Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 243-258
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Rapid Prediction of Wastewater Index Using CNN Architecture and PLS Series Statistical Methods

Qiushuang Mo[1], Lili Xu[2], Fangxiu Meng[1], Shaoyong Hong[3], Xuemei Lin[4*]

[1]School of Mathematics and Statistics, Guilin University of Technology, Guilin, China
[2]College of Marine Sciences, Beibu Gulf University, Qinzhou, China
[3]School of Data Science, Guangzhou Huashang College, Guangzhou, China
[4]Community Management Office, Guilin University of Technology, Guilin, China
Email: *linxuemei0773@foxmail.com

How to cite this paper: Mo, Q.S., Xu, L.L., Meng, F.X., Hong, S.Y. and Lin, X.M. (2024) Rapid Prediction of Wastewater Index Using CNN Architecture and PLS Series Statistical Methods. *Open Journal of Statistics*, **14**, 243-258.
https://doi.org/10.4236/ojs.2024.143012

Received: April 30, 2024
Accepted: May 27, 2024
Published: May 30, 2024

Abstract

Chemical oxygen demand (COD) is an important index to measure the degree of water pollution. In this paper, near-infrared technology is used to obtain 148 wastewater spectra to predict the COD value in wastewater. First, the partial least squares regression (PLS) model was used as the basic model. Monte Carlo cross-validation (MCCV) was used to select 25 samples out of 148 samples that did not conform to conventional statistics. Then, the interval partial least squares (iPLS) regression modeling was carried out on 123 samples, and the spectral bands were divided into 40 subintervals. The optimal subintervals are 20 and 26, and the optimal correlation coefficient of the test set (R_T) is 0.58. Further, the waveband is divided into five intervals: 17, 19, 20, 22 and 26. When the number of joint intervals under each interval is three, the optimal R_T is 0.71. When the number of joint subintervals is four, the optimal R_T is 0.79. Finally, convolutional neural network (CNN) was used for quantitative prediction, and R_T was 0.9. The results show that CNN can automatically screen the features inside the data, and the quantitative prediction effect is better than that of iPLS and synergy interval partial least squares model (SiPLS) with joint subinterval three and four, indicating that CNN can be used for quantitative analysis of water pollution degree.

Keywords

Wastewater, Near-Infrared Spectroscopy, Chemistry Oxygen Demand, Partial Least Squares, Convolutional Neural Network, Statistical Optimization

1. Introduction

Near-infrared spectral band is to be measured in the object of hydrogen-containing groups such as O-H, C-H, and N-H, etc. on the near-infrared combined frequency and octave absorption [1] [2]. Through the near-infrared spectral band, you can indirectly determine and analyze the content of the components in the substance, this method was first discovered by the British scholar Herschel W. Near-infrared spectroscopy of the substance measured can be a solid, liquid, or gas in any kind of form, this method does not contaminate the sample, and the measurement speed is relatively fast and the measurement accuracy is high [3] [4] and [5].

However, in the experimental process, due to the influence of the experimental conditions or the surface distribution of the collected samples is not uniform resulting in errors in the collected NIR spectra. Therefore, on this basis, the screening of samples that do not conform to the conventional statistical distribution is considered, which have a certain impact on the whole, and the NIR spectra will be more consistent with the trend after the removal of these samples. Monte Carlo cross-validation is a method of removing abnormal samples [6] [7], and has been used in many applications in NIR spectral analysis. It builds multiple models and counts the frequency of each sample participating in the modeling of the training set in each model. In general, the frequency of each sample participating in the modeling does not differ greatly, and if some samples do not conform to this pattern, they are identified as abnormal samples.

After removing the abnormal samples, due to the weak absorption signals of the samples in the near-infrared region and the overlapping problem of the spectral bands, the collected spectra of the samples contain not only the information related to the material components, but also irrelevant noise [8] [9]. If it is difficult to achieve the desired effect by applying the full spectral bands of the samples for near-infrared spectral analysis, therefore feature screening can reduce the complexity of the model, select the feature variables related to the target components among the many near-infrared spectral bands, reduce the influence of irrelevant variables on the model, and thus improve the prediction accuracy [10] [11] and [12]. The main methods for feature band screening are Moving Window PLS (MWPLS) [13], Interval PLS (iPLS) [14], and Synergy Interval PLS (SiPLS) [15], which are developed by partial least squares (PLS).

The relationship between spectral variables and dependent variables is often nonlinear, and linear methods alone may not be sufficient to fully explain the relationship between the independent and dependent variables, so a series of nonlinear statistical models representing the relationship between the independent and dependent variables are also used in near-infrared spectral analysis, such as Support Vector Machine (SVM) model, Artificial Neural Network (ANN) model and Convolutional Neural Network (CNN) model. Compared with other nonlinear modeling methods, CNN has been increasingly applied to high-latitude and high-complexity problems in recent years due to their local

sensing and weight-sharing effects, which can reduce the number of parameters and computational complexity during data processing, and are more effective in prediction than traditional nonlinear models and neural networks [16] [17].

Water is a basic natural resource for agricultural production. Water flow provides various elements around the ecosystem for cyclic irrigation and water security is an important issue that affects the quality of agricultural resources for crop cultivation. Water pollution is a growing problem that is likely to undermine the sustainability of the agro-environment [18]. Chemical oxygen demand (COD) of wastewater is a measure of the degree of pollution in wastewater, which is the amount of oxidant consumed when oxidizing with a strong oxidant. The larger the COD value, the stronger oxidant is needed, which represents the more serious pollution. Wang *et al.* explained the relationship between the near-infrared (NIR) spectra and the COD values using PLS model [19]. Zhang *et al.* predicted COD values in ethanol wastewater by backpropagation-artificial neural network model (BP-ANN) and SVM [20].

The experimental data in this paper are the spectral data of wastewater, and the content of COD value in wastewater is used as a predictor to quantitatively predict the degree of wastewater pollution. Monte Carlo cross-validation was used to screen out the abnormal samples, based on which the PLS model was used as the basic model for prediction, and the interval partial least squares and joint interval partial least squares (with the number of joint subintervals of three and four) were used to reduce the modeling variables and improve the prediction effect, and at the same time, the CNN was used to compare the prediction effect of the above two nonlinear models on the COD value.

2. Material and Methods

2.1. Material

In the experiment, 148 industrial wastewater samples were collected and the near-infrared spectra of the target water samples were detected with the FOSS grating spectrometer (NIRSystems 5000 produced in Denmark) with InGaAs accessory standard. The surrounding environment is controlled at constant temperature (25˚C ± 1˚C) and constant humidity (46% ± 1% relative humidity). The samples were detected one by one, and spectra were obtained in the scanning wavelength range of 780 - 2500 nm with a resolution of 2 nm, so the sample spectrum consisted of 860 wavelength variables. The near-infrared absorption spectra of 148 contaminated water samples are shown in **Figure 1**.

2.2. The Architecture of CNN

CNN architecture mainly consists of Convolution layer, Pooling layer, Flatten layer and fully connected layer. Each convolutional layer consists of several convolutional filters to extract feature information from the input data. The weights of the convolution filters are formed by random initialization, and when the convolution operation is performed on the original data, the weights of the

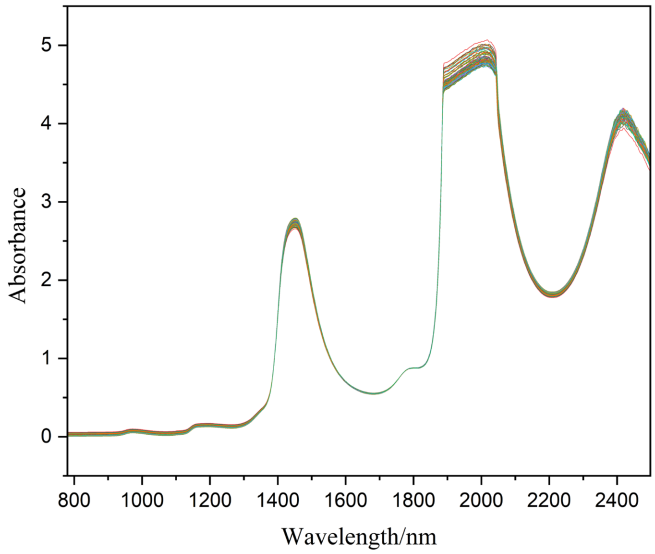

Figure 1. Near-infrared spectra of 148 samples.

convolution filters are multiplied with the original data, and finally summed to get the result. One convolution is performed for each move until all the data is convolved. The output feature map obtained by convolving the feature data by the convolution filter, the feature map is nonlinearly changed after convolution, which makes the convolutional neural network has nonlinear mapping capability. Equation (1), (2) and (3) are ReLU, Sigmoid and tanh activation function.

Since the dimensionality of the data after convolution is too large, the pooling layer reduces the dimensionality of the data while maintaining the useful information of the data. This operation reduces computation time and prevents overfitting [21]. Pooling methods are maximum pooling method and average pooling method. The average pooling method takes the average value within the pooling range as the pooling result, while the maximum pooling method selects the maximum value within the pooling range as the pooling result. Then, the output data from the pooling layer is turned into one-dimensional data through the Flatten layer, and then passed to the fully connected layer to generate output.

$$\mathrm{ReLU} = \begin{cases} x, x > 0 \\ 0, x < 0 \end{cases} \tag{1}$$

$$\mathrm{Sigmoid} = \frac{1}{1+e^{-x}} \tag{2}$$

$$\tanh = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3}$$

During convolutional network training, the batch size (Batch_size) is the number of training samples used during each iteration. When all the small sample datasets are forward and back propagated through the convolutional neural network once is a training rounds (epoch). Epoch improves the computational power by parallel processing, and when training the model, if all the samples are used as the training samples for each training, the computation time of each

epoch is too long. Another critical factor is learning rate, it determines the parameter update step of the model during the training process, too large learning rate may cause the model to oscillate around the optimal solution and fail to converge, while too small may lead to too slow a training process.

2.3. PLS Series Method

PLS is a method that combines principal component analysis and multiple linear regression, which is widely used in the field of chemometrics. It is able to remove the covariance between multivariate variables and extract the information between independent variables and dependent variables to a great extent, which is very effective in the case of small samples and multivariate variables. The principle is as follows:

Suppose there are n samples $x_1, x_2, \ldots, x_n$ with their corresponding dependent variables $y_1, y_2, \ldots, y_n$, then find the link between the independent and dependent variables of the samples. First, extract the corresponding component p_1 of the independent variable, p_1 carries the information in the independent variable, and then extract the component q_1 of the dependent variable, q_1 carries the information in the dependent variable, p_1 and q_1 should have the largest degree of correlation and maximize the extraction of their respective information, and then perform the regression of X on p_1 and the regression of Y on q_1 until the desired precision of the regression is achieved.

The iPLS regression method divides the full near-infrared spectral bands into n intervals, and carries out the partial least square regression analysis in the interval band range, and then output the correlation coefficient between the predicted value and the true value of each band range, the maximum correlation coefficient interval is selected as optimal interval . On the basis of dividing the infrared spectral band into n intervals, the SiPLS regression is to further combine m intervals, so that there are C_n^m intervals in total. The partial least squares regression analysis is carried out on these C_n^m intervals respectively, and the R within each interval range is output, the joint intervals corresponding to largest R the selected as optimal interval, This two methods can reduce the irrelevant information variables and noise during modeling and improve the accuracy and speed of modeling.

2.4. Monte Carlo Cross-Validation

Monte Carlo cross-validation is a statistical simulation method that can detect outliers effectively. This method uses the data difference between outliers and normal samples to screen outliers that do not comply with conventional statistics by building a large number of calibration models. In general, the cumulative frequency ratio of each sample collected is not very different, and if individual samples do not conform to the conventional statistical distribution, it means that these samples have a strong influence on the model, and therefore are not representative of the conventional samples. MCCV also considers the internal rela-

tionship between near-infrared data and reference chemical values to achieve a more comprehensive identification of outlier samples. The specific operations are as follows:

1) The sample is randomly divided into the training and test set at a ratio of 3:1, the training set samples are selected to build the model, and the test set are used to test the prediction ability of the model. This is repeated 1000 times to build a large number of calibration models.

2) Arrange the test set of the model in ascending order of root-mean-square error.

3) Calculate the cumulative frequency of samples in each model participating in the training set in this order.

4) Outlier samples are excluded from samples that do not conform to the cumulative frequency distribution.

The cumulative frequency calculation formula is shown in Equation (4), $\text{sample}(i,k)$ indicating whether the ith sample is a training set sample in the k model. If it is a training set sample, it is 1; otherwise, it is 0. $\text{cummulation\_rate}(i,j)$ represents the first j cumulative frequency of sample i.

$$\text{cummulation\_rate}(i,j)=\sum_{k=1}^{j}\text{sample}(i,k)/j\times 100\% \tag{4}$$

2.5. SPXY Algorithm

SPXY algorithm is developed from KS algorithm. The classical KS algorithm is to select a representative subset of the specified sample size, and ensure that this subset is widely distributed in the data space. KS follows a step-by-step process of selecting new sample selections in areas of sample space that are already far away. To this end, the algorithm uses the Euclidean distance between x vectors of each pair of samples (p, q) to calculate as:

$$d_x(p,q)=\sqrt{\sum_{j=1}^{J}x_p(j)-x_q(j)^2} \tag{5}$$

where $x_p(j)$ and $x_q(j)$ are the response of sample p and sample q at wavelength j, and J is the number of wavelengths. First, the sample with the largest distance (p_1, p_2) is selected as the sample of the calibration set, the algorithm selects the sample with the largest minimum distance relative to the selected sample, and the process is repeated until the number of samples specified for analysis is reached.

KS algorithm uses $d_x(p,q)$ distance, while SPXY algorithm uses $d_{xy}(p,q)$ distance. SPXY is based on KS and the considered dependent variable (y), combining the Euclidean distance of the dependent variable with the Euclidean distance of the independent variable. The distance of the dependent variable y_p and y_q is calculated as follows:

$$d_y(p,q)=\sqrt{(y_p-y_q)^2}=|y_p-y_q| \tag{6}$$

Combining the *x* space and the *y* space, dividing the distance $d_x(p,q)$ and $d_y(p,q)$ by the shortest distance of the sample in the data set, the standardized $d_{xy}(p,q)$ distance is:

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max\left(d_x(p,q)\right)} + \frac{d_y(p,q)}{\max\left(d_y(p,q)\right)} \tag{7}$$

2.6. Model Indicator

The samples were divided into training and test set for follow-up analysis, and the selected samples were further screened for feature bands with greater correlation coefficient R through iPLS and SiPLS. Here, the statistical indicator *R* is shown in Equation (8), where y_i represents the reference chemical value of the *i* sample, y_i' represents the predicted value of the *i* sample, y_m represents the average value of the reference chemical value of all samples, y_m' represents the average value of the predicted value of all samples, and the larger the *R*, the little difference between y_i and y_i', the greater the degree of linear correlation and the better the prediction.

$$R = \frac{\sum_i (y_i - y_m)(y_i' - y_m')}{\sum_i (y_i - y_m)^2 \sum_i (y_i' - y_m')^2} \tag{8}$$

3. Results and Discussion

3.1. iPLS and SiPLS Analysis

MCCV modeling was repeated 1000 times, and 1000 training models and corresponding Root mean squares error of test set (RMSET) were obtained, as shown in Equation (9). The models were sorted in ascending order of RMSET, that is, from good prediction performance to poor prediction performance. Finally, the cumulative frequency of each sample divided into training set was calculated with the sorted model, as shown in **Figure 2**.

$$\text{RMSET} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i')}{n-1}} \tag{9}$$

where y_i represent the reference chemical value of the *i*th sample in test set, y_i' represent the predicted valued of the *i*th sample in test set, *n* is the total sample number in test set.

As can be seen from the cumulative frequency, the probability of each sample gradually approached the normal probability distribution with the increase of the cumulative frequency, the specific value is 0.75 (111/148).

In selecting outliers, we take cumulative frequencies such as the first 20, first 40, and first 60 until the first 1000 frequencies reached. We achieved a total of 50 stages, and samples that did not obey the conventional statistical distribution were screened in these 50 cases. The samples outside the 75% ± 5% range were defined as abnormal samples, and the total number of these samples outside the

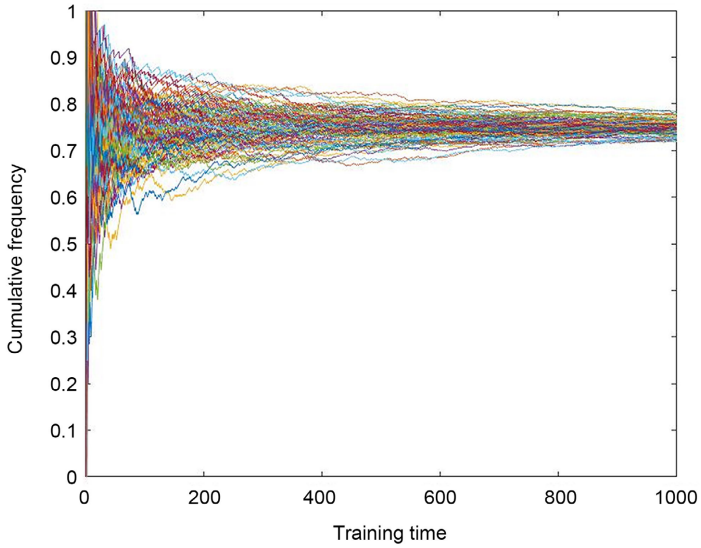

Figure 2. The cumulative frequency variation of samples with the number of trainings.

range was recorded and more than 10 times were used as the threshold to screen outliers, with a total of 25 outlier samples.

The descriptive statistical analysis of the COD values of the original samples and the 123 samples after removed 23 samples that do not meet the statistical distribution after SPXY division into training and test set are shown in **Table 1** and **Table 2**, from which it can be seen that the maximum and minimum values of the training set of the original samples and the 123 samples are the same and the standard deviation of the samples does not differ much, whereas the standard deviation of the test set of the 123 samples is larger than the original samples.

Then, the interval partial least square method was used to screen the feature bands, which divided the full waveband into 40 subintervals. The correlation coefficient of the test set (R_T) corresponding to each subinterval is shown in **Table 3**. As the intervals were divided from 1 to 40, each interval exhibits variation with an increment in the number of intervals, as depicted in **Figure 3**. It can be seen that the optimal R_T is 0.58 under the sub-intervals 20 and 26.

To investigate the impact of combining different bands in partial least squares regression models, we tested 17, 19, 20, 22, and 26 divided intervals. When divided the band into 17 subintervals and joint subintervals of three, the optimal R_T is 0.6324 and the corresponding joint subinterval is [561 - 610, 661 - 710, 811 - 860]. For 19 subintervals and three joint subintervals, the optimal R_T is 0.5333 and the corresponding joint subinterval is [411 - 455, 546 - 590, 681 - 725]. For 20 subintervals and three joint subintervals, the optimal R_T is 0.6981 and the corresponding joint subinterval is [87 - 129, 259 - 301, 431 - 473]. When using 22 subintervals and three joint subintervals, the optimal R_T is 0.7133 and the corresponding joint subintervals is [120 - 158, 159 - 197, 276 - 314]. For 26 subintervals and three joint subintervals, the R_T is 0.6401 and the corresponding

joint subintervals is [287 - 319, 432 - 464, 564 - 596]. The bar chart of optimal values for subintervals 17, 19, 20, 22, and 26, when using three joint subintervals, is shown in Figure 4 and the optimal joint feature bands are listed in Table 4.

Table 1. Statistical result of COD values for 148 samples.

	Number of samples	COD Value			
		Maximum	Minimum	Average	Standard deviation
Training set	111	382	52	218.24	93.74
Test set	37	381	60.0	241.82	96.14

Table 2. Statistical result of COD values for 123 samples.

	Number of samples	COD Value			
		Maximum	Minimum	Average	Standard deviation
Training set	92	382	52	222.89	93.34
Test set	31	381	57	219.65	100.45

Table 3. R_T corresponding to different subintervals.

Subinterval	R_T	Subinterval	R_T
1	0.49	21	0.32
2	0.27	22	0.48
3	0.30	23	0.48
4	0.38	24	0.38
5	0.40	25	0.42
6	0.24	26	0.58
7	0.24	27	0.53
8	0.33	28	0.46
9	0.32	29	0.41
10	0.34	30	0.54
11	0.22	31	0.48
12	0.30	32	0.49
13	0.33	33	0.47
14	0.31	34	0.41
15	0.26	35	0.38
16	0.28	36	0.46
17	0.50	37	0.48
18	0.37	38	0.46
19	0.38	39	0.48
20	0.58	40	0.52

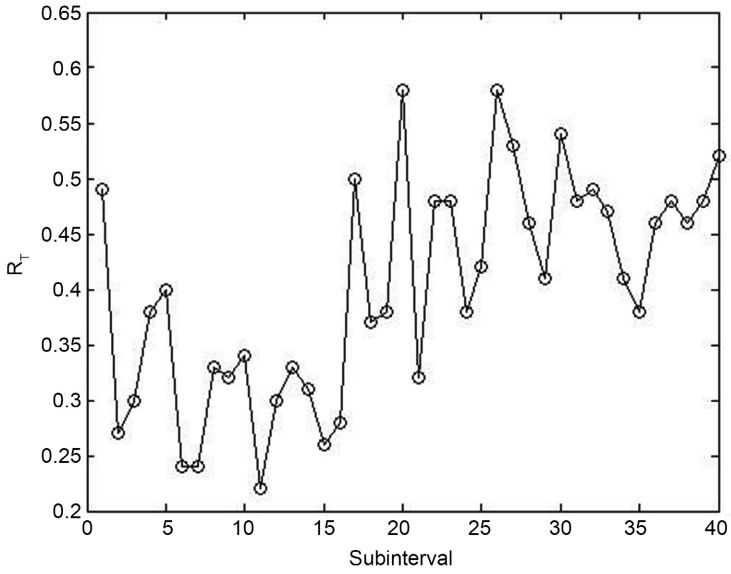

Figure 3. Line graph of R_T corresponding to different subintervals.

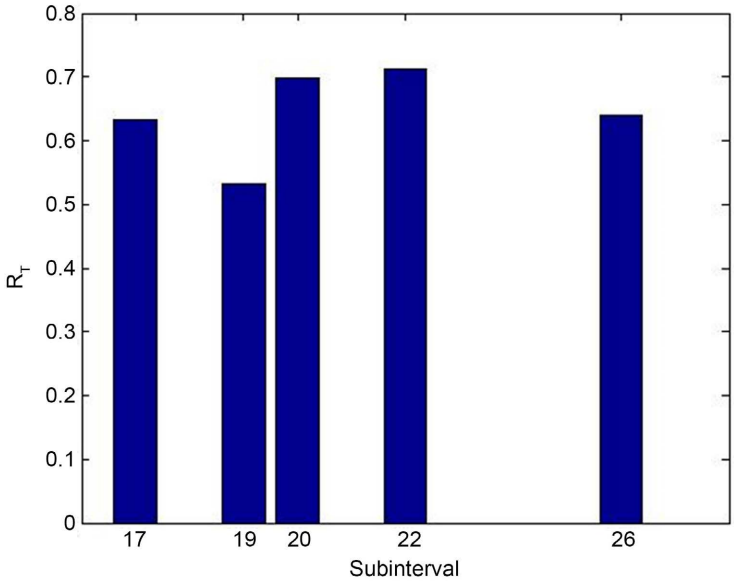

Figure 4. R_T corresponding to subinterval optimal joint band (joint subinterval is three).

Table 4. Optimal joint band corresponding to each divided subinterval when the number of joint subintervals is three.

Number of subintervals	Optimal joint band
17	1900 nm - 1998 nm, 2100 nm - 2198 nm, 2400 nm - 2498 nm
19	1600 nm - 1688 nm, 1870 nm - 1958 nm, 2140 nm - 2228 nm
20	952 nm - 1036 nm, 1296 nm - 1380 nm, 1640 nm - 1724 nm
22	1018 nm - 1094 nm, 1096 nm - 1172 nm, 1330 nm - 1406 nm
26	1312 nm - 1376 nm, 1642 nm - 1706 nm, 1906 nm - 1970 nm

When the number of subintervals is 17 and the number of joint subintervals is four, the optimal R_T is 0.6653, with the corresponding joint subinterval is [103 - 153, 154 - 204, 256 - 306, 460 - 510]. When the number of subintervals is 19 and the number of joint subintervals is four, the optimal R_T improve to 0.6018, and the corresponding joint subinterval is [185 - 229, 411 - 455, 546 - 590, 771 - 815]. When the number of subintervals is 20 and the number of joint subintervals is four, the optimal R_T is slightly higher at 0.6607, with joint subinterval of [1 - 43, 259 - 301, 431 - 473, 560 - 602]. When the number of subintervals is 22 and the number of joint subintervals is four, there is a notable increase in the optimal R_T of 0.7914, and the corresponding joint subinterval is [432 - 470, 549 - 587, 666 - 704, 783 - 821]. When the number of subintervals is 26 and the number of joint subintervals is four, the optimal R_T is 0.6563, and the corresponding joint subinterval is [267 - 299, 432 - 464, 498 - 530, 564 - 596]. When the number of joint sub-intervals is four, the bar chart of optimal R_T for sub-intervals 17, 19, 20, 22, 26 is shown in Figure 5, and the optimal joint feature bands are shown in Table 5.

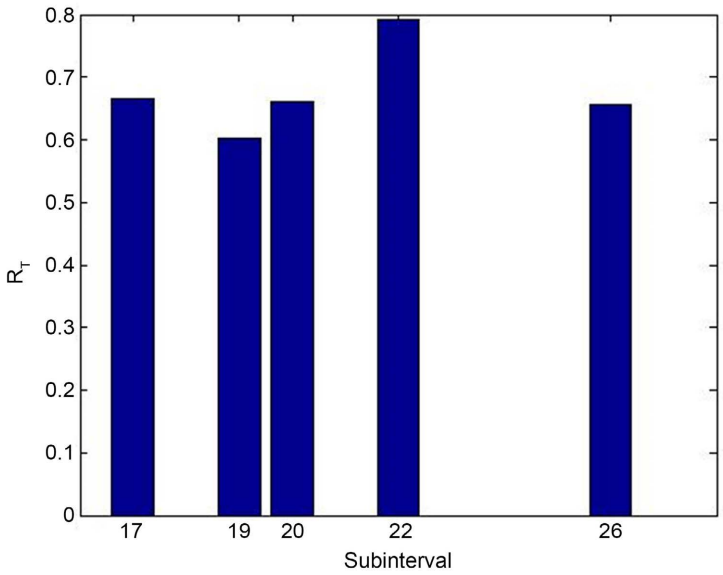

Figure 5. R_T corresponding to subinterval optimal joint band (joint subinterval is four).

Table 5. Optimal joint band corresponding to each divided subinterval when the number of joint subintervals is four.

Number of subintervals	The optimal joint band
17	984 nm - 1084 nm, 1086 nm - 1186 nm, 1290 nm - 1390 nm, 1698 nm - 1798 nm
19	1148 nm - 1238 nm, 1600 nm - 1688 nm, 1870 nm - 1958 nm, 2320 nm - 2408 nm
20	780 nm - 864 nm, 1296 nm - 1380 nm, 1640 nm - 1724 nm, 1898 nm - 1982 nm
22	1642 nm - 1718 nm, 1876 nm - 1952 nm, 2110 nm - 2186 nm, 2344 nm - 2420 nm
26	1312 nm - 1376 nm, 1642 nm - 1706 nm, 1774 nm - 1838 nm, 1906 nm - 1970 nm

3.2. CNN Analysis

In order to compare the prediction effect of SiPLS, CNN was used to predict the COD value of wastewater, the model parameters of CNN are shown in Table 6, the activation function of convolutional and input layer was selected as tanh function, the input value of tanh function was compressed to the range of −1 to 1, which made it more popular than Sigmoid and ReLU in some cases, and the number of training rounds of the model was set at 100 times, Adam optimizer is chosen for the model and the learning rate is set to 0.01.

As shown in Figure 6 below, R shows intense convergence, gradually converged to 1 in rounds 0 - 20 during the training process. This indicates that the model showed significant changes and adjustments in the initial training phase. With the progress of training, the model gradually learned the characteristics of the data, the error value is effectively reduced, and the correlation between the prediction results and the true value was also enhanced. In rounds 20 - 80, the model has learned the characteristics of the data stable, the parameter adjustments were subtle, the model's performance remained more stable, and the prediction accuracy is further improved. Near the end of the training, the model has basically learned the intrinsic laws of the data, the parameter adjustments were smaller, the model's generalization ability is verified, and the prediction accuracy is stabilized at a relatively superior level, the correlation coefficient reached 0.9.

The results of comparing the predicted and true values of the training are shown in Figure 7 below. The dots represented the correspondence between the true and predicted values, and the red line represents the straight line where the true value is equal to the predicted value in the ideal case. By comparing the relationship between the distribution of blue dots and the red line, the fitting effect of the model can be visually assessed. The red line can be used as a baseline reference for comparing the degree of deviation between the true and predicted values. The blue dots represent the true value and the corresponding predicted value for each sample, and their position reflected the accuracy of the model's prediction for each sample. If all the blue dots are distributed on the red line, then it means that the model's prediction is accurate and there is an appreciable linear relationship.

Observing the distribution of blue dots in the scatterplot can help us judge the fitting effect of the model. As shown in the figure, some of the blue dots were concentrated around the red line, but there are still some results distributed in the part far away from the red line, which indicated that the model has a certain degree of accuracy in fitting, and there is still room for improvement, but the correlation coefficient between the predicted value and the true value has already reached 0.9, so further optimization of the CNN parameters is no longer considered here. The results show that the CNN structure can internally feature spectral screening, which reduces the complex process of feature screening for spectra, and the prediction correlation coefficient is also higher than that of SiPLS.

Table 6. CNN structural parameters.

	Filter number	Filter size	Pooling size	Pooling stride	Activation
Convolution	201	8	-	-	tanh
Max Pooling	-		3	1	-
Flatten	-	-	-	-	-
Input	-	-	-	-	tanh
Output	-	-	-	-	Linear

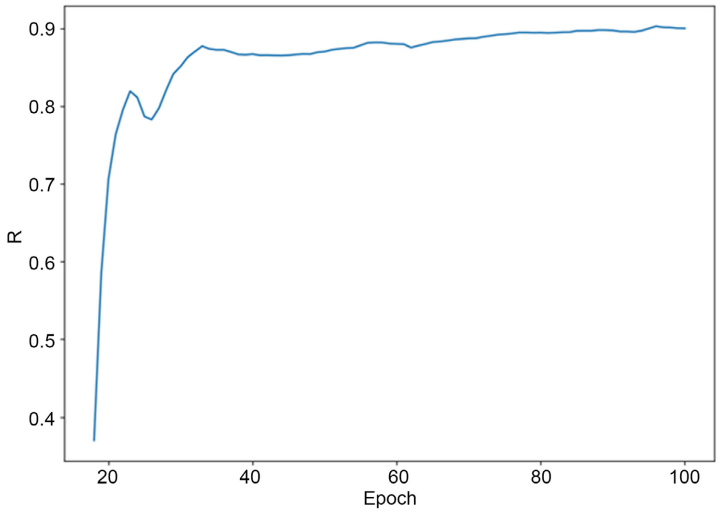

Figure 6. Changes of *R* value in different training rounds.

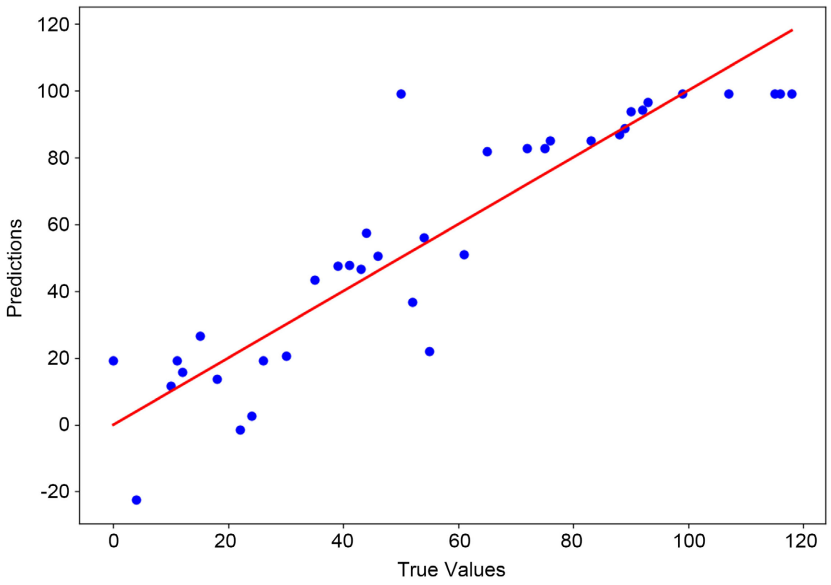

Figure 7. Scatter plot of the predicted and true values.

4. Conclusions

In this paper, after eliminating 25 samples that did not conform to the conventional statistical distribution through Monte Carlo cross-validation, the remain-

ing 123 spectra samples were analyzed to predict the COD content. First, the samples were divided into training and test set by SPXY in the ratio of 7:3, and then iPLS regression analysis was carried out. When 860 bands were divided into 40 sub-intervals, the optimal R_T was obtained under subintervals 20 and 26, and the R_T is 0.58. Further, in order to reduce the interference of unnecessary feature bands, the wavebands were divided into five subintervals: 17, 19, 20, 22 and 26. In these subintervals, we conducted SiPLS with a number of joint subintervals three or four. When the number of joint subintervals is three, the optimal R_T is 0.71, and the corresponding joint band is [1018 nm - 1094 nm, 1096 nm - 1172 nm, 1330 nm - 1406 nm]. When the number of joint subintervals is four, the optimal R_T is 0.79, and the corresponding joint band is [1642 nm - 1718 nm, 1876 nm - 1952 nm, 2110 nm - 2186 nm, 2344 nm - 2420 nm]. This indicated that the joint subintervals can reduce unnecessary information interference of feature bands and improve the prediction effect. Finally, the R_T of COD prediction by CNN model is 0.9, and the results showed that CNN has a better prediction effect than iPLS and SiPLS methods and can be further generalized to quantitative near-infrared spectroscopy.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62365008) and special project in key areas of ordinary universities and colleges in Guangdong Province (2023ZDZX4069).

Conflicts of Interest

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

References

[1] Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez C., Edmond, A. and Jent, N. (2007) A Review of Near Infrared Spectroscopy and Chemometrics in Pharmaceutical Technologies. *Journal of Pharmaceutical and Biomedical Analysis*, **44**, 683-700. https://doi.org/10.1016/j.jpba.2007.03.023

[2] Chadha, R. and Haneef, J. (2015) Near-Infrared Spectroscopy: Effective Tool for Screening of Polymorphs in Pharmaceuticals. *Applied Spectroscopy Reviews*, **50**, 565-83. https://doi.org/10.1080/05704928.2015.1044663

[3] Li, X.Y., Chen, H.Z., Xu, L.L., Mo, Q.S., Du, X.R. and Tang, G.Q. (2024) Multi-Model Fusion Stacking Ensemble Learning Method for the Prediction of Berberine by FT-NIR Spectroscopy. *Infrared Physics & Technology*, **137**, Article 105169. https://doi.org/10.1016/j.infrared.2024.105169

[4] Cui, P.D., Wang, Q.Y., Li, Z., Wu, C.L., Li, G., Zhao, J., *et al.* (2022) A Feasibility Study on Improving the Non-Invasive Detection Accuracy of Bottled Shuanghuanglian Oral Liquid Using Near Infrared Spectroscopy. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **274**, Article 121120. https://doi.org/10.1016/j.saa.2022.121120

[5] Radney, J.G. and Zangmeister, C.D. (2015) Measurement of Gas and Aerosol Phase

Absorption Spectra across the Visible and Near-IR Using Supercontinuum Photoacoustic Spectroscopy. *Analytical Chemistry*, **87**, 7356-7363. https://doi.org/10.1021/acs.analchem.5b01541

[6] Qu, Y.X. and Cai, Z.Y. (2018) Identification of Singular Samples in Near Infrared Spectrum of Starch Water Content Prediction by Using Monte Carlo Cross Validation Combined with T Test. *IOP Conference Series*: *Earth and Environmental Science*, **186**, Article 012035. https://doi.org/10.1088/1755-1315/186/3/012035

[7] Ye, D.D., Sun, L.J., Zou, B., Zhang, Q., Tan, W.Y. and Che, W.K. (2018) Non-Destructive Prediction of Protein Content in Wheat Using NIRS. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **189**, 463-472. https://doi.org/10.1016/j.saa.2017.08.055

[8] Ling, M.X., Bian, X.H., Wang, S.S., Huang, T., Liu, P., Wang, S.Y., *et al.* (2022) A Piecewise Mirror Extension Local Mean Decomposition Method for Denoising of Near-Infrared Spectra with Uneven Noise. *Chemometrics and Intelligent Laboratory Systems*, **230**, Article 104655. https://doi.org/10.1016/j.chemolab.2022.104655

[9] Liu, C., Yang, S.X., Li, X.F., Xu, L.J. and Deng, L. (2020) Noise Level Penalizing Robust Gaussian Process Regression for NIR Spectroscopy Quantitative Analysis. *Chemometrics and Intelligent Laboratory Systems*, **201**, Article 104014. https://doi.org/10.1016/j.chemolab.2020.104014

[10] Zhao, Z.N., Liu, Y.H., Yang, S., Li, Y.R., Zhang, Y.S. and Yan, H. (2023) Fast Detection of the Tenderness of Mulberry Leaves by a Portable Near-Infrared Spectrometer with Variable Selection. *Infrared Physics & Technology*, **133**, Article 104818. https://doi.org/10.1016/j.infrared.2023.104818

[11] Nawar, S., Mohamed, E.S., Sayed S.E.-E., Mohamed, W.S., Rebouh, N.Y. and Hammam, A.A. (2023) Estimation of Key Potentially Toxic Elements in Arid Agricultural Soils Using Vis-NIR Spectroscopy with Variable Selection and PLSR Algorithms. *Frontiers in Environmental Science*, **11**, Article 1222871. https://doi.org/10.3389/fenvs.2023.1222871

[12] Chen, M.J., Yin, H.L., Liu, Y., Wang, R.R., Jiang, L.W. and Li, P. (2022) Non-Destructive Prediction of the Hotness of Fresh Pepper with a Single Scan Using Portable Near Infrared Spectroscopy and a Variable Selection Strategy. *Analytical Methods*, **14**, 114-124. https://doi.org/10.1039/D1AY01634B

[13] Zhang, L.N., Tian, H., Wang, L.R., Li, H. and Pu, Z.Y. (2023) Selection and Validation of the Best Detection Location for Hemoglobin Determination by Spatially Resolved Diffuse Transmission Spectroscopy. *Infrared Physics & Technology*, **133**, Article 104839. https://doi.org/10.1016/j.infrared.2023.104839

[14] Marañón, M., Fernández-Novales, J., Tardaguila, J., Gutiérrez, S. and Diago, M.P. (2023) NIR Attribute Selection for the Development of Vineyard Water Status Predictive Models. *Biosystems Engineering*, **229**, 167-178. https://doi.org/10.1016/j.biosystemseng.2023.04.001

[15] Miao, X.X., Miao, Y., Liu, Y., Tao, S.H., Zheng, H.B., Wang, J.M., *et al.* (2023) Measurement of Nitrogen Content in Rice Plant Using Near Infrared Spectroscopy Combined with Different PLS Algorithms. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **284**, Article 121733. https://doi.org/10.1016/j.saa.2022.121733

[16] Wang, J.B., Lu, Z.X., Wang, G.M., Hussain, G., Zhao, S.H., Zhang, H.J., *et al.* (2023) Research on Fault Diagnosis of HMCVT Shift Hydraulic System Based on Optimized BPNN and CNN. *Agriculture*, **13**, Article 461. https://doi.org/10.3390/agriculture13020461

[17] Harini Chandana, S. and Senthil Kumar, R. (2022) A Deep Learning Model to Identify Twins and Look Alike Identification Using Convolutional Neural Network (CNN) and to Compare the Accuracy with SVM Approach. *ECS Transactions*, **107**, Article 14109. https://doi.org/10.1149/10701.14109ecst

[18] Kotir, J.H., Smith C., Brown G., Marshall, N. and Johnstone, R. (2016) A System Dynamics Simulation Model for Sustainable Water Resources Management and Agricultural Development in the Volta River Basin, Ghana. *Science of the Total Environment*, **573**, 444-457. https://doi.org/10.1016/j.scitotenv.2016.08.081

[19] Wang, X.D., Ratnaweera, H., Holm, J.A. and Olsbu, V. (2017) Statistical Monitoring and Dynamic Simulation of a Wastewater Treatment Plant: A Combined Approach to Achieve Model Predictive Control. *Journal of Environmental Management*, **193**, 1-7. https://doi.org/10.1016/j.jenvman.2017.01.079

[20] Zhang, L.H., Chao, B. and Zhang, X. (2020) Modeling and Optimization of Microbial Lipid Fermentation from Cellulosic Ethanol Wastewater by *Rhodotorula glutinis* Based on the Support Vector Machine. *Bioresource Technology*, **301**, Article 122781. https://doi.org/10.1016/j.biortech.2020.122781

[21] Lee, C.Y., Gallagher, P. and Tu, Z.W. (2018) Generalizing Pooling Functions in CNNs: Mixed, Gated, and Tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 863-875. https://doi.org/10.1109/TPAMI.2017.2703082