear mapping ability, adaptive ability and generalization ability, but it is very sensitive to the initial weights and tends to converge to local minima. GA is used to perform a global search for the weights and thresholds of the BPNN and combined with the ability of the local search of the BPNN, the global optimal solution of the problem can be obtained.

2.7. PLSR Model

In this work, partial least squares regression algorithm (PLSR) was selected to establish the regression prediction model of SOM content. PLSR is a multivariate regression analysis method that combines PCA, multiple linear analysis and typical correlation analysis.

2.8. Assessing the Model Performance

The model's performance was evaluated by root mean square error (*RMSE*), coefficient of determination (R^2) and residual prediction deviation (*RPD*). The *RMSE* (Equation (6)) reflects the error between model predictions and observations. The R^2 (Equation (7)) and *RPD* (Equation (8)) both reflect the variation in the response variable explained by the model; however, *RPD* predicts better for non-linear models compared to R^2 and can be used to further measure the predictive performance of the model [45].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{7}$$

$$RPD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}} \tag{8}$$

where y_i is the observed SOM content, $\hat{y}$ is the predicted SOM content, $\overline{y}$ is the mean of observed SOM content, n is the number of data points in the dataset.

3. Results and Analysis

3.1. Feature Optimization Results

3.1.1. The Result of Single Feature Ranking Methods

PCC and One-way ANOVA were used to analyze the correlation between features and SOM content, and the features were ranked in descending order of correlation, as shown in Table 3.

Table 3. Feature importance ranking in single feature ranking methods.

Analytical Method	Feature Importance Ranking
PCC	M_6, D_6, D_2, D_8, M_8, D_7, E_1, M_1, R_1, M_2, D_1, D_4, D_9, E_6, M_{10}, R_{10}, D_3, E_9, R_9, R_6, D_{10}, E_2, M_4, R_2, M_9, E_4, E_{10}, R_4, E_8, R_8, M_5, E_5, R_5, D_5, E_7, R_7, M_3, G_1, G_4, V_1, G_9, G_7, G_3, V_6, V_{10}, M_7, V_5, G_{10}, E_3, R_3, V_3, V_2, V_7, G_6, V_8, V_4, G_8, G_2, V_9, G_5
One-way ANOVA	D_4, D_6, D_9, E_1, R_1, M_6, D_8, V_2, V_1, D_{10}, G_{10}, M_2, E_2, R_2, M_5, V_6, M_4, E_6, R_6, M_1, M_9, D_7, G_2, M_8, D_2, D_1, M_{10}, M_3, R_5, E_5, M_7, G_1, V_9, R_9, E_9, E_4, R_4, E_{10}, R_{10}, D_3, V_{10}, R_8, E_8, V_3, G_9, D_5, V_7, V_5, G_7, G_6, E_3, R_3, V_4, R_7, E_7, G_4, G_5, V_8, G_8, G_3

According to the importance of features, the prediction accuracy of PLSR prediction models established by different single feature ranking methods was obtained, as shown in Figure 5.

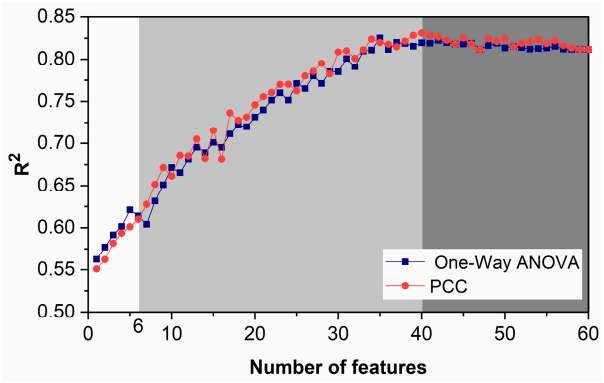

Figure 5. The accuracy of PLSR model established by single feature ranking algorithms.

When fewer features were used, the prediction accuracy of the PLSR model established by One-Way ANOVA was higher than that of PCC. When more than six features were used, the prediction accuracy of the PLSR model tended to increase first and then stabilize as the number of features increased. When the number of features selected was 40, the prediction accuracy of the PLSR model established by PCC was the highest. When the number of features selected was 42, the prediction accuracy of the PLSR model established by One-Way ANOVA was the highest.

The predictive performance of the PLSR models established by the single feature ranking algorithms for SOM content on the training and testing sets is summarized in Table 4. The observed and the predicted SOM content are shown in Figure 6. The model prediction results showed that the R^2 of training and testing sets of the PLSR prediction model established by single feature ranking algorithms was not less than 0.83, *RMSE* was not more than 7.48, and *RPD* was not less than 2.28, indicating that the model had the ability to quantitatively predict SOM content. Compared to the One-Way ANOVA, the testing set of the PLSR prediction model established by PCC showed 1.75% improvement in R^2, 3.93% improvement in *RPD* and 3.78% reduction in *RMSE*, indicating that the model built with PCC was more effective in prediction.

Table 4. The predictive performance of PLSR models established by single feature ranking algorithms.

Single Feature Ranking Method	Dimension of Feature Space	Training Set			Testing Set		
		R^2	*RMSE*	*RPD*	R^2	*RMSE*	*RPD*
PCC	40	0.83	5.75	2.40	0.84	7.19	2.37
One-Way ANOVA	42	0.84	5.51	2.51	0.83	7.48	2.28

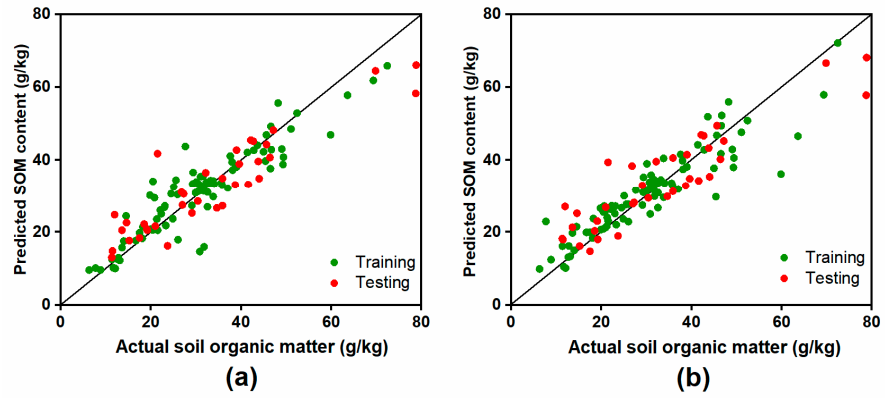

Figure 6. The prediction plots of SOM content by the PLSR model established by (**a**) PCC; (**b**) One-Way ANOVA.

According to the prediction performance of the model and the dimension of feature space, PCC is the best feature optimization method in single feature ranking algorithms. However, the single feature ranking algorithms ignore the connection and correlation between individual features, resulting in a certain amount of redundant information in modelling, which is not conducive to building an efficient and accurate predictive model.

3.1.2. The Result of Feature Dimensionality Reduction Algorithms

PCA was used to optimize the feature space, and the cumulative contribution rate of variance information was set to 95%. The cumulative contribution result of the principal components was obtained, as shown in Figure 7.

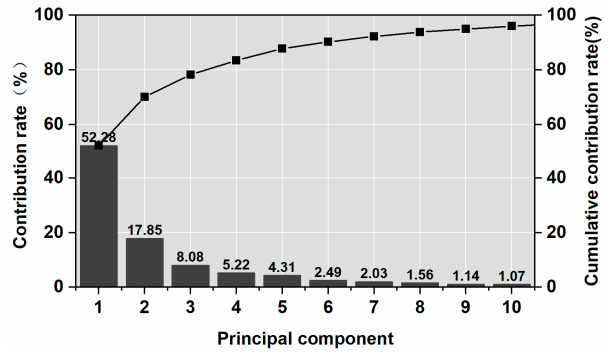

Figure 7. Result of PCA algorithm principal component cumulative contribution.

The variance information contribution rate of the first principal component (α_1), the second principal component (α_2), the third principal component (α_3), etc., and the tenth principal component (α_{10}) were 52.28%, 17.85%, 8.08%, 5.22%, 4.31%, 2.49%, 2.03%, 1.56%, 1.14% and 1.07%, respectively. The cumulative contribution of the variance information of the first 10 principal components reached 96.02%, indicating that 10 principal components can reflect the basic information of the feature space.

The PLSR model was built using the feature space optimized by PCA algorithms, and the predictive performance of the model for SOM content on the training and testing sets is summarized in Table 5. The observed and the predicted SOM content are shown in Figure 8. The R^2 of the training and testing set of the PLSR prediction model established by the PCA algorithm was not less than 0.82, *RMSE* was not more than 7.74, and *RPD* was not less than 2.20, indicating that the model had the ability to quantitatively predict SOM content. In the testing set, when the SOM content was high, the prediction results were lower than the actual values. The reason was that there were relatively few samples with high SOM content in the training set, and soil samples in this content range did not allow the model to be adequately trained, resulting in a reduction in the generalization ability of the model.

Table 5. The predictive performance of PLSR models built with PCA algorithm.

Dimension of Feature Space	Training Set			Testing Set		
	R^2	RMSE	RPD	R^2	RMSE	RPD
10	0.82	5.99	2.31	0.83	7.74	2.20

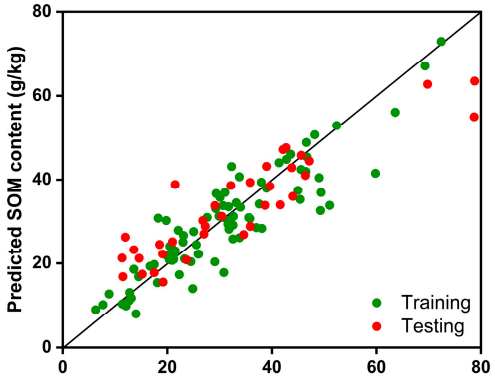

Figure 8. The prediction results of PLSR model established by PCA algorithm.

The SOM data in this study were continuous regression data. To obtain the best dimensionality reduction for LDA, soil samples were divided into 5, 7, 9 and 11 categories according to the SOM content. The predictive performance of the PLSR model established by LDA algorithm for SOM content on the training and testing sets is summarized in Table 6. The observed and the predicted SOM content are shown in Figure 9.

Table 6. The performance parameters for PLSR models established by LDA algorithm.

SOM Category	Dimension of Feature Matrix	Training Set			Testing Set		
		R^2	*RMSE*	*RPD*	R^2	*RMSE*	*RPD*
5	4	0.81	6.19	2.23	0.81	8.57	1.99
7	6	0.84	5.56	2.48	0.84	7.02	2.42
9	8	0.88	4.86	2.84	0.88	5.88	2.89
11	10	0.87	5.08	2.72	0.86	6.35	2.68

When the soil samples were classified into five categories according to SOM content, the model built with the training set was R^2 = 0.81, *RPD* = 2.23, *RMSE* = 6.19, and the testing set R^2 = 0.81, *RPD* = 1.99, *RMSE* = 8.57, indicating that the model had a low predictive performance. The predictive performance of the model was significantly improved when the soil samples were divided into seven categories, with a 3.17% improvement in R^2, 11.31% improvement in *RPD* and 10.16% reduction in *RMSE* for the training set, 3.76% improvement in R^2, 22.00% improvement in *RPD* and 18.03 reduction in *RMSE* for the testing set. When the soil samples were classified into 9 classes, the model built with the training set was R^2 = 0.88, *RPD* = 2.84, *RMSE* = 4.86, and R^2 = 0.88, and the testing set R^2 = 0.88, *RPD* = 2.89, *RMSE* = 5.88. The model achieved the best prediction performance, and the feature space was reduced to 8 dimensions. Compared to 9 classes, when the soil samples were classified into 11 classes according to SOM content, with 1.24% reduction in R^2, 4.22% reduction in *RPD*, and 4.40% improvement in *RMSE* for the model built with the training set, 2.04% reduction in R^2, 7.30% reduction in *RPD*, 7.87% improvement in RMSE for the testing set, indicating that the prediction performance of the model decreased due to over-fitting as the number of features per sample increased.

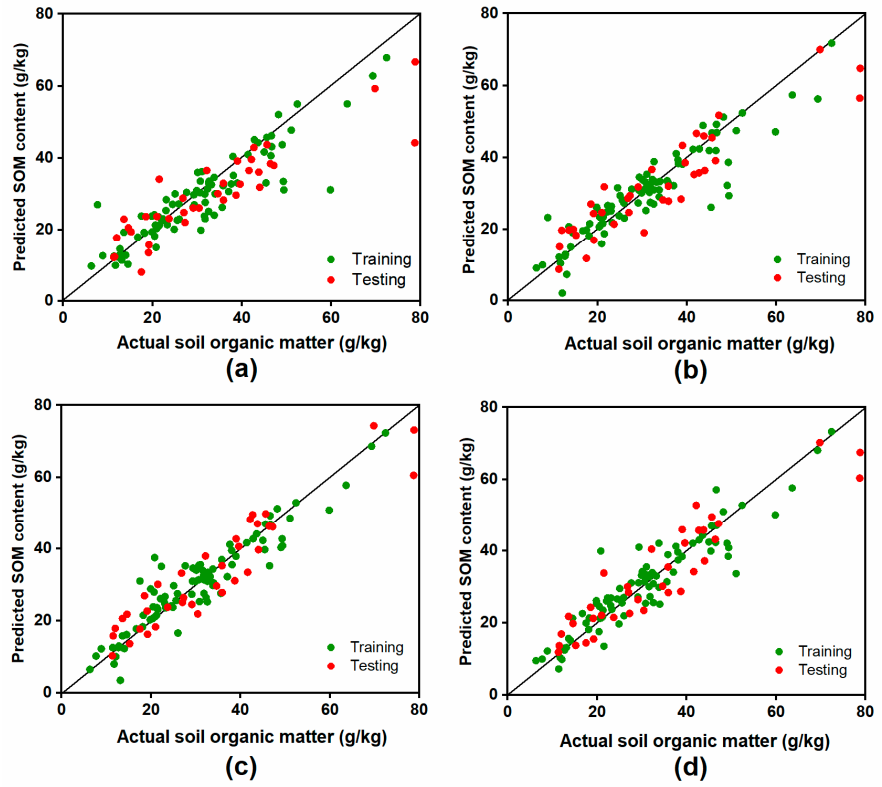

Figure 9. The prediction plots of SOM content by the PLSR model established by LDA after dividing the range of SOM content into (**a**) 5 categories, (**b**) 7 categories, (**c**) 9 categories, (**d**) 11 categories.

GA-BP was used to optimize the feature space, and the population size was set to 50. The individual length was set to 60, and the output condition was set to 100 iterations. After 27 iterations, the best fitness value of the feature space remains unchanged, and the electronic nose feature space achieved the best dimensionality reduction effect. In this case, the selected optimal feature numbers were E_1, E_2, E_3, E_6, E_7, E_8, E_{10}, G_6, G_8, G_9, M_1, M_4, M_6, M_9, M_{10}, R_1, R_2, R_4, R_5, R_6, R_8, R_9, V_2, V_3, V_4, V_8, D_1, D_3, D_6, D_7, and the dimension of the feature was reduced from 60 to 30.

The PLSR model was built using the feature space optimized by GA-BP algorithms, and the predictive performance of the model for SOM content on the training and testing sets is summarized in Table 7. The observed and the predicted SOM content are shown in Figure 10. The R^2 of training and testing set of the PLSR prediction model established by GA-BP algorithm was not less than 0.90, RMSE was not more than 5.98, and RPD was not less than 2.84, indicating that the model could accurately predict the SOM content.

Table 7. The predictive performance of PLSR models built with GA-BP algorithm.

Dimension of Feature Space	Training Set			Testing Set		
	R^2	RMSE	RPD	R^2	RMSE	RPD
30	0.90	4.60	3.00	0.90	5.98	2.85

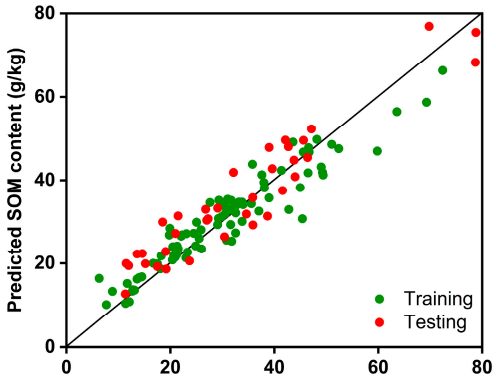

Figure 10. The prediction results of PLSR model established by GA-BP algorithm.

3.2. Comparison of the Results of Different Feature Optimization Algorithms

After optimization by different algorithms, the dimension of feature space and the prediction performance of the PLSR model are shown in Table 8. The single feature ranking methods (PCC and One-way ANOVA) optimized the highest dimension in the feature space (40 and 42, respectively), with PCC optimization being more effective than One-way ANOVA. PCA mapped the 60-dimensional data in feature space to the 10-dimensional space, and the model had R^2 = 0.83, *RPD* = 2.20, *RMSE* = 7.74, indicating that the prediction accuracy of the model was better than PCC. LDA used supervised learning to reduce the feature space to 8 dimensions, and the model had R^2 = 0.88, *RPD* = 2.89, *RMSE* = 5.88, indicating that the prediction accuracy of the model was better than PCA algorithm. GA-BP reduced the dimensionality of the feature space to 30 by using the global search capability of GA and the non-linear mapping capability, strong adaptive capability and generalization capability of BPNN, and the model had R^2 = 0.90, *RPD* = 2.85, *RMSE* = 5.98, indicating that the model predicted the best results and could achieve the quantitative prediction of SOM content. From the above analysis, the GA-BP-optimized feature space performed best.

Table 8. Comparison of optimized results of different feature optimization algorithms.

Feature Selection Algorithms	Dimension of Feature Space	Prediction Performance of PLSR		
		R^2	*RMSE*	*RPD*
PCC	40	0.84	7.19	2.37
One-way ANOVA	42	0.83	7.48	2.28
PCA	10	0.83	7.74	2.20
LDA	8	0.88	5.88	2.89
GA-BP	30	0.90	5.98	2.85

3.3. Results of Feature Optimization Based on GA-BP Algorithm

After optimization by GA-BP, the sensors were ranked according to the contributing feature dimensions and the results are shown in Table 9. The contributions of the six feature parameters to the feature space were as follows: $V_{rav} > V_{mean} > V_{max} > V_{mdc} = V_{3S} > V_{mgv}$, and provided 7, 6, 5, 4 and 3-dimensional features, respectively. The dimension contribution of the V_{mgv} was smallest. The V_{rav}, V_{mean} and V_{max} were important features reflecting the internal relationship between the electronic nose system and the SOM content.

In addition, different sensors contribute different dimensions to the feature space. The TGS2603 sensor provided 5-dimensional features, the TGS826 sensor provided 4-dimensional features, the TGS821 sensor and the TGS2612 sensor provided 2-dimensional features, the TGS2602 sensor provided only 1-dimensional features, and the rest of the sensors provided 3-dimensional features. The gas sensors used in this work all contributed to the composition of the electronic nose feature space, indicating that the sensor array was effective and non-redundant for the detection method of SOM content based on pyroly-

sis and electronic nose. Different sensors had different sensitivities to soil pyrolysis gas, indicating that various gases constituted pyrolysis gas with different concentrations. The TGS2603 sensor and TGS826 sensor contributed the most to the feature space, indicating that ammonia, trimethylamine, methyl mercaptan, etc., had higher concentrations in the pyrolysis gas, and the concentrations of these gases could best represent the SOM content. The results can provide a reference for the further optimization of subsequent sensor arrays.

Table 9. The corresponding relationship between sensor model, feature number and feature parameters after GA-BP optimization.

Sensor Type	Feature Number	Feature Parameters
TGS2603	E_6, G_6, M_6, R_6, D_6	V_{mean}, V_{mgv}, V_{max}, V_{rav}, V_{mdc}
TGS826	E_1, M_1, R_1, D_1	V_{mean}, V_{max}, V_{rav}, V_{mdc}
TGS2610	E_3, R_3, V_3	V_{mean}, V_{rav}, $V_{3\mathrm{s}}$
TGS2620	R_4, V_4, D_4	V_{rav}, $V_{3\mathrm{s}}$, V_{mdc}
TGS2611	E_7, G_7, D_7	V_{mean}, V_{mgv}, V_{mdc}
TGS823	E_8, R_8, V_8	V_{mean}, V_{rav}, $V_{3\mathrm{s}}$
TGS2600	G_9, M_9, R_9	V_{mgv}, V_{max}, V_{rav}
TGS821	M_5, R_5	V_{max}, V_{rav}
TGS2612	E_{10}, R_{10}	V_{mean}, V_{rav}
TGS2602	V_2	$V_{3\mathrm{s}}$

4. Conclusions

This paper proposed a method of soil organic matter (SOM) content detection based on pyrolysis and electronic nose combined with multi-feature data fusion optimization. Two single feature ranking algorithms (PCC and One-Way ANOVA) and three feature dimensionality reduction algorithms (PCA, LDA and GA-BP) were used for multi-feature data fusion and optimization. The PLSR prediction model was established through the optimized feature space, and the predictive performance of the model for SOM content was used as an evaluation metric. The main conclusions were as follows:

(1) Among the two single feature ranking algorithms (PCC, One-Way ANOVA), the feature space optimized by PCC could improve the model prediction performance. The PLSR prediction model established by PCC had R^2 = 0.84, and the dimension of the feature space was 40.

(2) Among the three feature dimensionality reduction algorithms (PCA, LDA, and GA-BP), the feature space optimized by GA-BP could best improve the model prediction performance. The PLSR prediction model established by the GA-BP algorithm had the best prediction performance (R^2 = 0.90, *RPD* = 2.85, *RMSE* = 5.98), and the dimension of the feature space was 30.

(3) Among the feature space optimized by GA-BP, the response area value (V_{rav}), the mean value (V_{mean}) and the maximum value (V_{max}) were important features that reflected the internal relationship between the detection system and SOM content. The TGS2603 sensor and TGS826 sensor contributed the most to the electronic nose feature space, which indicated that ammonia, trimethylamine, methyl mercaptan, etc. had higher concentrations in the pyrolysis gas, and the concentrations of these gases could best represent the level of SOM content. The results can be used as a reference for the further optimization of the sensor array.

Author Contributions: Conceptualization, X.X. and M.L.; methodology, X.X. and M.L.; software, Q.Z. and X.X.; validation, Q.Z.; investigation, M.L. and D.H.; resources, X.X.; visualization, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Science and Technology Development Program of Jilin Province, grant number 20220508113RC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Yu, L.; Hong, Y.; Geng, L.; Nie, Y. Hyperspectral estimation of soil organic matter content based on partial least squares regression. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 103–109. [CrossRef]
2. Zhu, L.; Jia, H.; Chen, Y.; Wang, Q.; Li, M.; Huang, D.; Bai, Y. A Novel Method for Soil Organic Matter Determination by Using an Artificial Olfactory System. *Sensors* **2019**, *19*, 3417. [CrossRef] [PubMed]
3. Lehmann, J.; Kleber, M. The contentious nature of soil organic matter. *Nature* **2015**, *528*, 60–68. [CrossRef] [PubMed]
4. Nowkandeh, S.M.; Noroozi, A.A.; Homaee, M. Estimating soil organic matter content from Hyperion reflectance images using PLSR, PCR, MinR and SWR models in semi-arid regions of Iran. *Environ. Dev.* **2018**, *25*, 23–32. [CrossRef]
5. Wu, C.; Xia, J.; Duan, Z. Review on detection methods of soil organic matter. *Soils* **2015**, *47*, 453–460. [CrossRef]
6. Li, M.; Zhu, Q.; Xia, X.; Liu, H.; Huang, D. Detection Method of Soil Organic Matter Based on Multi-sensor Artificial Olfactory System. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 109–119. [CrossRef]
7. Beltrame, K.K.; Souza, A.M.; Coelho, M.R.; Winkler, T.C.B.; Souza, W.E.; Valderrama, P. Soil Organic Carbon Determination Using NIRS: Evaluation of Dichromate Oxidation and Dry Combustion Analysis as Reference Methods in Multivariate Calibration. *J. Braz. Chem. Soc.* **2016**, *27*, 1527–1532. [CrossRef]
8. Zhao, L.; Liu, X.; Wang, Y.; Ren, T. Thermal analysis determining soil organic matter content and thermal stability. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 105–114. [CrossRef]
9. Zhang, Y.; Hartemink, A.E. Data fusion of vis-NIR and PXRF spectra to predict soil physical and chemical properties. *Eur. J. Soil Sci.* **2020**, *71*, 316–333. [CrossRef]
10. Rakotonindrina, H.; Kawamura, K.; Tsujimoto, Y.; Nishigaki, T.; Razakamanarivo, H.; Andrianary, B.H.; Andriamananjara, A. Prediction of Soil Oxalate Phosphorus using Visible and Near-Infrared Spectroscopy in Natural and Cultivated System Soils of Madagascar. *Agriculture* **2020**, *10*, 5. [CrossRef]
11. Foroughi, H.; Naseri, A.A.; Nasab, S.B.; Hamzeh, S.; Sadeghi, M.; Tuller, M.; Jones, S.B. A new mathematical formulation for remote sensing of soil moisture based on the Red-NIR space. *Int. J. Remote Sens.* **2020**, *41*, 8034–8047. [CrossRef]
12. Shen, Z.; D'Agui, H.; Walden, L.; Zhang, M.; Yiu, T.M.; Dixon, K.; Nevill, P.; Cross, A.; Matangulu, M.; Hu, Y.; et al. Miniaturised visible and near-infrared spectrometers for assessing soil health indicators in mine site rehabilitation. *Soil* **2022**, *8*, 467–486. [CrossRef]
13. Chen, Q.; Wu, Y.; Wu, T.; Si, C.; Zhang, G. Study on the fingerprints of soil organic components in alpine grassland based on Py-GC- MS/MS Technology. *Acta Ecol. Sin.* **2018**, *38*, 2864–2873. [CrossRef]
14. Ma, S.; Chen, Y.; Lu, X.; Wang, X. Soil Organic Matter Chemistry: Based on Pyrolysis-Gas Chromatography-Mass Spectrometry (Py-GC/MS). *Mini Rev. Org. Chem.* **2018**, *15*, 389–403. [CrossRef]
15. Zhang, Z.; Wang, J.; Lyu, X.; Jiang, M.; Bhadha, J.; Wright, A. Impacts of land use change on soil organic matter chemistry in the Everglades, Florida—A characterization with pyrolysis-gas chromatography–mass spectrometry. *Geoderma* **2019**, *338*, 393–400. [CrossRef]
16. Jiang, Y.; Yu, W.; Wang, J.; Kang, H. Analysis of Chemical Composition of Soil Organic Matter Using Pyrolysis Gas Chromatography Mass Spectrometry. *Chin. J. Anal. Chem.* **2018**, *48*, 1526–1534. [CrossRef]
17. Seesaard, T.; Goel, N.; Kumar, M.; Wongchoosuk, C. Advances in gas sensors and electronic nose technologies for agricultural cycle applications. *Comput. Electron. Agric.* **2022**, *193*, 106673. [CrossRef]
18. Peng, P.; Zhao, X.; Pan, X.; Ye, W. Gas Classification Using Deep Convolutional Neural Networks. *Sensors* **2018**, *18*, 157. [CrossRef]
19. Martinez-Garcia, R.; Moreno, J.; Bellincontro, A.; Centioni, L.; Puig-Pujol, A.; Peinado, R.A.; Mauricio, J.C.; Garcia-Martinez, T. Using an electronic nose and volatilome analysis to differentiate sparkling wines obtained under different conditions of temperature, ageing time and yeast formats. *Food Chem.* **2021**, *334*, 127574. [CrossRef]
20. Tozlu, B.H.; Okumuş, H.İ. A new approach to automation of black tea fermentation process with electronic nose. *Automatika* **2018**, *59*, 373–381. [CrossRef]
21. Tatli, S.; Mirzaee-Ghaleh, E.; Rabbani, H.; Karami, H.; Wilson, A.D. Rapid Detection of Urea Fertilizer Effects on VOC Emissions from Cucumber Fruits Using a MOS E-Nose Sensor Array. *Agronomy* **2021**, *12*, 35. [CrossRef]
22. Avian, C.; Mahali, M.I.; Putro, N.A.S.; Prakosa, S.W.; Leu, J.S. Fx-Net and PureNet: Convolutional Neural Network architecture for discrimination of Chronic Obstructive Pulmonary Disease from smokers and healthy subjects through electronic nose signals. *Comput. Biol. Med.* **2022**, *148*, 105913. [CrossRef] [PubMed]
23. MacDougall, S.; Bayansal, F.; Ahmadi, A. Emerging Methods of Monitoring Volatile Organic Compounds for Detection of Plant Pests and Disease. *Biosensors* **2022**, *12*, 129. [CrossRef] [PubMed]
24. Lavanya, S.; Deepika, B.; Narayanan, S.; Krishna Murthy, V.; Uma, M.V. Indicative extent of humic and fulvic acids in soils determined by electronic nose. *Comput. Electron. Agric.* **2017**, *139*, 198–203. [CrossRef]
25. Bieganowski, A.; Jaromin-Glen, K.; Guz, L.; Lagod, G.; Jozefaciuk, G.; Franus, W.; Suchorab, Z.; Sobczuk, H. Evaluating Soil Moisture Status Using an e-Nose. *Sensors* **2016**, *16*, 886. [CrossRef]

26. Zhu, L.; Li, M.; Xia, X.; Huang, D.; Jia, H. Soil Organic Matter Detection Method Based on Artificial Olfactory System. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 171–179. [CrossRef]
27. Huang, D.; Liu, H.; Zhu, L.; Li, M.; Xia, X.; Qi, J. Soil organic matter determination based on artificial olfactory system and PLSR-BPNN. *Meas. Sci. Technol.* **2020**, *32*, 035801. [CrossRef]
28. Han, L.; Chen, M.; Li, Y.; Wu, S.; Zhang, L.; Tu, K.; Pan, L.; Wu, J.; Song, L. Discrimination of different oil types and adulterated safflower seed oil based on electronic nose combined with gas chromatography-ion mobility spectrometry. *J. Food Compos. Anal.* **2022**, *114*, 104804. [CrossRef]
29. Men, H.; Liu, M.; Shi, Y.; Yuan, H.; Liu, J.; Wang, Q. Ultra-lightweight dynamic attention network combined with gas sensor for distinguishing the quality of rice. *Comput. Electron. Agric.* **2022**, *197*, 106939. [CrossRef]
30. Xv, K.; Wang, J.; Deng, F.; Wei, Z.; Cheng, S. Optimization of sensor array of electronic nose for aging time detection of pecan. *Trans. Chin. Soc. Agric. Mach.* **2017**, *33*, 281–287. [CrossRef]
31. Wei, Z.; Wang, J.; Zhang, W. Detecting internal quality of peanuts during storage using electronic nose responses combined with physicochemical methods. *Food Chem.* **2015**, *177*, 89–96. [CrossRef] [PubMed]
32. Cevoli, C.; Cerretani, L.; Gori, A.; Caboni, M.F.; Gallina Toschi, T.; Fabbri, A. Classification of Pecorino cheeses using electronic nose combined with artificial neural network and comparison with GC-MS analysis of volatile compounds. *Food Chem.* **2011**, *129*, 1315–1319. [CrossRef] [PubMed]
33. Sun, H.; Tian, F.; Liang, Z.; Sun, T.; Yu, B.; Yang, S.X.; He, Q.; Zhang, L.; Liu, X. Sensor Array Optimization of Electronic Nose for Detection of Bacteria in Wound Infection. *IEEE Trans. Ind. Electron.* **2017**, *64*, 7350–7358. [CrossRef]
34. Zhang, S.; Han, S.; Xiong, L.; Hou, Y.; Gao, X.; Tang, X. Detection of stored grain pests Tribolium castaneum(Herbst) based on the feature optimization of gas sensor array. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 303–309. [CrossRef]
35. Qiu, S.; Wang, J.; Gao, L. Discrimination and characterization of strawberry juice based on electronic nose and tongue: Comparison of different juice processing approaches by LDA, PLSR, RF, and SVM. *J. Agric. Food Chem.* **2014**, *62*, 6426–6434. [CrossRef]
36. Jiang, S.; Wang, J.; Sun, Y. Qualitative and quantitative analysis of fatty acid profiles of Chinese pecans (Carya cathayensis) during storage using an electronic nose combined with chemometric methods. *RSC Adv.* **2017**, *7*, 46461–46471. [CrossRef]
37. Liu, H.; Zhu, Q.; Xia, X.; Li, M.; Huang, D. Multi-Feature Optimization Study of Soil Total Nitrogen Content Detection Based on Thermal Cracking and Artificial Olfactory System. *Agriculture* **2021**, *12*, 37. [CrossRef]
38. Derenne, S.; Quénéa, K. Analytical pyrolysis as a tool to probe soil organic matter. *J. Anal. Appl. Pyrolysis* **2015**, *111*, 108–120. [CrossRef]
39. He, Z.; Ma, Z.; Li, M.; Zhou, Y. Selection of a calibration sample subset by a semi-supervised method. *J. Near Infrared Spectrosc.* **2018**, *26*, 87–94. [CrossRef]
40. Zhao, M.; Ren, J.; Ji, L.; Fu, C.; Li, J.; Zhou, M. Parameter selection of support vector machines and genetic algorithm based on change area search. *Neural Comput. Appl.* **2011**, *21*, 1–8. [CrossRef]
41. Sun, X.; Shi, Z.; Lei, G.; Guo, Y.; Zhu, J. Multi-Objective Design Optimization of an IPMSM Based on Multilevel Strategy. *IEEE Trans. Ind. Electron.* **2021**, *68*, 139–148. [CrossRef]
42. Tanner-Smith, E.E.; Tipton, E. Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and spss. *Res. Synth. Methods* **2014**, *5*, 13–30. [CrossRef] [PubMed]
43. Crutch, S.J.; Schott, J.M.; Rabinovici, G.D.; Murray, M.; Snowden, J.S. Consensus classification of posterior cortical atrophy. *Alzheimer's Dement.* **2017**, *13*, 870–884. [CrossRef] [PubMed]
44. Tharwat, A.; Gaber, T.; Ibrahim, A.; Hassanien, A.E. Linear discriminant analysis: A detailed tutorial. *AI Commun.* **2017**, *30*, 169–190. [CrossRef]
45. Vohland, M.; Besold, J.; Hill, J.; Fründ, H.-C. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* **2011**, *166*, 198–205. [CrossRef]