



Article

# Simultaneous Determination of Metal Ions in Zinc Sulfate Solution Using UV–Vis Spectrometry and SPSE-XGBoost Method

Fei Cheng <sup>1</sup>, Chunhua Yang <sup>1</sup>, Can Zhou <sup>1,2,\*</sup> , Lijuan Lan <sup>1</sup>, Hongqiu Zhu <sup>1</sup>  and Yonggang Li <sup>1</sup>

<sup>1</sup> School of Automation, Central South University, Changsha 410083, China; feicheng@csu.edu.cn (F.C.); ychh@csu.edu.cn (C.Y.); lijuan.lan@csu.edu.cn (L.L.); hqcsu@csu.edu.cn (H.Z.); liyonggang@csu.edu.cn (Y.L.)

<sup>2</sup> State Key Laboratory of High Performance Complex Manufacturing, Changsha 410083, China

\* Correspondence: zhocan@csu.edu.cn; Tel.: +86-731-8883-0700

Received: 8 July 2020 ; Accepted: 25 August 2020; Published: 31 August 2020



**Abstract:** Excessive discharge of heavy metal ions will aggravate environment pollution and threaten human health. Thus, it is of significance to real-time detect metal ions and control discharge in the metallurgical wastewater. We developed an accurate and rapid approach based on the singular perturbation spectrum estimator and extreme gradient boosting (SPSE-XGBoost) algorithms to simultaneously determine multi-metal ion concentrations by UV–vis spectrometry. In the approach, the spectral data is expanded by multi-order derivative preprocessing, and then, the sensitive feature bands in each spectrum are extracted by feature importance (VI score) ranking. Subsequently, the SPSE-XGBoost model are trained to combine multi-derivative features and to predict ion concentrations. The experimental results indicate that the developed “Expand-Extract-Combine” strategy can not only overcome problems of background noise and spectral overlapping but also mine the deeper spectrum information by integrating important features. Moreover, the SPSE-XGBoost strategy utilizes the selected feature subset instead of the full-spectrum for calculation, which effectively improves the computing speed. The comparisons of different data processing methods are conducted. It outcomes that the proposed strategy outperforms other routine methods and can profoundly determine the concentrations of zinc, copper, cobalt, and nickel with the lowest RMSE<sub>P</sub>. Therefore, our developed approach can be implemented as a promising mean for real-time and on-line determination of multi-metal ion concentrations in zinc hydrometallurgy.

**Keywords:** zinc hydrometallurgy; metal ion measurement; UV–vis spectroscopy; feature selection and combination; singular perturbation spectrum estimator; extreme gradient boosting

## 1. Introduction

Zinc metal smelting wastewater contains multiple toxic metal ions, such as zinc, copper, cobalt, and nickel. Irrational discharge of heavy metal ions will cause serious harm to the ecological environment [1,2]. At present, the concentrations of metal ions are mostly acquired via off-line analysis in the laboratory, which is laborious, time-consuming, connects with many errors and chemical costs, and leads to blind control of wastewater discharge. Hence, the real-time and accurate detection of metal ions is urgently needed [3].

As for better online monitoring methods, optical detection methods are widely used because of its high efficiency and low laboriousness, such as ultraviolet-visible (UV–vis) spectroscopy [4,5], atomic absorption spectroscopy (AAS) [6], near-infrared spectroscopy (NIRS) [7], surface-enhanced raman spectroscopy (SERS) [8], laser-induced breakdown spectroscopy (LIBS) [9], and so on. Among these, the UV-vis spectrophotometry can achieve online analysis on multi-ions without expensive

sample pretreatment and is easily operated, making it cheaper and faster in the applications [10–12]. Our previous work focused on detecting copper and cobalt concentrations using UV–vis spectroscopy and multivariate regression model based on the wavelet denoising and locally weighted partial least squares methods [13,14]. However, due to the complex background and similar chemical properties of detected ions, the spectra are excessively overlapped and exist severe nonlinearity. The denoising method and regression model based on full spectrum will become invalid. The characteristic information of each ion is difficult to distinguish and extract when the ion species increased. Moreover, the external environment unavoidably generates noise interference, resulting in inconsistent intensity of spectral signals. All these problems make it arduous for the spectral quantitative analysis of complex mixed solution and seriously restrict the application of spectral technology.

To establish a quantitative analysis model, the works of predecessors can be roughly divided into three parts: spectral preprocessing, feature selection, and multivariate calibration. For spectral preprocessing, the commonly used methods are denoising and derivatives. The derivative method can reconstruct the spectral peak and eliminate the background signal interference [15,16]. Moreover, the ability to distinguish subtle changes in similar spectra is considerably enhanced in the derivative spectrum [17]. However, most studies usually select a single derivative approach, which may not be sufficient for analysis of severe overlapped spectra. Li et al. proposed the singular perturbation spectral estimator (SPSE) based on the singular perturbation technique and Taylor series to obtain high quality derivative spectra from the measured spectrum with noise [18–20]. Since the obtained spectrum is relatively simple, lacks detail information, and contains a large amount of arbitrary noises, it is vital to adopt diversified preprocessing methods that provide abundant and accurate information.

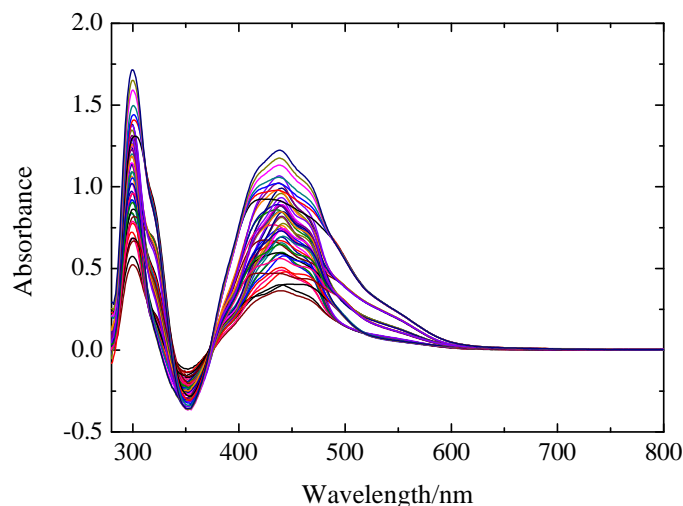
Frequently-used feature selection methods in spectroscopy are uninformative variable elimination (UVE) [21] and competitive adaptive reweighted sampling (CARS) [22]. It is considered that variable combination has a great influence on prediction performance [23]. Even when the subsets containing less important variables are combined, they can achieve a good predictive performance [24,25]. Therefore, the idea of variable combination is introduced into the spectrum analysis. For multivariate calibration, the commonly applied approaches are the linear method (e.g., partial least squares (PLS)) and the nonlinear modeling method (e.g., support vector machine (SVM)) [26,27]. At present, ensemble learning becomes a common technology to enhance the generalization ability by combining the prediction results of multiple base learners [28–30]. Extreme gradient boosting (XGBoost) is an iconic ensemble learning algorithm proposed by Chen et al. [31]. XGBoost has many advantages in processing nonlinear data and can extract features from variables containing noise and redundant information. Numerous studies demonstrate that it has promoted prediction accuracy and performed remarkable results for spectral analyses in different domains [32–34]. However, there are still few works to incorporate this sophisticated strategy into spectral quantitative analysis of heavy metal ions in solution.

Motivated by the above factors, this article introduces SPSE and XGBoost into UV–vis spectrometry for the first time to measure multi-metal ion concentrations. In view of the redundant noise and intricate correlation, the SPSE is employed to expand the multi-order derivative spectra with high accuracy and strong resistance of disturbance. The ensemble XGBoost model is used to extract the feature variables and rank the importance score. The sensitive feature bands in each spectrum are integrated to form new characteristic variable sets and the ion concentrations are predicted. Afterwards, the multi-derivative feature subset combination is considered to further promote the prediction precision. Finally, to validate the performance of the “Expand-Extract-Combine” strategy in SPSE-XGBoost, the comprehensive analyses among CARS-PLS, UVE-LS-SVM, and XGBoost are carried out. The remainder of this article is organized in the following sections. Section 2 describes the experimental procedure, in which the basic concepts of SPSE and XGboost are given, respectively. Then the proposed modeling framework and procedure are introduced. In Section 3, the validation of comparative results and overall performance of each model are discussed. Conclusions are drawn in Section 4.

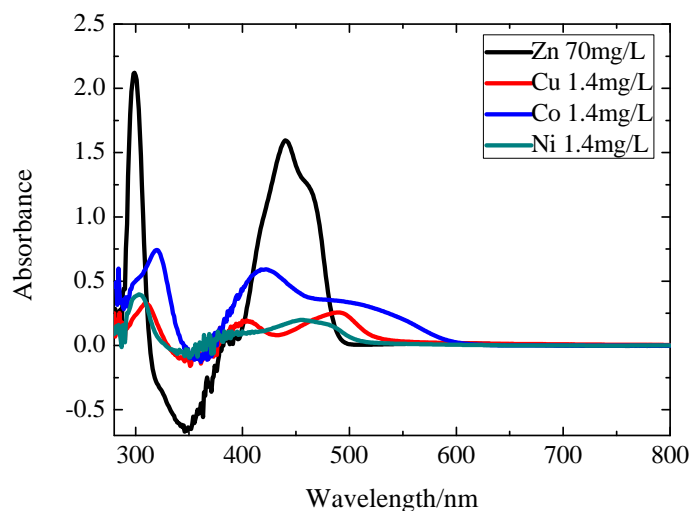
## 2. Materials and Methods

### 2.1. Experimental Apparatus and Samples

A T9 UV–vis spectrophotometer (Beijing Purkinje General Instrument Co., Ltd., Beijing, China) is used to measure the spectrum. The T9 spectrophotometer utilizes a high-performance xenon lamp and double beam optical system, which can achieve spectral scanning over a wide wavelength range of 185 nm to 900 nm. A computer (Lenovo Group, Beijing, China) receives the spectral data via a UV-Win Software (Beijing Purkinje General Instrument Co., Ltd., Beijing, China). UV-Win software provides complete instrument control and a set of mathematical tools to analyze the measurement results.



**Figure 1.** The original absorption spectrum of 49 samples for training and testing.



**Figure 2.** The single ion absorption spectrum for zinc, copper, cobalt, and nickel, respectively.

The main metal ions in the hydrometallurgy wastewater of Zhuzhou Smelter Company are Zn(II), Cu(II), Co(II), and Ni(II), in which the concentration of Zn(II) is 20–250 times that of the other metal ions. Huge difference of ion concentrations can lead to inconsistent intensity of spectral and severe masking problems. Hence, it is extremely significant to select appropriate experimental reagents to ensure the precision of simultaneous determination for the multiple ions. The reagents and their optimized dosage are as follows: 0.4% Nitroso R salt chromogenic agent solution: 2.5 mL; HAc-NaAc buffer solution: pH = 5.5, 5 mL. The concentrations of zinc standard solutions are 1 g/L. Copper, cobalt, and nickel standard solutions are all 12.5 mg/L. All reagents are of analytical grade and added in a 25

mL colorimetric tube. The specific operation procedure of the experiment is as follows: In the 25 mL colorimetric tube, add 5.0 mL HAc-NaAc buffer solution. Add zinc standard solution and proper amount of copper, cobalt, nickel ion standard solution. Add 2.5 mL 0.40% Nitroso R salt chromogenic solution, shake well to make the chromogenic reaction fully react. Add distilled water to make the volume up to 25 mL. Prepare reagent blank. Adjust the instrument to zero. Place the sample in a 1 cm quartz cuvette, and use the reagent blank as a reference.

**Table 1.** The concentration of Zn(II), Cu(II), Co(II), and Ni(II) of 49 samples (mg/L).

NO.	Zn(II)	Cu(II)	Co(II)	Ni(II)	NO.	Zn(II)	Cu(II)	Co(II)	Ni(II)
1	10	0.2	0.4	0.6	26	40	1.0	0.6	0.2
2	10	0.4	0.8	1.2	27	40	1.2	1.0	0.8
3	10	0.6	1.2	0.4	28	40	1.4	1.4	1.4
4	10	0.8	0.2	1.0	29	50	0.2	0.4	0.6
5	10	1.0	0.6	0.2	30	50	0.4	0.8	1.2
6	10	1.2	1.0	0.8	31	50	0.6	1.2	0.4
7	10	1.4	1.4	1.4	32	50	0.8	0.2	1.0
8	20	0.2	0.4	0.6	33	50	1.0	0.6	0.2
9	20	0.4	0.8	1.2	34	50	1.2	1.0	0.8
10	20	0.6	1.2	0.4	35	50	1.4	1.4	1.4
11	20	0.8	0.2	1.0	36	60	0.2	0.4	0.6
12	20	1.0	0.6	0.2	37	60	0.4	0.8	1.2
13	20	1.2	1.0	0.8	38	60	0.6	1.2	0.4
14	20	1.4	1.4	1.4	39	60	0.8	0.2	1.0
15	30	0.2	0.4	0.6	40	60	1.0	0.6	0.2
16	30	0.4	0.8	1.2	41	60	1.2	1.0	0.8
17	30	0.6	1.2	0.4	42	60	1.4	1.4	1.4
18	30	0.8	0.2	1.0	43	70	0.2	0.4	0.6
19	30	1.0	0.6	0.2	44	70	0.4	0.8	1.2
20	30	1.2	1.0	0.8	45	70	0.6	1.2	0.4
21	30	1.4	1.4	1.4	46	70	0.8	0.2	1.0
22	40	0.2	0.4	0.6	47	70	1.0	0.6	0.2
23	40	0.4	0.8	1.2	48	70	1.2	1.0	0.8
24	40	0.6	1.2	0.4	49	70	1.4	1.4	1.4
25	40	0.8	0.2	1.0					

In this study, 49 groups of mixed solutions are analyzed. Figure 1 shows the original spectra of the 49 mixed ion solutions. The measured UV–vis spectrum wavelength ranges from 280 nm to 800 nm with 1 nm scanning resolution. Each sample is scanned three times and the averaged spectrum is obtained for calculation. Among them, the concentration range of Zn(II) is between 10 mg/L and 70 mg/L and the concentrations of Cu(II), Co(II), and Ni(II) all range from 0.2 mg/L to 1.4 mg/L. The concentration of Zn(II), Cu(II), Co(II), and Ni(II) in the solutions are shown in Table 1. Figure 2 is the spectra of single ion solution of Zn(II), Cu(II), Co(II), and Ni(II), which exhibits that the peaks of ions are adjacent to each other. This is because the competitive reaction between Zn(II) and other impurity metal ions aggravates the spectrum overlapping and masking. Meanwhile, the peak shape and movement tendency of ions are similar due to their resemble chemical properties. Besides, a negative peak occurs mainly because the absorbance value of Zn(II) complex in the measured solution is less than of the reagent blank at 320–380 nm. A large amount of noise appear at 280–300 nm and 300–400 nm mainly because of the high absorbance of the reference solution, making the spectrum nonlinear at these ranges. Therefore, the traditional calibration method based on the full spectrum can hardly achieve a high detection accuracy.

## 2.2. Multi-Derivative Spectral Reconstruction by SPSE

Due to the random noise and overlapping problems in the original spectra of complex mixtures, the derivative spectra method is widely used in spectral analysis of multicomponent calibration.

To decrease the background error and separate the overlapping absorption band, the singular perturbation spectrum estimator (SPSE) [19,20] is applied. The estimator is based on inverse Taylor series, which takes the advantage of scale separation to obtain the simplified original problems [35].

Assuming that the spectral signal  $u(v_1)$  is given at any wavelength  $v_1$  and that  $u(v_1)$  is differentiable  $k + 1$  times at any  $v_1$ . Define  $v_1 = v + \varepsilon$ , where  $\varepsilon$  is the system perturbation parameter and sufficiently small. For higher-order derivative estimators, because  $\lim_{\varepsilon \rightarrow 0} x_i = u^{(i-1)}(v)$  ( $i = 1, 2, \dots, n$ ), the Taylor series of spectral signal can be approximated as linear differential system and the description of SPSE is obtained as follows:

$$\begin{cases} \dot{x}_1(v) = x_2(v) \\ \dot{x}_2(v) = x_3(v) \\ \dot{x}_3(v) = -\frac{6}{\varepsilon^3}(x_1(v) - \hat{u}(v)) - \frac{6}{\varepsilon^2}x_2(v) - \frac{3}{\varepsilon}x_3(v) \\ y(v) = x_1(v), \end{cases} \quad (1)$$

where  $\hat{u}(v) = u(v_1)$  is the measured spectral signal;  $(\dot{x}_1, \dot{x}_2, \dot{x}_3)$  are the state items of the differentiator;  $x_1(v)$  is the denoising spectrum of the measured signal  $u(v)$ ; and  $x_2(v)$  and  $x_3(v)$  are the first-order and second-order derivative spectrum, respectively.

Since the SPSE only has the perturbation parameter  $\varepsilon$  to adjust, it can overcome the restriction of inconsistent parameter selection. Meanwhile, a large amount of additive noise is eliminated by the successive multiple integral parts. Thus, it is concluded that the denoising spectrum and multi-order derivative spectrum can be estimated, effectively suppressing random noise and redundant background signals in the spectrum.

### 2.3. eXtreme Gradient Boosting Based on Feature Importance Ranking

Extreme gradient boosting (XGBoost) [31] is a novel tree learning algorithm which achieves considerable result for sparse data processing. It takes classification and regression tree (CART) as the base learner. Figure 3 illustrates the basic structure of XGBoost, in which  $X$  is the spectral absorbance matrix in this model and  $y$  is the concentration of a certain metal ion. According to the additive training strategy of boosting, each tree is constructed based on learning from the residual  $\delta$  of the previous tree.  $\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i)$  is the prediction of the  $k$ -th iteration. At every iteration, XGBoost optimizes the model and decreases the prediction error. The final prediction output  $\hat{y}_i$  is generated by the weighted summation of trees as follows:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}, \quad (2)$$

where  $\mathcal{F}$  is the space of functions containing all regression trees;  $K$  denotes the number of trees. To learn function  $f_k$  of each tree, XGBoost establishes an objective function with regularization:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (3)$$

where  $\phi$  is all learnable parameters in XGBoost;  $l(y_i, \hat{y}_i)$  is the loss function representing the error between the predicted concentration  $\hat{y}_i$  and the actual concentration  $y_i$ , the smaller the  $l$  is, the better the performance of the algorithm;  $\Omega(f_k)$  is the regularization term to penalize the model complexity and prevent over-fitting. When XGBoost uses the square loss function to measure error, the second derivative Taylor expansion of the loss function can assist the model to optimize the objective quickly. The second derivative Taylor expansion of the loss function after  $k$ -th iteration is given as follows:

$$\mathcal{L}(\phi)^{(t)} \simeq \sum [l(y_i^{(t)}, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t), \quad (4)$$

where  $g_i$  and  $h_i$  are the first and second derivative of the loss function. It can be learned that the loss function only depends on the first and second derivatives of each data point. To predict the ion concentrations, the essential step in the XGBoost learning algorithm is to optimize the XGBoost algorithm parameters, booster parameters, and learning parameters.

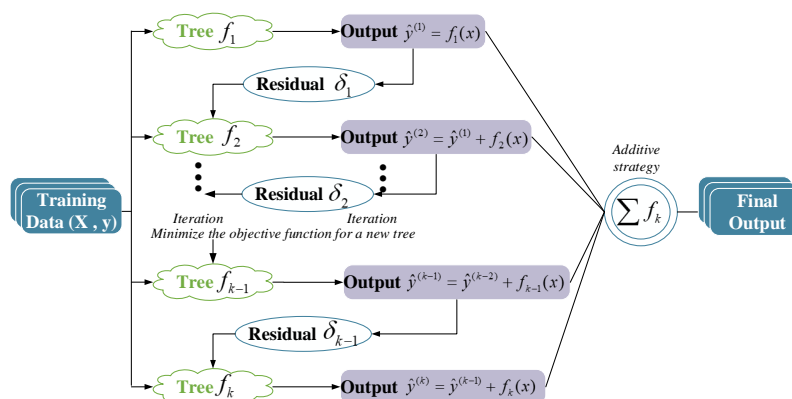


Figure 3. The structure of extreme gradient boosting.

Additive tree boosting model enables XGBoost to flexibly use variables in different areas of the output space. This model can perform effective feature selection and capture higher-order interactions. Thus, XGBoost in this paper is used not only for feature selection but also for prediction. After all boosting trees are built, XGBoost can calculate out the importance of each feature. XGBoost generates the ranking of all features based on variable importance (VI). The VI score measures the frequency of individual feature that is used to build trees. The more times a feature is selected for splitting trees, the more valuable it proves to be in the model. In this paper, the feature importance ranking of VI score is regarded as the basis of feature selection.

#### 2.4. The Proposed SPSE-XGBoost Approach

In view of the merits of SPSE and XGBoost, in this paper, they are integrated to establish a novel calibration model, shorted as SPSE-XGBoost. The focus of this approach is to explore the benefits of combining feature subsets of multi-order derivative spectrum and, meanwhile, introduce the ensemble XGBoost algorithm as key model in feature selection and prediction of metal ion concentrations such as zinc, copper, cobalt, and nickel. To assess the quality of SPSE-XGBoost model, the root mean square error (RMSE), the coefficient of determination ( $R^2$ ), the mean absolute percentage error (MAPE), and the maximum absolute percentage error (MaxAPE) are utilized as the main evaluation criteria in the proposed approach. Smaller RMSE, MAPE, and MaxAPE represent better model precision. Figure 4 illustrates the flow chart of the proposed SPSE-XGBoost model that is comprised of the following four steps.

- Step 1:** The samples are divided into training set and test set. The X matrix contains all variables of training set. Then, select different singularly perturbation parameter  $\varepsilon$  of SPSE to obtain the denoising spectrum and the first-order and second-order derivative spectra.
- Step 2:** Perform XGBoost modeling with cross validation for each spectrum. Since different derivative spectra have different predictive capability, the results are further compared and analyzed to select preferable derivative order for each ion.
- Step 3:** Calculate the VI score and rank to extract features. The sensitive feature in each spectrum are integrated to form new feature subsets.
- Step 4:** Combine the feature subsets and build SPSE-XGBoost model to predict ions concentration in the test set. Determine the optimum variables combination via  $RMSE_p$  and  $R^2$ .

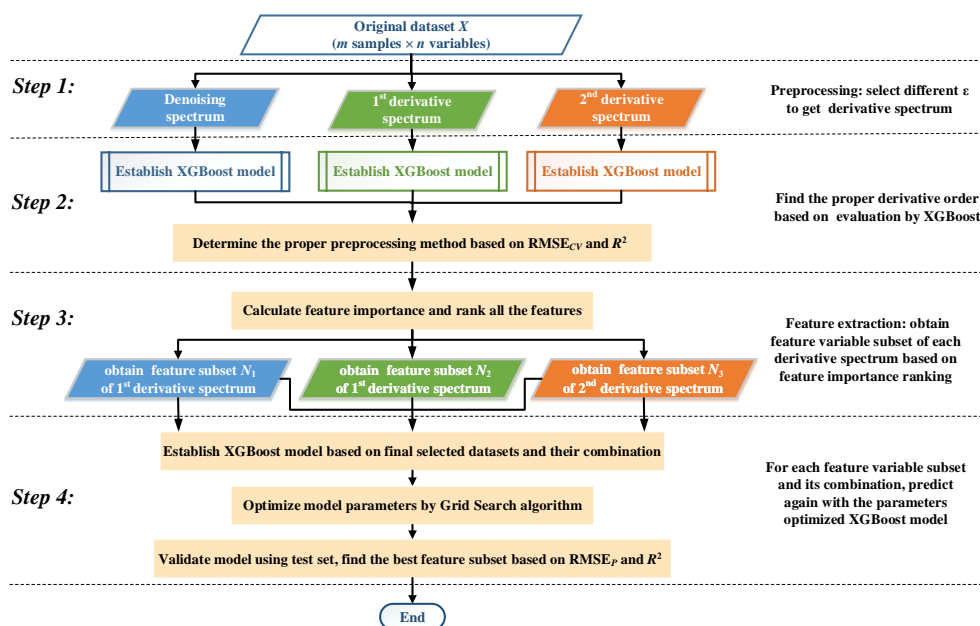


Figure 4. Flow chart of the proposed singular perturbation spectral estimator (SPSE)-XGBoost model.

In brief, the proposed method aims to find the best subset of features for multi-metal ions prediction and analyze the effects of variable selection and combination by an “Expand-Extract-Combine” strategy. “Expand” refers to the derivative preprocessing procedure that expands the spectral space of original data. “Extract” means that individual variable is ranked and selected by the VI score and “Combine” defines that multi-derivative feature subset combination is considered to promote prediction performance.

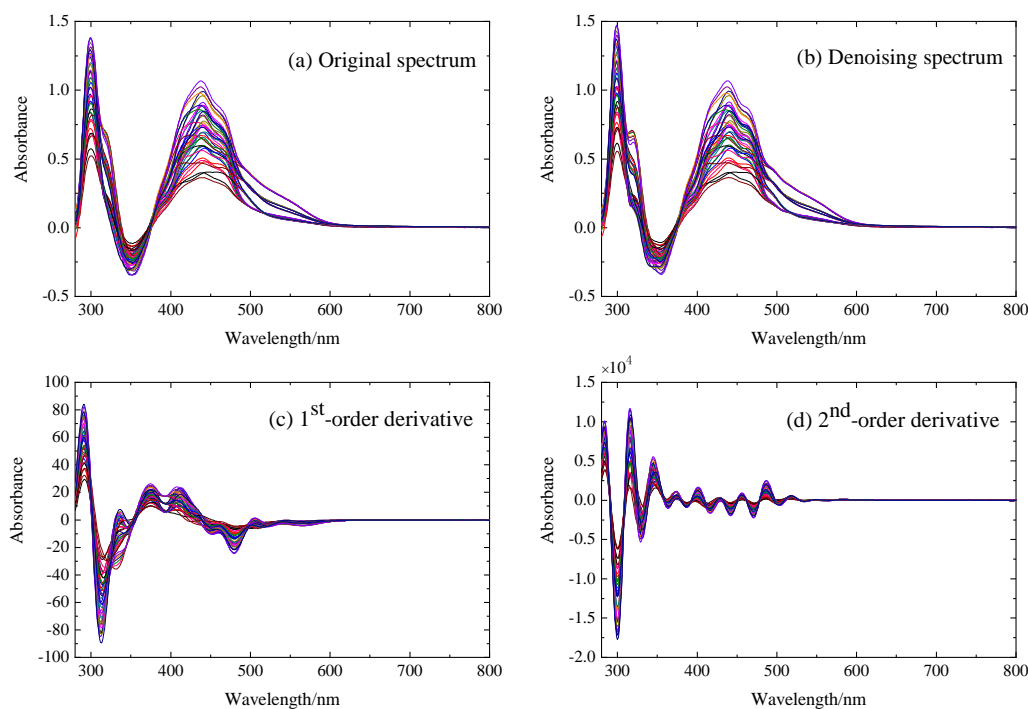
### 3. Results and Discussion

#### 3.1. Multi-Order Derivative Reconstruction Pretreatment

The samples were firstly divided into a training set (39 samples) and a test set (10 samples) using the Kennard-Stone algorithm [36]. To highlight the characteristic information of each ion, different singularly perturbation parameter  $\varepsilon$  is selected in the SPSE to get the denoising spectrum and the first-order and second-order derivative spectra. To obtain the best preprocessing results, the  $\varepsilon$  in our work were set as 0.007 for denoising spectrum, 0.013 for first-order spectrum, and 0.016 for second-order spectrum, respectively. It can be seen from Figure 5 that subtle changes in the original spectrum is obviously reinforced after pretreatment, highlighting the ionic difference. The denoising spectrum does not make significant difference with the original spectrum in the shape. The absorbance and resolution are greatly enhanced in the derivative spectra, which reasonably expands the data space and thus provides abundant features for variables selection.

To ensure the reliability of preprocessing, the prediction results of XGBoost model with 10-fold cross-validation using full-spectrum (280–800 nm) is preliminarily analyzed as shown in Table 1. The  $RMSE_{CV}$  and  $R^2$  are used to evaluate the predictive ability of the model. The MAPE and MaxAPE are also calculated. Table 2 reveals that different derivative spectrum models provide different prediction results. For zinc, the accuracy of models established by the first-order and second-order spectrum are improved compared with the original spectrum. Although the noise signal is suppressed in the denoising spectrum, some important characteristics of zinc may be inevitably weakened, so the predicting result of the denoising spectrum is the worst. For copper, most of the original signals are masked by zinc, so the denoising spectrum and the first-order derivative spectrum have the strongest prediction ability, and their MAPE are greatly reduced. Whereas for cobalt, only the model established by the second-order derivative spectrum has a impressive accuracy. By calculating the

variation in absorbance change rate, spectral peaks become exceedingly sharp and the overlapping spectral bands are in a way separated. Therefore, the characteristic information of cobalt stands out conspicuously. For nickel, the model with the denoising spectrum and the first-order derivative is preferable. The second-order derivative model is the worst. This is because the signal of nickel is the weakest in zinc sulfate solution. The second-order derivative spectrum inevitably amplifies the instability of signal and reduces the prediction precision.



**Figure 5.** The original, denoising, first-order, and second-order derivative spectra of the training set.

**Table 2.** Error evaluations of full-spectrum among different preprocessing methods for metal ions concentration prediction.

	Preprocessing	MAPE(%)	MaxAPE(%)	RMSE <sub>CV</sub> (mg/L)	R <sup>2</sup>
<b>Zinc</b>	Raw	13.777	21.534	4.785	0.792
	Denoising	16.718	24.621	4.947	0.749
	<b>1st derivative</b>	<b>8.744</b>	<b>13.220</b>	<b>4.242</b>	<b>0.892</b>
	<b>2nd derivative</b>	<b>8.235</b>	<b>11.395</b>	<b>3.941</b>	<b>0.921</b>
<b>Copper</b>	Raw	13.762	18.781	0.145	0.746
	<b>Denoising</b>	<b>8.023</b>	<b>12.789</b>	<b>0.115</b>	<b>0.924</b>
	<b>1st derivative</b>	<b>7.236</b>	<b>11.604</b>	<b>0.105</b>	<b>0.938</b>
	2nd derivative	11.819	14.355	0.139	0.824
<b>Cobalt</b>	Raw	11.213	23.958	0.131	0.773
	Denoising	11.051	21.545	0.146	0.737
	1st derivative	10.792	17.897	0.134	0.756
	<b>2nd derivative</b>	<b>6.254</b>	<b>12.032</b>	<b>0.099</b>	<b>0.901</b>
<b>Nickel</b>	Raw	12.463	23.057	0.143	0.779
	<b>Denoising</b>	<b>8.529</b>	<b>14.573</b>	<b>0.114</b>	<b>0.907</b>
	<b>1st derivative</b>	<b>9.322</b>	<b>15.496</b>	<b>0.118</b>	<b>0.894</b>
	2nd derivative	17.758	25.619	0.150	0.738

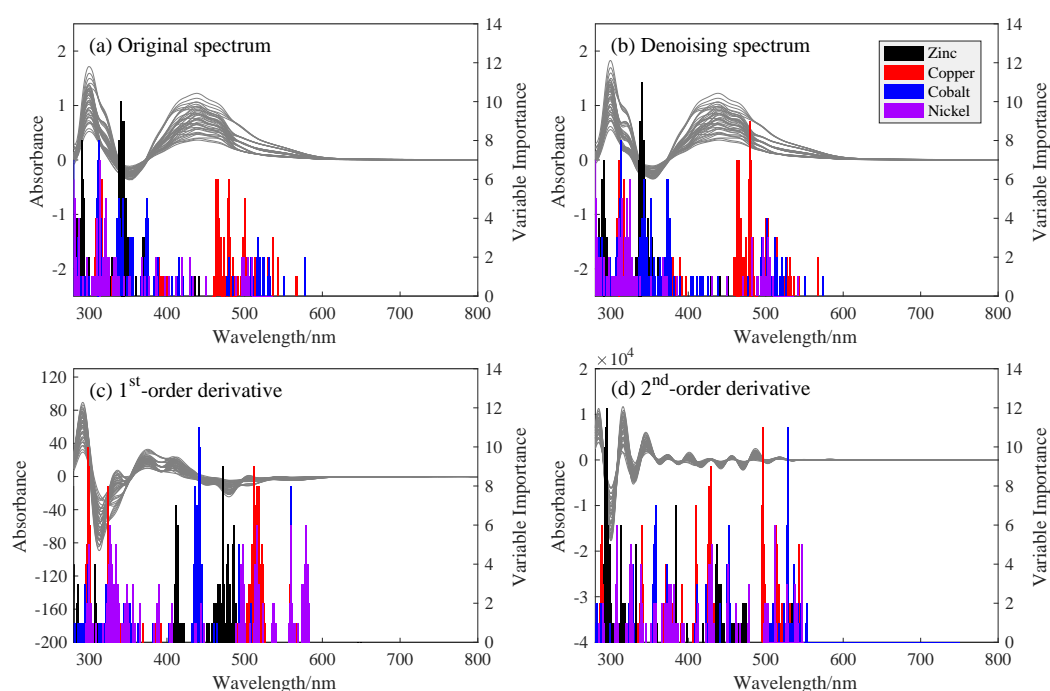
Overall, we have three interesting findings: (i) Different preprocessing method has different effects on the prediction ability of different ions in the same solution. For example, the denoising spectrum has a positive effect on copper, cobalt, and nickel but negative on zinc; (ii) derivative pretreatment



provides more abundant and effective data for spectral prediction, which can improve the accuracy of model; (iii) a single derivative full-spectrum cannot meet the industrial requirements that the average error of the measurement of should not exceed 5%, and the maximum error should not exceed 10%. Therefore, the derivative variables extraction and combination will be considered to maximize the effective information in the following subsections.

### 3.2. Variable Selection and Feature Importance Analysis

After preprocessing, we obtain the high-dimension spectral data. Before modeling the spectral data with small sample yet high dimensions, it is extremely necessary to lessen dimension with appropriate variable selection method. For VI score ranking, XGBoost extracts feature variables by calculating the importance ranking of all variables. The times of a feature selected as a splitting tree node are regarded as the VI scores to measure the feature importance.



**Figure 6.** Characteristic variable selection and feature importance (VI) score results of zinc, copper, cobalt, and nickel ions in the original (a), denoising (b), first-order (c), and second-order derivative (d) spectra. The gray curves in each subplot are the spectra of the training set (Figure 6).

**Table 3.** The number of selected characteristic variables by XGBoost of zinc, copper, cobalt, and nickel ions under different preprocessing methods.

Preprocessing	Number of Characteristic Variables			
	Zinc	Copper	Cobalt	Nickel
Raw	57	77	78	68
Denoising	62	83	82	71
1st derivative	63	86	73	72
2nd derivative	59	75	81	87

The variables selection for multi-metal ions and their VI scores of different spectra are shown in Figure 6. From Figure 6a,b, we can find that the original and denoising spectra have approximately the same range of wavelength bands, indicating that the denoising pretreatment does not change feature position in the selected variables. The selected variables of zinc are mainly located in the range of 280–380 nm. Copper has a strong concentrated feature band located at 460–510 nm. This is consistent

with the characteristic peak at 490 nm in the single ion spectrum of copper, which can be clearly seen at Figure 2. In the original and denoising spectra, the cobalt ions are distributed in a dispersed manner, ranging from 280 nm to 580 nm. Nickel ions are mainly distributed around 280–400 nm and 500 nm, which is related to the three characteristic peaks of nickel ions at 310 nm, 410 nm, and 500 nm in the single ion spectrum. In general, the VI scores of the denoising spectrum is higher than that of the original spectrum, especially for copper and nickel, which reflects that the elimination of random noise is conducive to feature selection and information mining.

When taking the first-order and second-order derivations, the selected variables are distributed to a wide range and move backwards. The selected feature bands of the different ions become distinguished and more concentrated, suggesting that the derivations can effectively expand the spectral data space for feature selection and separate overlapping spectral peaks of ions. For example, in the first-order derivative spectrum, the characteristic regions of zinc move to 400–420 nm and 450–500 nm, while the variables of copper, cobalt, and nickel are relocated at 500–520 nm, 430–450 nm, and 550–580 nm, respectively. These characteristic bands are all independent of each other and skillfully avoid the high noise bands at 280–380 nm. In particular, the variable selection of the first-order derivative spectrum is more concentrated, and the second derivative has a scattered distribution before 600 nm due to the large number of characteristic peaks and narrow peak shape.

The number of the selected characteristic variables of the corresponding metal ions under different preprocessing methods is listed in Table 3. Compared with the full-spectrum prediction, the calculated variable for one ion in the XGBoost model is reduced to below 87. Thus, using the selected feature subset can not only highlight the characteristic information for a specific metal ion but also remarkably improve the calculation speed of the model.

### 3.3. XGBoost Model with Variable Combination

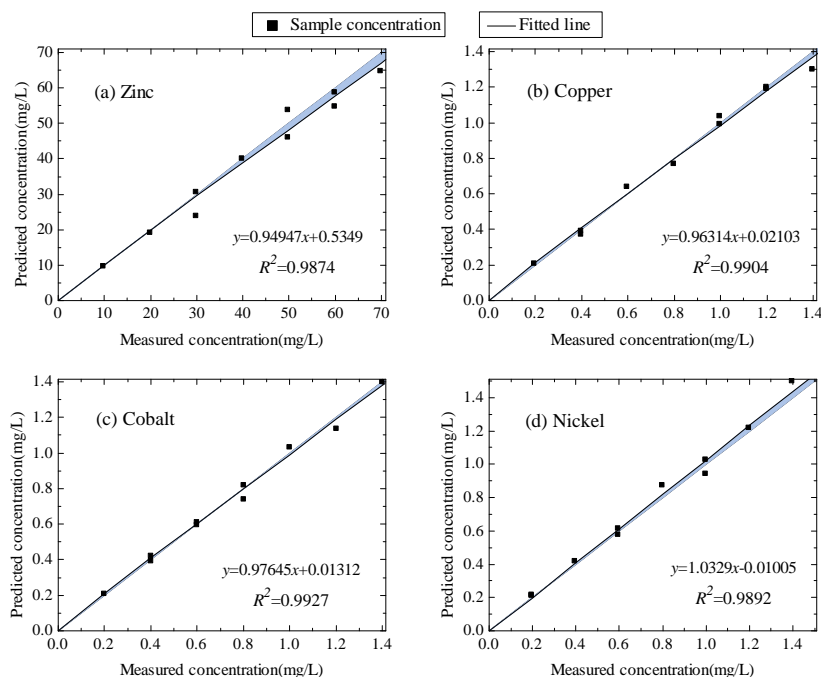
To meliorate the prediction ability of the model, this approach also adopts the variable combination strategy. Variable combination takes advantage of the diversity of feature variables. It is also worth mentioning that to our best knowledge, the effect of variable combination is not considered in any existing method to predict the metal ion concentrations. Hence, the selected subset of derivative variables and their combinations are applied to retrain the XGBoost model, respectively. The prediction results are compared and evaluated by the MAPE, MaxAPE, RMSE<sub>P</sub>, R<sup>2</sup> in Table 4. The optimal model usually exhibits the largest R<sup>2</sup> and the lowest RMSE<sub>P</sub> value. To meet the needs of industrial on-line detection, MAPE and MaxAPE are also required to be under 10%.

**Table 4.** Model prediction results of extracted variables subsets and their combinations under different preprocessing methods for metal ion concentration prediction.

	Feature Subset	MAPE(%)	MaxAPE(%)	RMSE <sub>P</sub> (mg/L)	R <sup>2</sup>
<b>Zinc</b>	Denoising	\	\	\	\
	1st derivative	6.942	9.433	3.569	0.948
	2nd derivative	6.271	9.185	3.412	0.956
	<b>Combination</b>	<b>4.098</b>	<b>7.986</b>	<b>3.107</b>	<b>0.987</b>
<b>Copper</b>	Denoising	4.241	9.452	0.056	0.977
	1st derivative	3.924	8.404	0.051	0.984
	2nd derivative	\	\	\	\
	<b>Combination</b>	<b>3.515</b>	<b>6.939</b>	<b>0.043</b>	<b>0.990</b>
<b>Cobalt</b>	Denoising	\	\	\	\
	1st derivative	\	\	\	\
	<b>2nd derivative</b>	<b>3.083</b>	<b>7.414</b>	<b>0.041</b>	<b>0.993</b>
	Combination	\	\	\	\
<b>Nickel</b>	Denoising	6.879	9.922	0.078	0.946
	1st derivative	4.612	9.067	0.072	0.953
	2nd derivative	\	\	\	\
	<b>Combination</b>	<b>4.331</b>	<b>8.323</b>	<b>0.054</b>	<b>0.989</b>

According to model results, it can be observed that the modeling outcome of selected variable subsets is superior to that based on the whole spectrum variables (Table 2). More interestingly, through the combination of subsets of characteristic region in the denoising and multi-order derivation spectra, the prediction ability of ions is further optimized. It is apparent from Table 4 that the  $RMSEP$  and MAPE are effectively reduced. The MAPE of zinc, copper, cobalt, and nickel are 4.098%, 3.515%, 3.083%, and 4.331%, respectively. The MaxAPE of the four ions is less than 8.323%. Hence, it concludes that feature variable subset combination can make better use of the richness and diversity of the multi-order derivative spectra and contain characteristic variables to the greatest extent. Moreover, it further proves that even when some of the less important variables are combined, they can achieve good predictive performance. Especially for nickel with the strongest instability and the weakest absorbance, the MaxAPE and MAPE of the model prediction are greatly reduced, which meet the industrial requirements of accuracy.

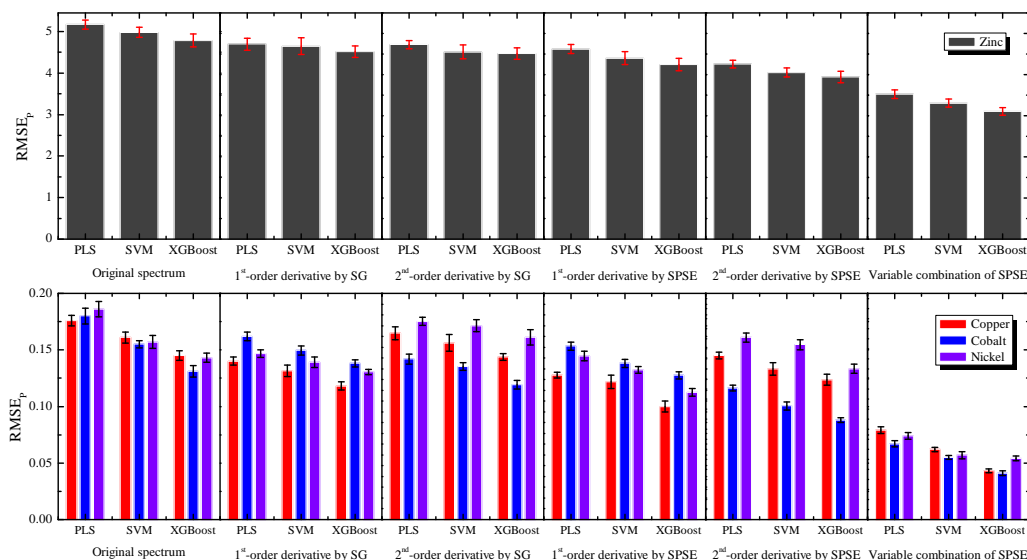
The scatter points in Figure 7 illustrate the comparison of actual concentration and predicted concentration in the test set. The black lines are the fitted lines of actual and predicted concentration scatter points. The blue shaded areas stand for the gap between the 1:1 lines and the fitted lines. When the fitted line is closer to the 1:1 line, the blue shaded area is smaller, indicating a better correlation between the prediction result and actual concentration. As exhibited, the fitted lines of each ion closely approach to the 1:1 lines and the  $R^2$  values of each ion are higher than 0.987, suggesting that the proposed method has favorable accuracy and promising effect in detecting the concentrations of zinc, copper, cobalt, and nickel ions in zinc sulfate solution.



**Figure 7.** The fitting curves of zinc (a), copper (b), cobalt (c), and nickel (d) ions in the test set (10 samples).

To further verify the performance of the proposed model, different derivatives pretreatment methods, such as Savitzky-Golay (SG) derivative algorithm and traditional modeling methods (CARS-PLS, UVE-LS-SVM), are also carried out and compared in Figure 8. Note that CARS-PLS and UVE-LS-SVM are shortened as PLS and SVM in the figure, respectively. All the methods are conducted 10 times to obtain the statistical results. Only the variable combination method of SPSE uses subset combination of characteristic variables, while the other methods are all based on the full spectral variables. From Figure 8, the performances of the different algorithms are distinct. For the single preprocessing methods, the original spectral data almost have the highest  $RMSEP$  due to the noise and masking problems in the raw spectrum. The exceptions are the first-order derivative for cobalt

and the second-order derivative for nickel, which have been interpreted in Section 3.1. The derivative spectra using SPSE always achieve lower  $RMSE_p$  than the corresponding SG, indicating that SPSE is superior to SG. For the modeling methods, the XGBoost outperforms the other two traditional calibration methods as shown in the figure. It is worth mentioning that the combination of variables has a significant effect on model promotion and yields the lowest  $RMSE_p$ , signifying the best prediction ability of SPSE-XGBoost in general.



**Figure 8.** Comparison results among different preprocessing and modeling methods for zinc, copper, cobalt, and nickel ions.

All in all, it is easily concluded that the SPSE and XGBoost algorithms have better capabilities in expanding the spectral information, extracting the effective variables, and predicting ion concentrations in the UV-vis spectrum analysis. When all of the feature subsets from the derivative spectra are combined, the developed SPSE-XGBoost method can further enhance the measurement accuracy. Therefore, our proposed “Expand-Extract-Combine” strategy of SPSE-XGBoost in this article has a great potential for the real-time and on-line detection of multi-metal ion concentrations in hydrometallurgy wastewater.

#### 4. Conclusions

In this work, we developed the SPSE-XGBoost approach to simultaneously measure the multi-metal ion concentrations by UV-vis spectroscopy in the complex zinc sulfate solutions. At first, the spectral data was expanded by the denoising and multi-order derivative preprocessing in SPSE. Then, the feature variables were extracted by accounting the VI score ranking using the ensemble XGBoost algorithm. Finally, the feature subsets from the derivative spectra were combined to further promote the accuracy in determining zinc, copper, cobalt, and nickel ion concentrations. The adequate analyses indicate that the “Expand-Extract-Combine” strategy in the SPSE-XGBoost approach has the properties of suppressing the redundant noises, extending the spectral feature data to a broad space, extracting the spectral feature band for a specific metal ion, improving the computing speed, and obtaining the high-precise results, and so on. The comparisons with the conventional SG preprocessing, CARS-PLS, and UVE-LS-SVM methods illustrate the superior performances of our proposed method. With such analyses, our developed approach was proven to be suitable for real-time and on-line detection of multi-metal ion concentrations in hydrometallurgy wastewater.

**Author Contributions:** F.C. performed the experiments, analyzed the data and wrote the manuscript; C.Y. helped with the proofreading of the manuscript; L.L. helped to interpret the results and revise the paper; C.Z. conceived and designed the experiments; H.Z. interpreted the data and revised the paper; Y.L. performed the analysis with constructive discussions. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** This work is supported by the State Key Program of National Natural Science Foundation of China (Grant No. 61533021), the National Natural Science Foundation of China (Grant No. 61773403), the project of State Key Laboratory of High Performance Complex Manufacturing in Central South University (Grant No. ZZYJKT2019-14), and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 2019zzts561).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chen, S.H.; Li, Y.X.; Li, P.H.; Xiao, X.Y.; Jiang, M.; Li, S.S.; Zhou, W.Y.; Yang, M.; Huang, X.J.; Liu, W.Q. Electrochemical spectral methods for trace detection of heavy metals: A review. *TrAC Trends. Anal. Chem.* **2018**, *106*, 139–150. [[CrossRef](#)]
2. Lu, Y.; Liang, X.; Niyungeko, C.; Zhou, J.; Xu, J.; Tian, G. A review of the identification and detection of heavy metal ions in the environment by voltammetry. *Talanta* **2018**, *178*, 324–338. [[CrossRef](#)]
3. Burakov, A.E.; Galunin, E.V.; Burakova, I.V.; Kucherova, A.E.; Agarwal, S.; Tkachev, A.G.; Gupta, V.K. Adsorption of heavy metals on conventional and nanostructured materials for wastewater treatment purposes: A review. *Ecotoxicol. Environ. Saf.* **2018**, *148*, 702–712. [[CrossRef](#)]
4. Hou, D.B.; Zhang, J.; Chen, L.; Huang, P.J.; Zhang, G.X. Water quality analysis by UV-Vis spectroscopy: A review of methodology and application. *Spectrosc. Spectral Anal.* **2013**, *33*, 1839–1844.
5. Rocha, F.S.; Gomes, A.J.; Lunardi, C.N.; Kaliaguine, S.; Patience, G.S. Experimental methods in chemical engineering: Ultraviolet visible spectroscopy—UV-Vis. *Can. J. Chem. Eng.* **2018**, *96*, 2512–2517. [[CrossRef](#)]
6. Viljanen, J.; Kalmankoski, K.; Contreras, V.; Sarin, J.K.; Sorvajärvi, T.; Kinnunen, H.; Enestam, S.; Toivonen, J. Sequential collinear photofragmentation and atomic absorption spectroscopy for online laser monitoring of triatomic metal species. *Sensors* **2020**, *20*, 533. [[CrossRef](#)] [[PubMed](#)]
7. Kirchler, C.G.; Henn, R.; Modl, J.; Münzker, F.; Baumgartner, T.H.; Meischl, F.; Kehle, A.; Bonn, G.K.; Huck, C.W. Direct determination of Ni<sup>2+</sup>-capacity of IMAC materials using near-infrared spectroscopy. *Molecules* **2018**, *23*, 3072. [[CrossRef](#)] [[PubMed](#)]
8. Guselnikova, O.; Svorcik, V.; Lyutakov, O.; Chehimi, M.M.; Postnikov, P.S. Preparation of selective and reproducible SERS sensors of Hg<sup>2+</sup> ions via a sunlight-induced thiol–Yne reaction on gold gratings. *Sensors* **2019**, *19*, 2110. [[CrossRef](#)]
9. Ma, S.; Tang, Y.; Ma, Y.; Chu, Y.; Chen, F.; Hu, Z.; Zhu, Z.; Guo, L.; Zeng, X.; Lu, Y. Determination of trace heavy metal elements in aqueous solution using surface-enhanced laser-induced breakdown spectroscopy. *Opt. Express* **2019**, *27*, 15091–15099. [[CrossRef](#)]
10. Stiedl, J.; Green, S.; Chassé, T.; Rebner, K. Characterization of oxide layers on technical copper material using ultraviolet visible (UV-Vis) spectroscopy as a rapid on-line analysis tool. *Appl. Spectrosc.* **2019**, *73*, 59–66.
11. Wang, K.; Yu, J.; Hou, D.; Yin, H.; Yu, Q.; Huang, P.; Zhang, G. Optical detection of contamination event in water distribution system using online Bayesian method with UV-Vis spectrometry. *Chem. Intel. Lab. Syst.* **2019**, *191*, 168–174. [[CrossRef](#)]
12. Yang, C.S.; Jin, F.; Trivedi, S.; Brown, E.; Hömmerich, U.; Nemes, L.; Samuels, A.C. In situ chemical analysis of geology samples by a rapid simultaneous ultraviolet/visible/near-infrared (UVN)+ longwave-infrared laser induced breakdown spectroscopy detection system at standoff distance. *Opt. Express* **2019**, *27*, 19596–19614. [[CrossRef](#)] [[PubMed](#)]
13. Zhou, F.; Li, Y.; Zhu, H.; Zhou, C.; Li, C. Signal enhancement algorithm for on-line detection of multi-metal ions based on ultraviolet-visible spectroscopy. *IEEE Access* **2020**, *8*, 16000–16008. [[CrossRef](#)]
14. Chen, J.; Yang, C.; Zhou, C.; Li, Y.; Zhu, H.; Gui, W. Multivariate regression model for industrial process measurement based on double locally weighted partial least squares. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 3962–3971. [[CrossRef](#)]
15. Ojeda, C.B.; Rojas, F.S. Recent applications in derivative ultraviolet/visible absorption spectrophotometry: 2009–2011: A review. *Microchem. J.* **2013**, *106*, 1–16. [[CrossRef](#)]
16. Yang, J.; Cheng, Y.; Du, L.; Gong, W.; Shi, S.; Sun, J.; Chen, B. Selection of the optimal bands of first-derivative fluorescence characteristics for leaf nitrogen concentration estimation. *App. Opt.* **2019**, *58*, 5720–5727. [[CrossRef](#)]

17. Simion, I.M.; Sârbu, C. The impact of the order of derivative spectra on the performance of pattern recognition methods. Classification of medicinal plants according to the phylum. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *219*, 91–95. [[CrossRef](#)]
18. Li, Z.; Ma, Z. A new approach for filtering and derivative estimation of noisy signals. *Circuits Syst. Signal Process.* **2014**, *33*, 589–598. [[CrossRef](#)]
19. Li, Z.; Wang, Q.; Lv, J.; Ma, Z.; Yang, L. Improved quantitative analysis of spectra using a new method of obtaining derivative spectra based on a singular perturbation technique. *App. Spectrosc.* **2015**, *69*, 721–732. [[CrossRef](#)]
20. Li, Z.; Li, T.; Lv, H.; Wang, Q.; Si, G.; He, Z. Quantitative analysis of biofluids based on hybrid spectra space. *Chem. Intel. Lab. Sys.* **2017**, *165*, 22–28. [[CrossRef](#)]
21. Andries, J.P.; Vander Heyden, Y.; Buydens, L.M. Improved variable reduction in partial least squares modelling by Global-Minimum Error Uninformative-Variable Elimination. *Anal. Chim. Acta* **2017**, *982*, 37–47. [[CrossRef](#)] [[PubMed](#)]
22. Li, S.; Zhang, X.; Shan, Y.; Su, D.; Ma, Q.; Wen, R.; Li, J. Qualitative and quantitative detection of honey adulterated with high-fructose corn syrup and maltose syrup by using near-infrared spectroscopy. *Food Chem.* **2017**, *218*, 231–236. [[CrossRef](#)]
23. Hong, Y.; Chen, S.; Liu, Y.; Zhang, Y.; Yu, L.; Chen, Y.; Liu, Y.; Cheng, H.; Liu, Y. Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy. *Catena* **2019**, *174*, 104–116. [[CrossRef](#)]
24. Yun, Y.H.; Wang, W.T.; Deng, B.C.; Lai, G.B.; Liu, X.b.; Ren, D.B.; Liang, Y.Z.; Fan, W.; Xu, Q.S. Using variable combination population analysis for variable selection in multivariate calibration. *Anal. Chim. Acta* **2015**, *862*, 14–23. [[CrossRef](#)] [[PubMed](#)]
25. Yun, Y.H.; Li, H.D.; Deng, B.C.; Cao, D.S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends Anal. Chem.* **2019**, *113*, 102–115. [[CrossRef](#)]
26. Erler, A.; Riebe, D.; Beitz, T.; Löhmansröben, H.G.; Gebbers, R. Soil nutrient detection for precision agriculture using handheld laser-induced breakdown spectroscopy (LIBS) and multivariate regression methods (PLSR, Lasso and GPR). *Sensors* **2020**, *20*, 418. [[CrossRef](#)]
27. Sha, W.; Li, J.; Xiao, W.; Ling, P.; Lu, C. Quantitative analysis of elements in fertilizer using laser-induced breakdown spectroscopy coupled with support vector regression model. *Sensors* **2019**, *19*, 3277. [[CrossRef](#)]
28. Zhou, Z.H. *Ensemble Methods: Found. Algorithm*; Chapman and Hall/CRC: London, UK, 2012.
29. Liu, H.; Zhang, L. Advancing ensemble learning performance through data transformation and classifiers fusion in granular computing context. *Expert Syst. Appl.* **2019**, *131*, 20–29. [[CrossRef](#)]
30. Bian, X.; Li, S.; Shao, X.; Liu, P. Variable space boosting partial least squares for multivariate calibration of near-infrared spectroscopy. *Chemom. Intel. Lab. Syst.* **2016**, *158*, 174–179. [[CrossRef](#)]
31. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 22–27 August 2016; pp. 785–794.
32. Nawar, S.; Mouazen, A.M. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors* **2017**, *17*, 2428. [[CrossRef](#)]
33. Yang, X.; Fang, T.; Li, Y.; Guo, L.; Li, F.; Huang, F.; Li, L. Pre-diabetes diagnosis based on ATR-FTIR spectroscopy combined with CART and XGBoots. *Optics* **2019**, *180*, 189–198. [[CrossRef](#)]
34. Wei, L.; Yuan, Z.; Yu, M.; Huang, C.; Cao, L. Estimation of arsenic content in soil based on laboratory and field reflectance spectroscopy. *Sensors* **2019**, *19*, 3904. [[CrossRef](#)] [[PubMed](#)]
35. Khalil, H.K.; Grizzle, J.W. *Nonlinear System*; Engineering Michigan State University, Prentice Hall: Upper Saddle River, NJ, USA, 2002; Volume 3.
36. Macho, S.; Rius, A.; Callao, M.; Larrechi, M. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model. *Anal. Chim. Acta* **2001**, *445*, 213–220. [[CrossRef](#)]

