Research Article

Machine Learning Framework for Intelligent Detection of Wastewater Pollution by IoT-Based Spectral Technology

Jianhong Li,1 Ken Cai,2 Huazhou Chen3, Lili Xu,4 Qinyong Lin2, and Feng Xu3

1School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China
2College of Automation, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
3College of Science, Guilin University of Technology, Guilin 541004, China
4College of Marine Sciences, Beibu Gulf University, Qinzhou 535011, China

Correspondence should be addressed to Huazhou Chen; hzhengut@foxmail.com

Received 24 June 2021; Accepted 3 August 2021; Published 11 February 2022

Industrial wastewater contains excessive micro insoluble solids (MIS) that probably cause environmental pollutions. Near-infrared (NIR) spectroscopy is an advanced technology for rapid detection of the complex targets in wastewater. An Internet of Things (IoT) platform would support intelligent application of the NIR technologies. The studies of intelligent chemometric methods mainly contribute to improve the NIR calibration model based on the IoT platform. With the development of artificial intelligence, the backward interval and synergy interval techniques were proposed in combination use with the least square support vector machine (LSSVM) method, for adaptive selection of the informative spectral wavelength variables. The radial basis function (RBF) kernel is applied for nonlinear mapping. The regulation parameter and the kernel width are fused together for smart optimization. In the design for waveband auto fittings, the total of digital wavelengths in the full scanning range was split into 43 equivalent subintervals, and then, the back interval LSSVM (biLSSVM) and the synergy interval LSSVM (siLSSVM) models were both established for the improvement of prediction results based on the adaptive selection of quasidiscrete variable combination. In comparison with some common linear and nonlinear models, the best training model was acquired with the siLSSVM method while the best testing model was obtained with biLSSVM. The intelligent optimization of model parameters indicated that the proposed biLSSVM and siLSSVM deep learning methodologies are feasible to improve the model prediction results in rapid determination of the wastewater MIS content by the IoT-based NIR technology. The machine learning framework is prospectively applied to the fast assessment of the environmental risk of industrial pollutions and water safety.

1. Introduction

Internet of Things (IoT) has gradually permeated all kinds of technological landscape, such as monitoring, agriculture, irrigation management, healthcare, and security [1]. An IoT framework with cloud centric storage and processing can analyze the data received from various sensors along with decision support network nodes [2]. The IoT platform would support intelligent applicable technologies, providing solutions that are helpful and fruitful in smart dimensional irrigations of the agricultural landscape [3, 4]. Intelligent spectral detection based on the IoT platform supports for sustainable eco-environment in combination with machine learning methodologies.

Near-infrared (NIR) spectroscopy has been developing as a sanctioned technology for rapid detection of complex analytes [5]. This technology uses the lights in the near-infrared frequency to illuminate the target sample. The components can be identified and quantitatively determined by studies of the intelligent chemoinformatic methods serving for model training based on the IoT cloud situations [6]. The NIR technology has advantages of reagent-free operation, nondestructive process, multicomponent simultaneous determination, and online implementation [7, 8]. It is widely applied in the fields of precise agriculture, digital food safety, smart environment, and intelligent medicine [9–11]. The model prediction effect is prospectively improved in...
In algorithm, the PLS regression method mainly establishes linear models for NIR analysis. But the linear regression models cannot give the prospective prediction model to extract comprehensive information when confronting the complex spectral data [22]. Modern machine learning methods have been used in NIR field to improve the prediction ability. Diago et al. [23] proved that the automated plant-based methods were feasible for fast assessment of the water status of a grapevine. Chen et al. [24] discussed some kernel network methods for NIR prediction of the industrial wastewater. Ahn and Park [25] promote the adaptive atmospheric correction procedure for VIS-NIR analysis of the turbid water. Chen et al. [26] applied convolutional neural network model for fast prediction of wastewater COD, with optimizing multiple parameters. Given the success of these previous works, we hypothesize that similar methodologies of NIR calibration would be able to predict the wastewater MIS for rapid assessment of environmental risk of water pollution. If the NIR spectral data information could be extracted by variable selection techniques such as backward interval selection and synergy interval combination, the nonlinear calibration models can be improved further with kernel transformations.

Support vector machine (SVM) is a relatively mature nonlinear machine learning algorithm for qualitative and quantitative analyses [27]. Previous studies showed that the least square SVM (LSSVM) model outperformed the common PLS model because the relationships between the spectral data and the target contents are practically nonlinear, accompanied with unknown components as well as complex noise interference [28]. When dealing with the NIR data of samples with heterogeneous properties, the LSSVM models can be optimized with in-depth tuning of the kernel functions. The radial basis function (RBF) is regarded as an appreciate kernel that is suitable for extract spectral information in the NIR field, because the nonlinear nature of data has been mapped into a linear space by the kernel function [29]. Therefore, to get robust and reliable results in NIR calibration and prediction, the LSSVM regulation and the selection of RBF kernel width is much important for successful prediction, especially in combined optimization of the variable selection techniques.

This study was conducted with industrial wastewater samples collected in south China, to establish intelligent analytical models for estimating the MIS content by using the IoT-based NIR technology. Machine learning framework is constructed based on the LSSVM algorithm, for adaptive detection of the complex wastewater samples. The NIR spectral information was extracted using the backward interval selection mode and using the synergy interval combination manner. The RBF-kernel LSSVM algorithm was used to establish the calibration models to achieve fast prediction. The LSSVM parameters and the kernel parameters are automatically optimized in combination of interval selection of NIR informative wavebands. The goal is to observe the optimal machine learning model for NIR fast estimating wastewater pollutions. The proposed methodological framework is expected to build up smart learning models for precise assessment of the prospective pollution level of wastewater, thus providing governance guidelines for environment sustainable with synchronous development in the development of the IoT cloud platform.

2. Methodologies and the Machine Learning Framework

2.1. Machine Learning Framework by Parametric LSSVM. A machine learning framework is built up based on the parametric LSSVM modeling, for the rapid estimation of
wastewater MIS using the IoT-based NIR technology. The model prediction accuracy for in situ measurement is influenced by the tuning of chemometric methods.

The theory of LSSVM is to transform the original nonlinear spectral data into a high dimensional space, in which the relationship between the transformed NIR data and the target analyte would become absolutely or approximately linear [27]. The distribution of samples in the high dimensional space is determinant by the mapping function $\varphi(\bullet)$. In the high dimensional data space, a decision function ($Q$) was constructed for minimizing the regulation of prediction errors, that is,

$$Q = \min \left( \frac{1}{2} \|w\|^2 + \gamma \|\epsilon\|^2 \right)$$

subject to $y = w^T \varphi(x) + b + \epsilon$.

where $\gamma$ is the regularization parameter. This is a convex optimization problem that can be solved by Lagrangian multiplier. Then, the prediction of the target analyte is determined by the formula,

$$y = \alpha_i K(x, x_i),$$

where $\alpha_i$ is the Lagrange multiplier, $y$ is the targeted analyte, $x$ represents the NIR spectrum, and $K(x, x_i)$ refers to the kernel function defined as $K(x, x_i) = \varphi(x) \varphi(x_i)$.

Machine learning by the parametric LSSVM has global optimality for nonstationary data. The parametric regularization may limit the possibility of overfitting. The radial basis function (RBF) is taken as the most effective kernel for quantitative regression, and its training process is much simple and fast [29], that is,

$$\text{RBF}K(x, x_i) = \exp \left( \frac{(x - x_i)^2}{2\sigma^2} \right),$$

where $\sigma^2$ represents the kernel width.

The NIR rapid estimation of wastewater target analytes (i.e., the MIS) needs to launch the large-scale optimization of the LSSVM model with RBF kernel (denoted as RBF-LSSVM for short). To find the suitable calibration-prediction strategy, the RBF-LSSVM model was trained by screening the regularization parameter ($\gamma$) and the RBF kernel width ($\sigma^2$) in the grid search mode. Grid search of these two parameters is particularly necessary to minimize the predictive error. When the model is well trained, the prediction on any new samples is simple, fast, and reliable.

### 2.2. Adaptive Modeling Using the Variable Selection Techniques

The IoT-based NIR calibration model for rapid estimation of wastewater pollutant analytes requires to train the parametric RBF-LSSVM model in combination of adaptive variable selections from the full-range spectral data, because the full spectrum contains kinds of noises and irrelevant information that may lead to low prediction accuracy and high calibration errors [30]. Waveband selection techniques are able to well address this problem. Several interval selection techniques have been proposed for the linear PLS regression, such as interval PLS (iPLS) [31], backward interval PLS (biPLS) [32], and synergy interval PLS (siPLS) [33]. The core idea of these algorithms is to select the optimal interval of variables and then use the selected variables to establish the prediction model. On this basis, the backward interval and synergy interval methods are investigated in combination optimization with the parametric RBF-LSSVM models for rapid assessment of environmental risk of wastewater pollution. As there is no other kernel utilized, the RBF-LSSVM models could be remarked simply as LSSVM, for the convenience of the next abbreviations in combination with backward interval and synergy interval.

#### 2.2.1. Backward Interval Selection.

Backward interval selection is to divide the full range spectrum into $n$ equal-width intervals. By targeting the $i$-th interval for prediction ($i = 1, 2 \cdots n$), the LSSVM model is established using the other $n - 1$ intervals. Each interval is predicted after a closed loop targeting and thus to obtain the prediction error of each interval. By comparison, the interval that has the worst prediction effect is removed, and then, the remaining intervals are used for remodeling. In cycles of modeling and interval removing, the prediction effect will finally stop decreasing so that the remaining intervals are taken as the informative wavebands for variable selection. The LSSVM models combined with the backward interval selection technique are operated with the grid search parametric optimization for regularization and the tuning of RBF kernel. The combination optimization model is defined as the back interval LSSVM (biLSSVM) model.

#### 2.2.2. Synergy Interval Combination.

Synergy interval combination is to evaluate the combination modeling effect of several intervals. Firstly, the full range spectrum is divided into $n$ equal-width intervals. For each $i = 1, 2 \cdots n$, each divided interval is predicted after a closed loop targeting, and then, the parametric LSSVM models were established for all possible combinations of $t$ intervals ($t = 2, 3 \cdots i$). By comparison, the most effective interval combination is selected for modeling. Moreover, the most optimal combination of intervals can be recognized in comparison of different $i$'s. The synergy interval combination technique is designed for combined optimization with the parametric LSSVM model and is defined as the synergy interval LSSVM (siLSSVM) model.

#### 2.3. Model Indicators of the IoT-Based NIR Application.

The NIR calibration model predictions of wastewater MIS were compared with the conventional analytical values. As the spectral data was acquired from various sensors along with IoT network nodes, the models are dynamically trained with the data flow. Models should be assessed using the indicators including the root mean square error (RMSE), the coefficient correlation (CC), and the relative RMSE (RRMSE) [34]. They
are calculated obeying the following formulae:

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{m - 1}},
\]

\[
\text{RRMSE} = \frac{\text{RMSE}}{y_{\text{ave}}},
\]

\[
\text{CC} = \frac{\sum_{i=1}^{m} (y_i - y_{\text{ave}})(\hat{y}_i - \hat{y}_{\text{ave}})}{\sqrt{\sum_{i=1}^{m} (y_i - y_{\text{ave}})^2 \sum_{i=1}^{m} (\hat{y}_i - \hat{y}_{\text{ave}})^2}},
\]

where \(y_i\) is the reference value from conventional analysis of the \(i\)-th sample, \(\hat{y}_i\) is the predictive value by NIR modeling, \(y_{\text{ave}}\) is the average value of the target samples, \(\hat{y}_{\text{ave}}\) is the average of the NIR predictive values, and \(m\) represents the number of target samples.

The IoT-based NIR calibration process requires to divide all of the wastewater samples into the calibration set, the validation set, and the test set. The calibration samples are used to train the LSSVM models; the validation samples are predicted and used to determine the optimal parameters for the LSSVM model, the biLSSVM model, and the siLSSVM model. Then, the test samples are used to test prediction abilities of the optimal models. With online detection by different sensors, the number of samples is dynamically changing in minutes. Experimental experiences show that the ratio of samples for calibration, validation, and test is suitable at about 2:1:1. Accordingly, the modeling indicators are denoted as \(\text{RMSE}_X\), \(\text{CC}_X\), and \(\text{RRMSE}_X\), where the subscript \(X\) represents \(C\) for calibration, \(V\) for validation, and \(T\) for test, respectively.

3. Experiments and Discussions

3.1. Data Acquisition. The study was conducted on the industrial production waste discharge line in south China. For data experiment, a total of 148 samples were collected from the 6 chemical plants. The sampling and monitoring operations obey the requirements of China’s national environmental monitoring technical specifications. Amount of 50 milliliter wastewater was saved for each sample. Then, the samples were sealed storage and transported to the lab.

As the conventional analysis of MIS costs consuming and destructive, the NIR spectra were measured in advance. The 148 wastewater samples were one by one scanned using the FOSS NIRS-Systems 5000 grating spectrometer (produced in Denmark) with InGaAs accessory. The experiment was operated under the constant temperature of 25 ± 1°C and constant humidity of 46 ± 1%RH. The spectrum was recorded in the waveband range of 780-2498 nm; the spectral resolution was set as 2 nm. Thus, there spread 860 digital wavelength variables for each sample spectrum. The NIR spectrum of wastewater shows the comprehensive response of chemical components, in which the pollution effect is mainly reflected by the indicators of MIS. The spectrum is much characterized by the peaks of water absorption around the frequencies of 1400 nm and 1950 nm, while the pollutants are weak-responded in the smooth spectral curve.

After the spectrum measurement, the samples were delivered for conventional analyses to detect the MIS content. In the traditional analysis of MIS, the wastewater sample was weighed in advance (denoted as \(T\)) and repeatedly filtered using a fine mesh, and the mesh was dried and weighed for several times until the weight became constant. Then, the residue weight of the filtered mesh over the clean mesh was taken as the weight of MIS (denoted as \(t_{\text{me}}\)). Finally, the MIS content was calculated by \(t_{\text{me}}/T\) and marked as the weight percentage (unit: %). The MIS content of the 148 samples falls in the range of 0.38-0.82 wt%, with the mean value and the standard deviation being 0.593 and 0.127, respectively. The data was used as the reference target values in NIR rapid modeling analysis.

As the experimental example, the whole of 148 wastewater samples were divided, with 72 samples for calibration, 38 for validation, and 38 for test.

3.2. Determination of MIS Using the Parametric RBF-LSSVM Model. The RBF-LSSVM models were established for the determination of wastewater MIS content. By the former machine learning experiences [35], it is worth to fuse the regularization parameter (\(\gamma\)) and the RBF kernel width (\(\sigma^2\)) together, for automatic tuning to search for the optimal combination of parameters. Deep training of (\(\gamma, \sigma\)) is necessary for observing a minimum predictive RMSEV in the calibration-validation process. The experiment was designed with presetting \(\gamma\) changing from 10 to 200 with the step of 10 and \(\sigma\) changing continuously from 1 to 15, such to have \(\sigma^2\) values of \(1^2, 2^2, \ldots, 15^2\), respectively. In grids, the RBF-LSSVM model was trained as obtaining all of the predictive RMSEV values for each combination of (\(\gamma, \sigma\)). The results are showed in Figure 1.

We observed from Figure 1 that serious frustration appeared with the variation of \(\gamma\) and gentle smooth change goes with \(\sigma\). Specifically, the optimal RBF-LSSVM model was identified with \(\gamma = 120\) and \(\sigma = 11\), respectively. The correspondent model predictive RMSEV was minimized as 0.055 wt%. Correspondingly, \(\text{CC}_V\) was acquired as 0.891 (the CC value over 0.85 is regarded as high enough for the NIR rapid detection in fields of agriculture and environment [36]). The validation results indicated that the fuse selection of \(\gamma\) and \(\sigma\) is on critical demand for the optimization of the RBF-LSSVM model in NIR quantitative analysis of wastewater MIS content. Nevertheless, more discussions about the variable selections are expected to have improved modeling effects. Hereafter, we could use LSSVM as short for RBF-LSSVM.

3.3. Model Improvement with Adaptive Variable Selection. In the way of subinterval division, the waveband variables were partially selected for the test of LSSVM models. Based on the total of 860 digital wavelengths, we design to have the full waveband range split into 43 subintervals for adaptive machine learning, such that each interval includes 20 wavelength variables. The intervals were equivalently split and set with serial numbers (see Figure 2). Then, all of the...
intervals were tested for LSSVM models by adopting the backward interval and synergy interval methods, as to form a smart way to train the models based on different variable combinations. The intelligent learning models are denoted as biLSSVM and siLSSVM.

3.3.1. Optimization of the biLSSVM Model. The biLSSVM method is to combine the grid search of LSSVM parameters with the backward removal of the “worst” waveband intervals. Targeting each of the 43 intervals, the variables in the other 42 intervals were used to establish the LSSVM model, to observe the prediction effect of the target interval. Based on the calibration and validation sample sets, the algorithmic parameters of $\gamma$ and $\sigma^2$ were optimized during the LSSVM process. In a closed loop modeling, all of the 43 intervals have their individual optimal modeling results. By searching for the best combination of intervals, the mono interval with the largest value of RMSE$_V$ was removed out of the variable set prepared for the next modeling, and then, the remaining intervals were used for remodeling. In our analysis of wastewater MIS, the No. 31 interval was firstly removed, No. 38 secondly, No. 6 thirdly, etc. The interval removal order is showed in Figure 3, and the correspondent changing RMSE$_V$ values by the biLSSVM prediction are also demonstrated. It is learnt from Figure 3 that the biLSSVM model became better at the beginning as the intervals removed. It reached the best after 30 times removal, and then, RMSE$_V$ goes back to increase. The model was trained with its minimum RMSE$_V$ of 0.044 wt%; the corresponding best combination of variables includes 13 intervals, which are particularly identified with the shape of the NIR spectrum of wastewater (showed in Figure 4). The selected interval combination of variables is taken as the informative quasidiscrete wavelengths for the NIR quantitative determination of wastewater MIS content using the biLSSVM optimization.

3.3.2. Optimization of the siLSSVM Model. The siLSSVM method is to combine the grid search of LSSVM parameters with the synergy modeling of several waveband intervals. As abovementioned, the full range spectrum is divided into
43 equal-width intervals. Each individual interval is predicted with LSSVM parametric optimization. After a closed loop, LSSVM models were further established based on possible combinations of \( t \) intervals \((t = 2, 3 \ldots 43)\). For example, when targeting \( t = 2 \), LSSVM models were established based on all possible synergy cases of two-interval combinations, and then, the best case was selected by minimizing the predictive \( \text{RMSE}_V \), such to identify the optimal combination of two intervals. As the number of synergy intervals changes, the optimal siLSSVM models based on each \( t \) value were identified with their \( \text{RMSE}_V \) results (see Figure 5). We observe from Figure 5 that the most informative variable combination was obtained with \( t = 13 \), namely, 13 specific intervals were selected by siLSSVM as including informative wavelength variables for the determination of wastewater MIS. They are located in the full range with the spectrum shape of wastewater (see Figure 6). The selected intervals are combined to establish the LSSVM model, thus acquiring \( \text{RMSE}_V \) of 0.041 wt% and \( \text{CC}_V \) of 0.923. It was slightly better than the result from the biLSSVM optimization model. Also, we acknowledge that some of the intervals selected by siLSSVM are repeated with those selected by biLSSVM.
3.4. Model Comparison and Evaluation. The backward interval and synergy interval techniques were proposed as the intelligent section in combination use with the kernel-based LSSVM parametric learning method. In order to examine the effect of the smart model, the model prediction effects were compared with the traditional biPLS and siPLS methods, for the model training process based on the validation sample set, and for the model evaluation based on the test sample set. Logically, the common LSSVM and the conventional PLS modes were also listed for comparison. The optimal results of each method are showed in Table 1.

Table 1 shows that the models based on the LSSVM framework algorithm generate better prediction results than the PLS-based models. And also, the models with variable selections by the backward interval and synergy interval techniques play better roles in predictions of wastewater MIS. Of all, the siLSSVM model has the best optimal training results with its RMSEV as 0.041 wt% and CCV as 0.923. The correlation coefficient is well high enough for NIR spectroscopic detections in the environmental field. On the contrary, the
best evaluation model was observed as established by biLSSVM, with RMSE\textsubscript{T} as 0.059 wt\% and CC\textsubscript{T} as 0.896. And the siLSSVM model was acquired as the second place in the evaluation. To display the evaluation effects of the best biLSSVM and siLSSVM model, the regression plots of the 38 test samples are illustrated in Figure 7. It is acknowledged that the evaluation correlations of biLSSVM and siLSSVM did not reach but were very near to 0.9, accompanied with the approximate 10% of RRMSE\textsubscript{T}. These are quite appreciating results for real applications. The conclusions demonstrated that the proposed intelligent methods of combining the backward interval and synergy interval techniques to the parametric LSSVM learning model are feasible to improve the prediction results in rapid determination of the wastewater MIS content using the IoT-based NIR technology.

4. Conclusions

In this work, the IoT-based NIR spectroscopic technology was utilized for quantitative detection of wastewater MIS contents. The goal is to observe the optimal machine learning model for fast estimating wastewater pollutions. For modeling innovation, the backward interval and synergy interval techniques were proposed as for adaptive variable selections, in combination use with the parametric LSSVM learning method. The machine learning framework utilized the RBF kernel as for common non-linear mapping. Thus, the regulation parameter and the kernel width are fused together for two-dimensional grid search optimization.

The total of 860 digital wavelengths in the full scanning range was split into 43 equivalent subintervals. Each interval has 20 continuous wavelength variables. Then, all of the intervals were tested for LSSVM models in the way of backward interval removal and synergy interval combination. By the biLSSVM method, the model was trained with the optimal RMSE\textsubscript{V} of 0.044 wt\% and the corresponding CC\textsubscript{V} of 0.918. The best combination of waveband variables includes the 13 intervals showed in Figure 4. By siLSSVM, the best training model was identified as the combination of the 13 intervals showed in Figure 6. These intervals are combined to form the optimal LSSVM model, acquiring the minimum RMSE\textsubscript{V} of 0.041 wt\% and a correspondent CC\textsubscript{V} of 0.923. In comparison with the common LSSVM, conventional PLS, biPLS, and siPLS models, the optimal training results of the siLSSVM model were recognized as the best prediction based on the validation sample set, while the best evaluation model was established by biLSSVM for the test sample set. These prediction results demonstrated that the proposed biLSSVM and siLSSVM frameworks with the adaptive variable selection methods are feasible to improve the model prediction results in the rapid determination of the wastewater MIS content by the intelligent IoT-based NIR technology.

The conclusions indicated that the proposed machine learning framework of combining the backward interval and synergy interval techniques with the RBF-kernel-based LSSVM algorithm is prospectively an intelligent improvement for the IoT application to estimate the complex target of wastewater, as to partially assess the environmental risk of water safety in the industrial fields.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.
Acknowledgments

This work was supported by the National Natural Science Foundations of China (61505037 and 62003379), Natural Science Foundation of Guangxi Province (2018GXNSFAA050045), and Science and Technology Program of Guangzhou (202002030246).

References

[1] A. Goap, D. Sharma, A. K. Shukla, and C. Rama Krishna, “An IoT based smart irrigation management system using machine learning and open source technologies,” Computers and Electronics in Agriculture, vol. 155, no. 5, pp. 41–49, 2018.

[2] J. Gubbini, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (IoT): a vision, architectural elements, and future directions,” Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[3] D. K. Sharma, S. K. Dhurandher, M. S. Obaidat, S. Pruthi, and B. Sadoun, “A priority based message forwarding scheme for opportunistic networks,” in 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–4, Kunning, China, 2016.

[4] J. Su, Z. Sheng, A. Liu, Z. Fu, and Y. Chen, “A time and energy saving-based frame adjustment strategy (TES-FAS) tag identification algorithm for UHF RFID systems,” IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 2974–2986, 2020.

[5] S. Areekij, P. Ritthiruangdej, S. Kasemsumran, N. Therdthai, V. Haruthaithanasan, and Y. Ozaki, “Rapid and nondestructive analysis of deep-fried taro chip qualities using near infrared spectroscopy,” Journal of Near Infrared Spectroscopy, vol. 25, no. 2, pp. 127–137, 2017.

[6] S. Violino, L. Ortenzi, F. Antonucci et al., “An artificial intelligence approach for Italian EVOO origin traceability through an open source IoT spectrometer,” Foods, vol. 9, no. 6, p. 834, 2020.

[7] S. Grassi and C. Alamprese, “Advances in NIR spectroscopy applied to process analytical technology in food industries,” Current Opinion in Food Science, vol. 22, pp. 17–21, 2018.

[8] H. Khayyam, A. Jamali, A. Bab-Hadiashar et al., “A novel hybrid machine learning algorithm for limited and big data modeling with application in Industry 4.0,” IEEE Access, vol. 8, pp. 111381–111393, 2020.

[9] H.-Z. Chen, L.-L. Xu, G.-Q. Tang, Q.-Q. Song, and Q.-X. Fang, “Rapid detection of surface color of Shatian pomelo using VIS-NIR spectrometry for the identification of maturity,” Food Analytical Methods, vol. 9, no. 1, pp. 192–201, 2016.

[10] D. Ramírez-Morales, E. F.-B. Rivero, and A. Pazo’s, “Optimization of NIR calibration models for multiple processes in the sugar industry,” Chemometrics and Intelligent Laboratory Systems, vol. 159, no. 7, pp. 45–57, 2016.

[11] S. Cipullos, S. Nawar, A. M. Mouazen, P. Campo-Moreno, and F. Coulon, “Predicting bioavailability change of complex chemical mixtures in contaminated soils using visible and near-infrared spectroscopy and random forest regression,” Scientific Reports, vol. 9, no. 1, p. 4492, 2019.

[12] C. Lodeiro, J. L. Capelo, E. Oliveira, and C. Nuñez, “Pollutant toxic ions and molecules. A global pollution problem: trends in detection and protection,” Environmental Science and Pollution Research, vol. 23, no. 24, pp. 24419–24421, 2016.

[13] W. He, N. Qin, X.-Z. Kong et al., “Water quality benchmarking (WQB) and priority control screening (PCS) of persistent toxic substances (PTSs) in China: necessity, method and a case study,” Science of the Total Environment, vol. 472, no. 5, pp. 1108–1120, 2014.

[14] L. Dvořák, M. Gómez, J. Dolina, and A. Černin, “Anaerobic membrane bioreactors—a mini review with emphasis on industrial wastewater treatment: applications, limitations and perspectives,” Desalination and Water Treatment, vol. 57, no. 41, pp. 19062–19076, 2016.

[15] D.-H. Kang, D. Kim, S. Wang, D. Song, and M.-H. Yoon, “Water-insoluble, nanocrystalline, and hydrogel fibrillar scaffolds for biomedical applications,” Polymer Journal, vol. 50, no. 8, pp. 637–647, 2018.

[16] J. Su, Z. Sheng, V. C. M. Leung, and Y. Chen, “Energy efficient tag identification algorithms for RFID: survey, motivation and new design,” IEEE Wireless Communications, vol. 26, no. 3, pp. 118–124, 2019.

[17] F. Omar, H. Atan, and M. Z. MatJafri, “NIR spectroscopic properties of aqueous acids solutions,” Molecules, vol. 17, no. 6, pp. 7440–7450, 2012.

[18] S. Le Caër, G. Klein, D. Ortiz et al., “The effect of myoglobin crowding on the dynamics of water: an infrared study,” Physical Chemistry Chemical Physics, vol. 16, no. 41, pp. 22841–22852, 2014.

[19] D.-H. Jung, H.-J. Kim, H. S. Kim, J. Choi, J. Do Kim, and S. H. Park, “Fusion of spectroscopy and cobalt electrochemistry data for estimating phosphate concentration in hydroponic solution,” Sensors (Basel), vol. 19, no. 11, p. 2596, 2019.

[20] C. Charmier, E. Latrille, J. Jimenez et al., “Fast characterization of solid organic waste content with near infrared spectroscopy in anaerobic digestion,” Waste Management, vol. 59, pp. 140–148, 2017.

[21] K.-I. Suehara, K. Owari, J. Kohda, Y. Nakano, and T. Yano, “Rapid and simple determination of oil and urea concentrations and solids content to monitor biodegradation conditions of wastewater discharged from a biodiesel fuel production plant,” Journal of Near Infrared Spectroscopy, vol. 15, no. 2, pp. 89–96, 2007.

[22] S. Janjarasjitt and K. A. Loparo, “Comparison of complexity measures using two complex system analysis methods applied to the epileptic EEG,” Journal of the Korean Physical Society, vol. 63, no. 8, pp. 1659–1665, 2013.

[23] M. P. Diago, A. Bellincontro, M. Scheidweller, J. Tardaguila, S. Tittmann, and M. Stoll, “Future opportunities of proximal near infrared spectroscopy approaches to determine the variability of vineyard water status,” Australian Journal of Grape and Wine Research, vol. 23, no. 3, pp. 409–414, 2017.

[24] H. Chen, L. Xu, W. Ai, B. Lin, Q. Feng, and K. Cai, “Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy,” Science of the Total Environment, vol. 714, p. 136765, 2020.

[25] J. H. Ahn and Y. J. Park, “Estimating water reflectance at near-infrared wavelengths for turbid water atmospheric correction: a preliminary study for goci-ii,” Remote Sensing, vol. 12, no. 22, p. 3791, 2020.

[26] H. Chen, A. Chen, L. Xu et al., “A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources,” Agricultural Water Management, vol. 240, p. 106303, 2020.
[27] C.-C. Chang and C.-J. Lin, “LIBSVM,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1–27, 2011.

[28] H. Chen, L. Xu, Z. Jia, K. Cai, K. Shi, and J. Gu, “Determination of parameter uncertainty for quantitative analysis of shaddock peel pectin using linear and nonlinear near-infrared spectroscopic models,” Analytical Letters, vol. 51, no. 10, pp. 1564–1577, 2018.

[29] M. Ring and B. M. Eskofier, “An approximation of the Gaussian RBF kernel for efficient classification with SVMs,” Pattern Recognition Letters, vol. 84, pp. 107–113, 2016.

[30] M. J. Anzanello and F. S. Fogliatto, “A review of recent variable selection methods in industrial and chemometrics applications,” European J. of Industrial Engineering, vol. 8, no. 5, pp. 619–645, 2014.

[31] H. Chen, C. Tan, Z. Lin, and T. Wu, “Rapid determination of cotton content in textiles by near-infrared spectroscopy and interval partial least squares,” Analytical Letters, vol. 51, no. 17, pp. 2695–2707, 2018.

[32] R. Leardl and L. Norgaard, “Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions,” Journal of Chemometrics, vol. 18, no. 11, pp. 486–497, 2004.

[33] Y. Li, T. Fang, S. Zhu, F. Huang, Z. Chen, and Y. Wang, “Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 189, pp. 37–43, 2018.

[34] K. D. Shepherd and M. G. Walsh, “Development of reflectance spectral libraries for characterization of soil properties,” Soil Science Society of America Journal, vol. 66, no. 3, pp. 988–998, 2002.

[35] L. Mo, H. Chen, W. Chen, Q. Feng, and L. Xu, “Study on evolution methods for the optimization of machine learning models based on FT-NIR spectroscopy,” Infrared Physics & Technology, vol. 108, no. 5, p. 103366, 2020.

[36] C. Nduwamungu, N. Ziadi, L. É. Parent, G. F. Tremblay, and L. Thurès, “Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: a review,” Canadian Journal of Soil Science, vol. 89, no. 5, pp. 531–541, 2009.