water

MDPI

Article

A Multidisciplinary Approach for Evaluating Spatial and Temporal Variations in Water Quality

Viet Thang Le [1], Nguyen Hong Quan [2], Ho Huu Loc [3,4,*,†], Nguyen Thi Thanh Duyen [2], Tran Duc Dung [2], Hiep Duc Nguyen [5,6] and Quang Hung Do [7]

1 Institute of Environment Science, Engineering and Management, Industrial University of Ho Chi Minh City, 12 Nguyen Van Bao Street, Go Vap District, Ho Chi Minh City 700000, Vietnam; levietthang@iuh.edu.vn
2 Center for Water Management and Climate Change, Vietnam National University, Ho Chi Minh City 700000, Vietnam; hongquanmt@yahoo.com (N.H.Q.); nttduyen31@gmail.com (N.T.T.D.); dungtranducvn@yahoo.com (T.D.D.)
3 Nanyang Environment & Water Research Institute, Nanyang Technological University, 1 Cleantech Loop, CleanTech One, Singapore #06-08637141, Singapore
4 NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City 700000, Vietnam
5 Environmental Quality, Atmospheric Science and Climate Change Research Group, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam; nguyenduchiep@tdt.edu.vn
6 Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
7 Faculty of Information Technology, University of Transport Technology, Hanoi 100000, Vietnam; hungdq@utt.edu.vn or quanghung2110@gmail.com
* Correspondence: hohl@ntu.edu.sg
† Corresponding author is currently at Nanyang Environment & Water Research Institute, Nanyang Technological University, Singapore (NEWRI/NTU).

Received: 14 March 2019; Accepted: 12 April 2019; Published: 24 April 2019

Abstract: The primary goal of this study is to investigate the classification capability of several artificial intelligence techniques, including the decision tree (DT), multilayer perceptron (MLP) network, Naïve Bayes, radial basis function (RBF) network, and support vector machine (SVM) for evaluating spatial and temporal variations in water quality. The application case is the Song Quao-Ca Giang (SQ-CG) water system, a main domestic water supply source of the city of Phan Thiet in Binh Thuan province, Vietnam. To evaluate the water quality condition of the source, the government agency has initiated an extensive sampling project, collecting samples from 43 locations covering the SQ reservoir, the main canals, and the surrounding areas during 2015–2016. Different classifying models based on artificial intelligence techniques were developed to analyze the sampling data after the performances of the models were evaluated and compared using the confusion matrix, accuracy rate, and several error indexes. The results show that machine-learning techniques can be used to explicitly evaluate spatial and temporal variations in water quality.

Keywords: water quality; temporal and spatial assessment; multilayer perceptron (MLP) network; radial basis function (RBF) network; decision tree (DT)

1. Introduction

As one of the most important elements responsible for life, water quality status has a large impact on human life and public health. It is necessary to identify whether the quality of water sources is suitable for a certain purpose, meeting the requirements of people [1].

Water quality data are very important for assessing the health of the environment of water bodies, i.e., water pollution [2]. However, their measurements are usually unavailable or limited due to the lack of monitoring systems, especially in developing countries [2,3]. Additionally, water quality

stations are usually too expensive to set up and maintain [4]. In recent years, collecting short-term water quality data has often been considered by these countries as an urgent solution to identify the causes of water pollution. However, these investigation activities are costly because it is challenging to explore polluted sources which might be affected by various driven factors. A number of water quality samples thus need to be carefully determined before the investigation starts in order to spend the lowest cost, but still provide enough data for statistical analysis.

Though located in the tropical region with potential water resources, Vietnam has been suffering from a water shortage in both quantity and quality [5]. Water is mostly being polluted by untreated wastewater from point and diffuse sources [6,7]. The study area Binh Thuan is one of those central arid provinces of Vietnam with low rainfall intensities, of which the total volume is unevenly distributed across the landscape [8]. In recent years, the local municipality has developed an extensive irrigation network system to improve the situation [9]; in particular, the Song Quao-Ca Giang (SQ-CG) system of irrigation works, completed in late 1997, which is the main source of water supply for the livelihood and economic activities of Phan Thiet City (administration center of Binh Thuan province), Ma Lam Town, and the surrounding area, with a total capacity of 53,360 m^3 per day. The water from the SQ reservoir is carried through the main channel to the domestic water supply system of Phan Thiet city. Despite its importance, the system has insufficient water quality monitoring programs in terms of the absence of automatic monitoring stations, scheduled sampling, etc. Systematic and holistic investigations of SQ-CG system water quality, as such, have never been performed prior to this study. With the uncontrolled developments of agriculture in the upstream and midstream areas, the water resources of SQ-CG could be left deteriorated.

Effective management of river water quality is crucial; hence, a notable body in the literature is devoted to this topic [10,11]. In order to classify water quality indexes, statistical and multivariate analyses, including cluster analysis, discriminant analysis, and principal component analysis. have been applied to analyze the water quality status [12–14]. However, to develop more effective classifying models rather than traditional water-quality assessment methods, it is better to utilize computational intelligence models [15]. The computational intelligence approach has the capability to explore the nonlinear relationship and discover hidden knowledge from the dataset. As a result, this approach has been applied to a number of practical problems in various scientific disciplines.

First, this research aims to contribute the primary systematic water-quality investigation of SQ-CG using data obtained from 43 sampling stations located in the SQ reservoir, its catchment, and the associated canals. Each sampling location contributes six samples collected during the wet season of 2015 and dry season of 2016. Our original thrust is to contribute a baseline understanding of the spatial and temporal variations of the water quality parameters to support appropriate management decisions, as referenced from [14]. In addition, the presented method is expected to contribute a practical method to effectively explore the river quality of sparsely gauged catchments such as SQ – CG. Other than that, this study examines the validity of several artificial intelligence (AI) techniques, including the decision tree (DT), multilayer perceptron (MLP) network, Naïve Bayes, radial basis function (RBF) network, and support vector machine (SVM), in classifying water quality.

The remainder of the paper is organized as follows. A brief introduction of the theoretical backgrounds of the five AI techniques used is represented in Section 2. This section also summarizes how the water quality samples were collected. Section 3 presents results and discussions, followed by conclusions in Section 4.

2. Methods

In this study, five of the most commonly used AI techniques were applied, each of which is introduced in the following sections. The aim of this section is to give some first-hand-knowledge of AI techniques in exploring the structure of datasets and how they can be applied in water quality studies. In essence, this study expands on previous findings in these respects, e.g., [16–19].

2.1. AI Techniques

2.1.1. Multilayer Perceptron (MLP) Network

Artificial neural network (ANNs) are a form of artificial intelligence based on the function of the human brain and nervous system. An artificial neural network has two types of basic components, including a neuron and link. A neuron is a processing element and a link is used to connect one neuron with another. Each link has its own weight. Each neuron receives stimulation from other neurons, processes the information, and produces an output. Neurons are organized into a sequence of layers. The first and last layers are called input and output layers, respectively, and the middle layers are called hidden layers. The input layer is a buffer that presents data to the network. It is not a neural computing layer because it has no input weights and no activation functions. The hidden layer has no connections to the outside world. The output layer presents the output response to a given input. The activation coming into a neuron from other neurons is multiplied by the weights on the links over which it spreads and is then added together with other incoming activations.

A neural network in which activations only spread in a forward direction from the input layer through one or more hidden layers to the output layer is known as a multilayer feed-forward network. For a given set of data, a multilayer feed-forward network can give a good non-linear relationship. Studies have shown that a feed-forward network, even with only one hidden layer, can approximate any continuous function [20,21]. Therefore, a feed-forward network is an attractive approach [22]. Figure 1 shows an example of a feed-forward network with three layers. In Figure 1, R, N, and S are the number of inputs, hidden neurons, and outputs, respectively; iw and hw are the input and hidden weights matrices, respectively; hb and ob are the bias vectors of the hidden and output layers, respectively; x is the input vector of the network; ho is the output vector of the hidden layer; and y is the output vector of the network. The neural network in Figure 1 can be expressed through the following equations:

$$h_{oi} = f\left(\sum_{j=1}^{R} iw_{i,j}.x_j + hb_i\right), \text{ for } j = 1,\ldots,N \quad (1)$$

$$y_i = f\left(\sum_{k=1}^{N} hw_{i,k}.ho_k + ob_i\right), \text{ for } i = 1,\ldots,S \quad (2)$$

where f is an activation function.

When implementing a neural network, it is necessary to determine the structure in terms of the number of layers and the number of neurons in the layers. The larger the number of hidden layers and nodes, the more complex the network will be. A network with a structure that is more complicated than necessary overfits the training data [21]. This means that it performs well on data included in the training set, but it may perform poorly on that in a testing set.

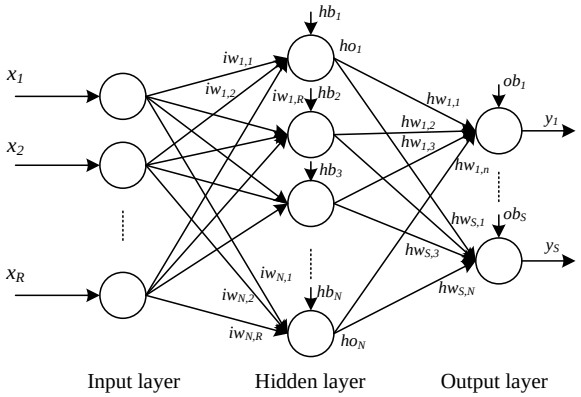

Figure 1. A feed-forward network with three layers.

Once a network has been structured for a particular application, it is ready for training. Training a network means finding a set of weights and biases that will give desired values at the network's output when presented with different patterns at its input. When network training is initiated, the iterative process of presenting the training data set to the network's input continues until a given termination condition is satisfied. This usually happens based on a criterion indicating that the current achieved solution is good enough to stop training. Some of the common termination criteria are the sum of squared error (SSE) and mean squared error (MSE). Through continuous iterations, the optimal or near-optimal solution is finally achieved, which is regarded as the weights and biases of a neural network. Suppose that there are m input-target sets, x_k–t_k for k = 1, 2, ... , m for neural network training. Thus, network variables arranged as *iw*, *hw*, *hb*, and *ob* are to be changed to minimize a cost function. E, such as the MSE between network outputs, y_k, and desired targets, t_k, is as follows:

$$\text{MSE} = \frac{1}{m}\sum_{k=1}^{m} e_k^2 = \frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2 \tag{3}$$

2.1.2. Radial Basis Function (RBF) Network

The RBF network is a kind of kernel function network that uses kernel functions, located in different neighborhoods of the input space. The architecture of the RBF network includes three layers: the input layer, the hidden layer, and the output layer, as shown in Figure 2. Although the structure of the Radial Basis Function (RBF) neural network is rather simple, the network has a strong generalization ability [23,24]. The RBF neural network has shown a good classification and approximation performance in various applications [25,26].

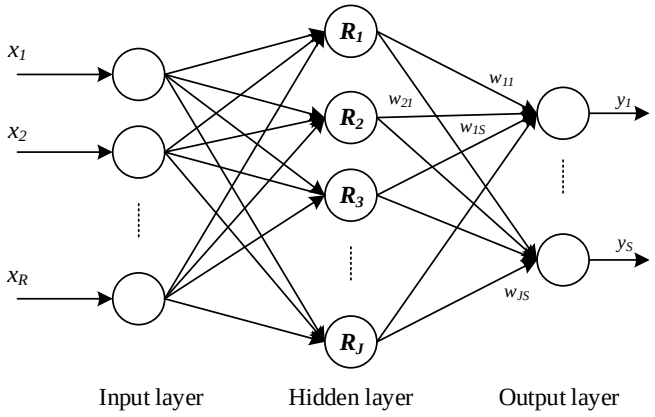

Figure 2. An RBF network.

As shown in Figure 2, the estimated output is a weighted summation utilizing the following equation:

$$y_s = \sum_{j=1}^{J} w_{js} R_j(x), \text{ s } = 1, 2, \ldots, S \tag{4}$$

where S denotes the number of outputs, J is the number of nodes in the hidden layer, and w_{js} is the connection weight between j—the node of the hidden layer and S—the node of the output layer. There are several radial basis functions; the most commonly used one is as follows:

$$R_j(x) = \exp\left(-\frac{\|x - c_j\|}{2\sigma_j^2}\right), \text{ j } = 1, 2, \ldots, J \tag{5}$$

where x is the input pattern vector, where each input is represented by the N -dimensional vector; c_j and σ_j are the center and width of RBF, respectively; and $\|x - c_j\|$ is the norm of the vectors x and c_j, which can be considered as the distance between the vectors x and c_j.

Through the RBF network, the relationship between the input and output is established. The design and training of an RBF are conducted through the estimation of three kinds of parameters, including the center and width of radial basis functions and the connection weights.

2.1.3. Decision Tree (DT)

The decision tree is also one of the most used intelligence techniques because of its simplicity in understanding and interpreting the results. A DT classifies original input variables into subgroups that construct a tree with a root node, internal nodes, and leaf nodes. A decision tree (DT) can be considered as a hierarchical model composed of decision rules that recursively split independent inputs into homogenous sections [27]. The aim of constructing a DT is to explore the set of decision rules that can be used to predict outcomes from a set of input variables. Applying a DT on a dataset would predict the target variable of a new dataset record. A DT is also called a regression or classification tree if the target variables are continuous or discrete, respectively [28]. The DT can give an idea of the importance of an attribute in a dataset.

2.1.4. Support Vector Machine (SVM)

SVM is a supervised learning method influenced by advances in statistical learning theory [29]. SVM has been successfully applied to various applications in classification and recognition problems. Using training data, SVM maps the input space into a high dimensional feature space. In the feature space, the optimal hyperplane is identified by maximizing the margins or distances of class boundaries. The training points that are closest to the optimal hyper plane are called support vectors. When the decision surface is obtained, it can then be used for classifying new data.

Consider a training dataset of feature-label pairs (x_i, y_i) with $i= 1, \ldots, n$. The optimum separating hyperplane is represented as

$$g(x) = sign\left(\sum_{i=1}^{n} y_i \alpha_i K\left(x_i, x_j\right) + b\right) \tag{6}$$

where $K(x_i, x_j)$ is the kernel function, α_i is a Lagrange multiplier, and b is the offset of the hyperplane from the origin. This is subject to constraints $0 \le \alpha_i \le C$ and $\sum \alpha_i y_i = 0$, where α_i is a Lagrange multiplier for each training point and C is the penalty. Only those training points lying close to the support vectors have non-zero α_i. However, in real-world problems, data are noisy and there will be no linear separation in the feature space. Hence, the optimum hyperplane can be identified as

$$y_i(w.x_i + b) \ge 1 - \zeta_i,\ \zeta_i \ge 0 \tag{7}$$

where w is the weight vector that determines the orientation of the hyperplane in the feature space and ζ_i is the i of the positive slack variable that measures the amount of violation from the constraints.

E. Naive Bayes Classifier

A Naive Bayes classifier is based on Bayes' theorem and the probability that a given data point belongs to a particular class [30]. Assume that we have m training samples $\left(x_i, y_j\right)$, where $x = (x_{i1}, x_{i2}, \ldots, x_{in})$ is a n-dimensional vector and y_i is the corresponding class. For a new sample x_{tst}, we wish to predict its class y_{tst} using Bayes' theorem:

$$y_{tst} = \underset{y}{\text{argmax}} P(y|x_{tst}) = \underset{y}{\text{argmax}} \frac{P(x_{tst}|y)P(y)}{P(x_{tst})} \tag{8}$$

However, the above equation requires an estimation of distribution $P(x|y)$, which is impossible in some cases. A Naive Bayes classifier makes a strong independence assumption on this probability distribution using the following equation:

$$P(x|y) = \prod_{j=1}^{n} P\left(x_j|y\right) \tag{9}$$

This means that individual components of x are conditionally independent given its label y. The task of classification now proceeds by estimating n one-dimensional distributions $P\left(x_j|y\right)$.

Table 1 summarizes some of the advantages and disadvantages of the AI techniques presented in this study.

Table 1. Advantages and disadvantages of the AI technique-based classifier.

Classifier	Advantages	Limitations
MLP Network	+ Easy to design + Few parameters	- Requiring high computational time - The training period may be slow - Difficult to identify the number of neurons and layers
RBF Network	+ Easy to design + Good generalization + More fast learning	- Sensitive to the dimensionality of data - Necessary of the preliminary setting of neurons and basic functions
DT	+ DT-based models are easily interpreted. + Easy to produce the model. + Can be used for both discrete and continuous values	- Not working well on the small training dataset - Overfitting problem. - A dataset with a small variation can produce different decision trees
SVM	+ High accuracy performance capability + Working well even if the dataset is not linearly separable	- The high cost of computation - High memory usage
Naïve Bayes Classifier	+ Simple to implement. + Providing accurate results in most of classification and prediction problems. + High computational efficiency	- The precision will decrease when the size of the dataset is small.

All models were coded in the Matlab 2015a environment. To avoid the over-fitting problem, 10-fold cross validation was utilized. For each technique, various sets of parameters were tried to obtain the best architecture of each classifying model. To evaluate the performance of the classifying model, several performance criteria were used. These criteria were applied to know how well the developed models worked. They are as follows: the percentage of accurate and inaccurate classification, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), root relative squared error (RRSE), and confusion matrix. Figure 3 shows the application framework of using the AI technique in the application case.

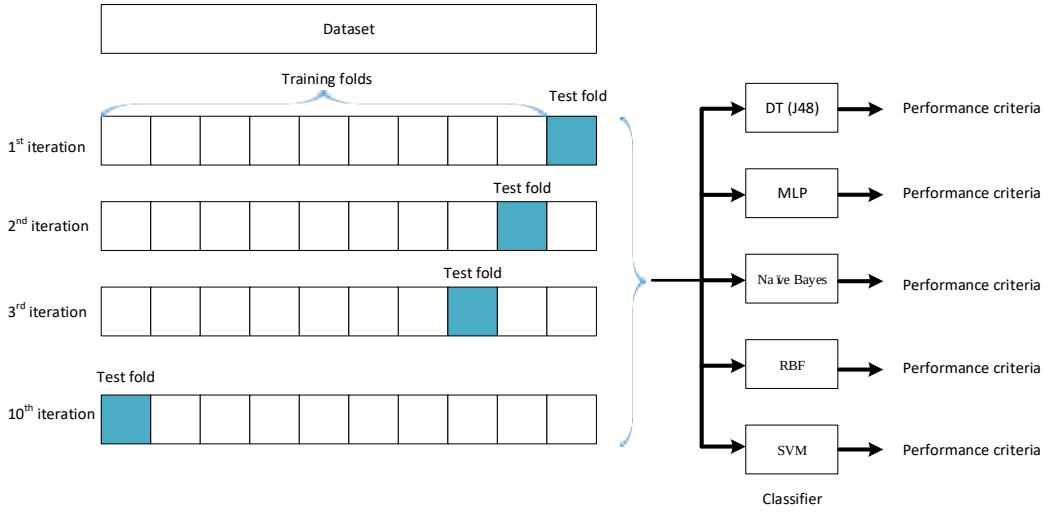

Figure 3. The application framework of using the AI technique in classification.

2.2. Dataset

Data on water quality were collected at 43 locations, as depicted in Figure 4. These include 23 stations on Song Quao Reservoir: R1 → R23; five stations in the surrounding areas, i.e., two upstream: SA1 & SA2 and three downstream: SA3 → SA5; and 15 stations along the main canal: C1 → C15.

We performed a total of six field trips to collect the samples during the wet season of 2015 and dry season of 2016, as follows:

- The wet season (November to April): three times (26–27 August 2015; 29–30 September 2015; and 26–27 October 2015);
- The dry season (November to April): three times (28–29 March 2016; 28–29 April 2016; and 26–27 May 2016).

The collected samples were analyzed at the Biochemical Laboratory of the Binh Thuan Centre of Standardization Metrology and Quality Control. We followed the international standard to collect, preserve, and analyze the samples, as regulated by [31–34]. In this study, the analyzed parameters include pH, TSS (Total Suspended Solids), DO (Dissolved Oxygen), COD (Chemical Oxygen Demand), BOD5 (Biological Oxygen Demand), Ammonium (N-NH_4^+), Nitrite (N-NO_2^-), Nitrate (N-NO_3^-),), Phosphate (P-PO_4^{3-}), Total (total Nitrogen), TP (total phosphorous), and Coliform. We also thrived to detect important heavy metals, i.e., Zinc (Zn), Cadmium (Cd), Arsenic (As), Lead (Pb), Crom VI (Cr), Manganese (Mn), Total Iron (Tot. Fe), Nickel (Ni), and Mercury (Hg). These analyses were performed via 36 samples collected from six stations: C1, C11, and C14 (canals); R19 (reservoir); and SA1 and SA4 (surrounding areas), hence solely subjected to descriptive statistics. All of the collected records were subsequently compared with respective national standards regulated in the national standards, widely quoted as QCVN 08-MT: 2015/BTNMT [35].

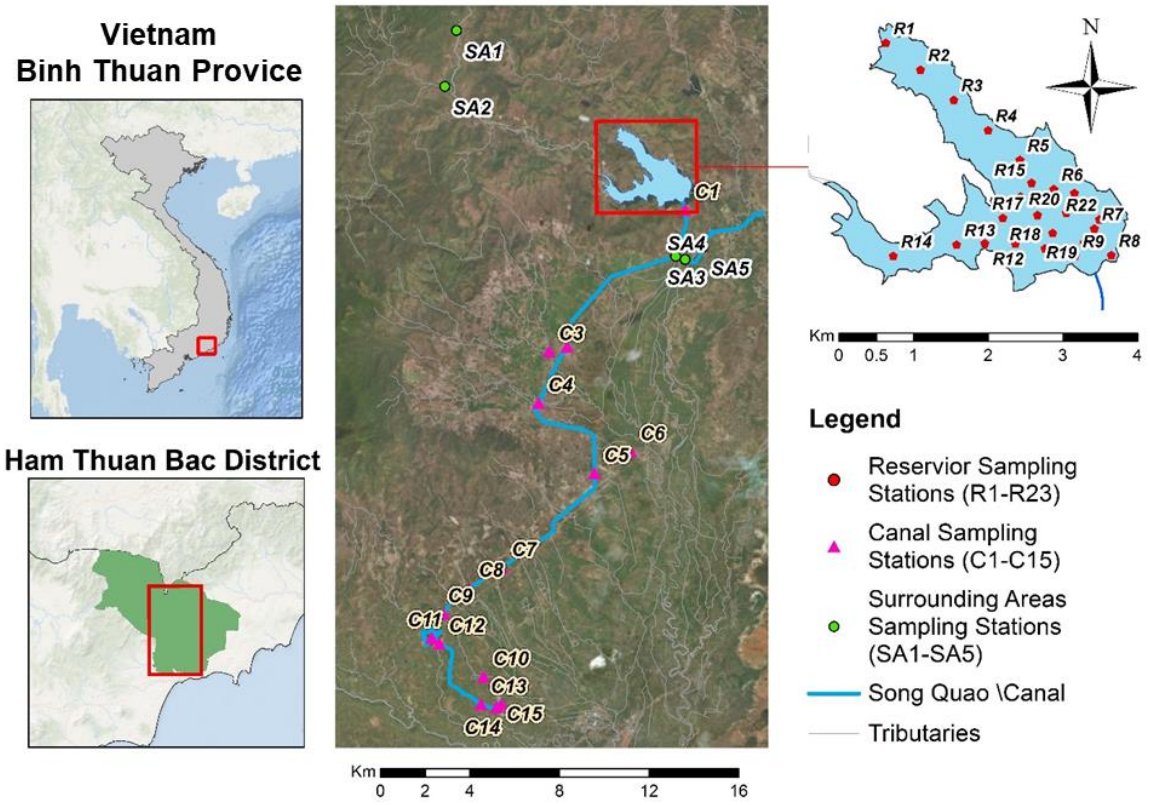

Figure 4. Map of study area and water quality sampling locations.

3. Results and Discussion

3.1. Preliminary Assessment of Water Quality

In general, the water quality of the SQ-CG system lies within the regulated criteria in terms of biochemical parameters, with the exceptions of BOD and COD; however, with marginal exceedance. More specifically, seven out of 43 sampling stations had mean BOD values exceeding the standard, all of which are located along the canals. Descriptive statistics, including the maximum, minimum, mean, and standard deviation of the samples, along with the regulated values of biochemical parameters, are summarized in Table 2.

Table 2. Descriptive statistics of observed variables.

Observed Variables	Minimum	Maximum	Mean	Standard Deviation	QCVN 08-MT: 2015/BTNMT
pH	6.75	8.9	7.57	0.315	6–8.5
DO	5	5.82	5.221	0.131	≥5
BOD	3	9	5.159	1.052	≤6
COD	8.3	17.1	10.819	1.802	≤15
TSS	2	32	11.28	5.342	≤50
NH_4	0.05	0.17	0.089	0.019	≤0.3
NO_2	0.01	0.04	0.012	0.005	≤0.05
NO_3	0.27	1.2	0.608	0.194	≤5
TN	0	4.2	0.815	0.999	Not Applicable
TP	0	0.33	0.111	0.071	Not Applicable
$P.PO_4^3$	0	0.3	0.064	0.049	≤0.2
Coliform	3	24,000	429.938	1938.732	≤5000

With regard to heavy metals, Mercury, Cadmium, and Lead were not detected, while Zinc and Manganese records were within the allowable ranges. Arsenic, Chrome VI, and Iron, however,

considerably exceeded the respective regulations. Table 3 summarizes the descriptive statistics of heavy metal concentrations at six sampling stations from 2015 to 2016. The results are more like exploratory evaluations so could not facilitate consolidated claims owing to the limited data available. Our findings nonetheless constitute warning notices regarding the potential contamination of water resources with hazardous heavy metals within the research area.

Table 3. Descriptive statistics of heavy metal concentrations at different sampling stations.

Parameters	QCVN 08-MT:2015/BTNMT		Site C1	Site C11	Site C14	Site SA2	Site SA4	Site R19
Arsenic (mg/L)	≤0.02	Range	0–2.5	0–3.1	0–2.6	0–4	1.8–12	0–3.5
		Mean	0.417	0.52	0.437	1.6	4.317	1.25
		S.D.	1.02	1.264	1.06	1.367	3.837	1.184
Mercury (mg/L)	≤0.001	UNDETECTED						
Cadmium (mg/L)	≤0.005							
Lead (mg/L)	≤0.02							
Zinc (mg/L)	≤1.0	Range	0–0.02	0–0.03	0–0.02	0–0.05	0–0.06	0–0.03
		Mean	0.007	0.012	0.01	0.013	0.023	0.01
		S.D.	0.103	0.013	0.011	0.02	0.023	0.013
Manganese (mg/L)	≤0.2	Range	0–0.04	0.02–0.09	0.02–0.05	0–0.07	0–0.14	0–0.02
		Mean	0.023	0.042	0.03	0.022	0.058	0.01
		S.D.	0.014	0.026	0.011	0.026	0.054	0.011
Chrome VI (mg/L)	≤0.02	Range	0–0.05	0–0.04	0–0.05	0–0.09	0–0.05	0–0.05
		Mean	0.0083	0.006	0.008	0.015	0.008	0.008
		S.D.	0.02	0.016	0.021	0.037	0.021	0.021
Nickel (mg/L)	≤0.1	UNDETECTED						
Iron (mg/L)	≤1.0	Range	0.26–1.61	0.25–2.22	0.27–1.3	0.26–1.33	0.28–5.33	0–0.52
		Mean	0.731	1.13	0.802	0.673	1.57	0.33
		S.D.	0.470	0.867	0.369	0.381	1.9	0.211

3.2. Spatial Variation

Regarding the spatial variation, the performance statistics of different techniques are represented in Tables 4 and 5. A model can be considered a good classifier when it achieves the smallest error values, including MAE, RMSE, RAE, and RRSE, as well as a bigger value of correctly classified samples. For calculating the correct classification rate, take DT (J48) as an example, where DT (J48) was able to accurately classify 137 out of 137 for the "reservoir", seven out of 20 for "surrounding areas," and 77 out of 101 for "canal." Therefore, the DT-based classifier provides an 85.66% accuracy. Figure 5 shows the decision tree for spatial variation derived from J48 method. According to Table 4, RBF achieved the highest correct classification rate of 86.82%, followed by DT, MLP, Naïve Bayes, and SVM. Other than that, RBF obtained the highest performance according to two out of the total four evaluation criteria. Figure 6 represents the three-class confusion matrixes obtained from the different techniques. There are 258 samples in total, including 138 collected from the reservoir, 30 from surrounding areas, and 90 from canals. Among the investigated techniques, MLP gives quite good results: all the samples from the reservoir are correctly classified. However, observing all the confusion matrixes of the three original classes, the majority of the samples from surrounding areas are classified as reservoirs or canals.

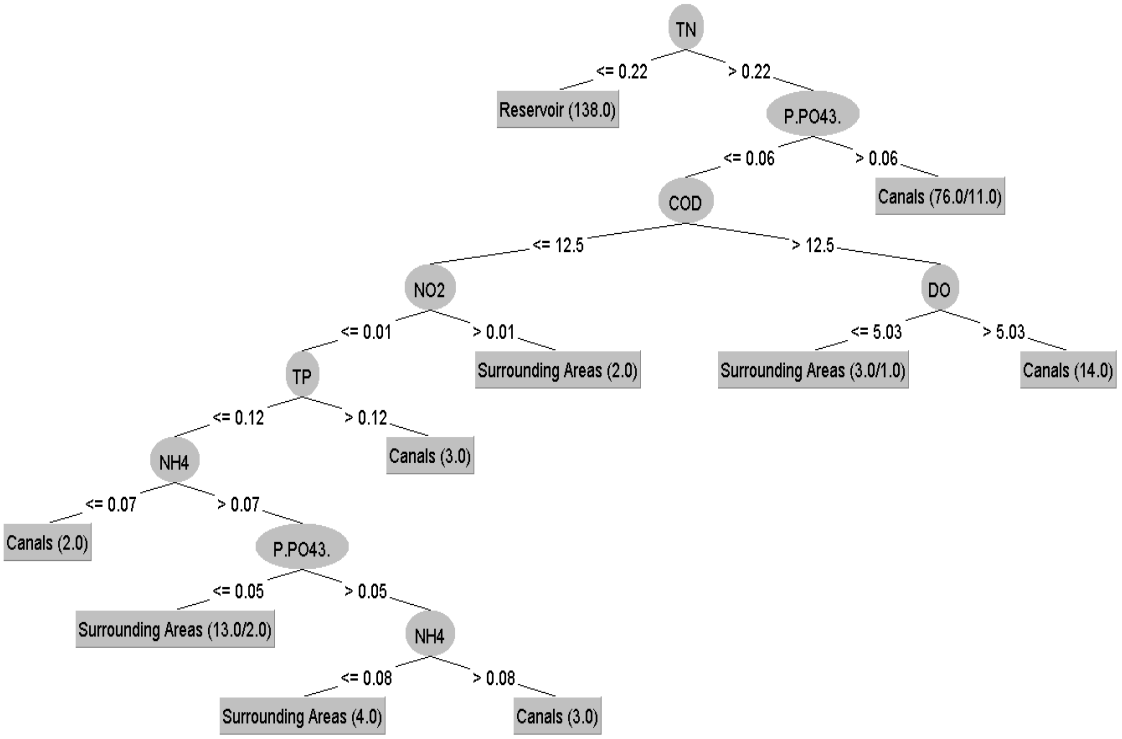

Figure 5. Visualization of the decision tree for spatial variation.

Table 4. The percentage of accurate and inaccurate classification for spatial variation.

Model	Correctly Classified Samples	Incorrectly Classified Samples
DT (J48)	221 (85.66%)	37 (14.34%)
MLP	217 (84.11%)	41 (15.89%)
Naïve Bayes	206 (79.84%)	52 (20.16%)
RBF	224 (86.82%)	34 (13.18%)
SVM	199 (77.13%)	59 (22.87%)

Table 5. Performance statistics for spatial variation.

Model	MAE	RMSE	RAE	RRSE
DT (J48)	0.11	0.29	27.64%	67.23%
MLP	0.11	0.28	29.16%	64.92%
Naïve Bayes	0.13	0.30	34.50 %	68.57%
RBF	0.20	0.28	52.98%	63.97%
SVM	0.15	0.39	39.44%	88.90%

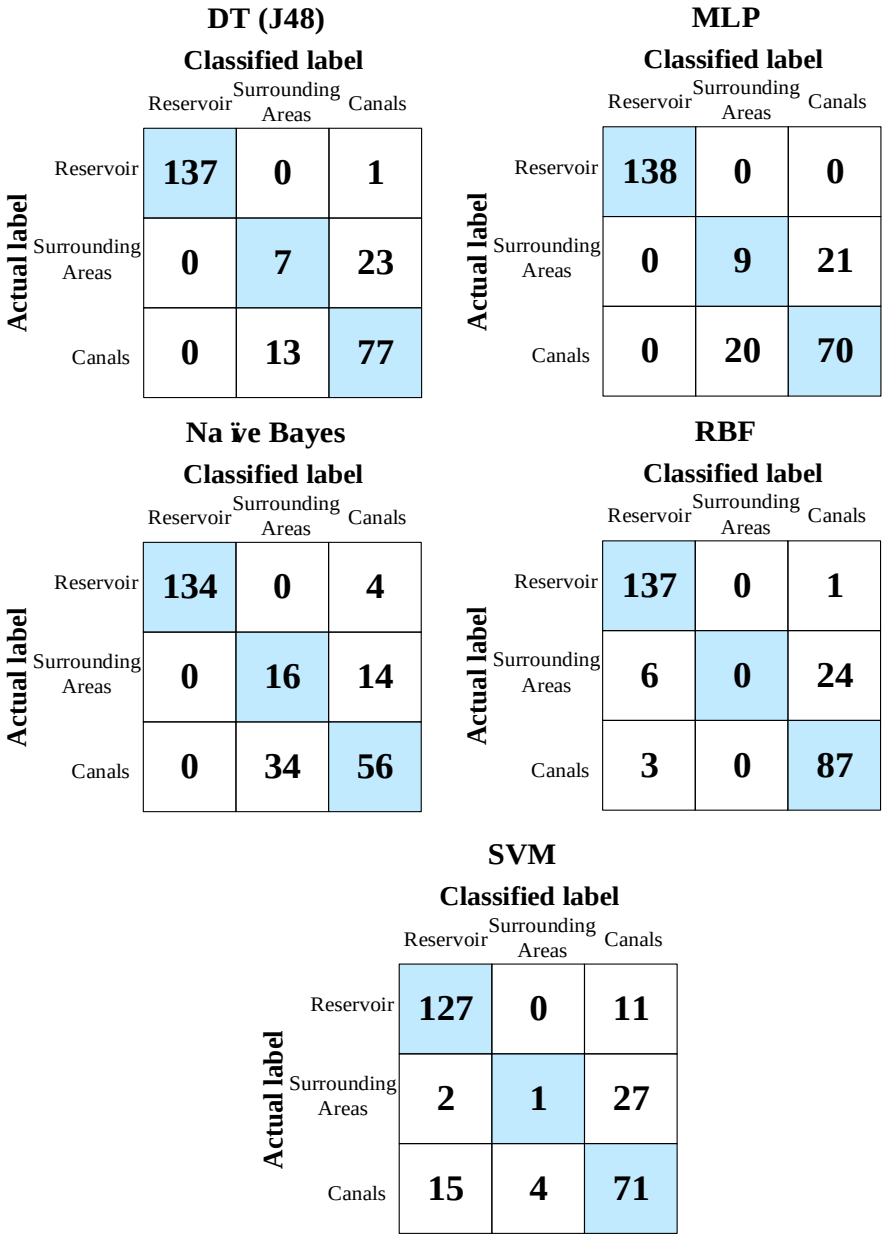

Figure 6. Obtained confusion matrixes for spatial variation.

3.3. Temporal Variation

For temporal variation, Tables 6 and 7 show the performance statistics of different techniques. Figure 7 shows the decision tree for temporal variation obtained from J48 method. According to Table 6, the DT, MLP, and Naïve Bayes together achieved the highest performance, followed by RBF and SVM. On the other hand, Table 7 reveals that MLP outperformed the other techniques, according to all four criteria. The two-class confusion matrixes in Figure 8 showed that except for SVM, the others obtained quite good results.

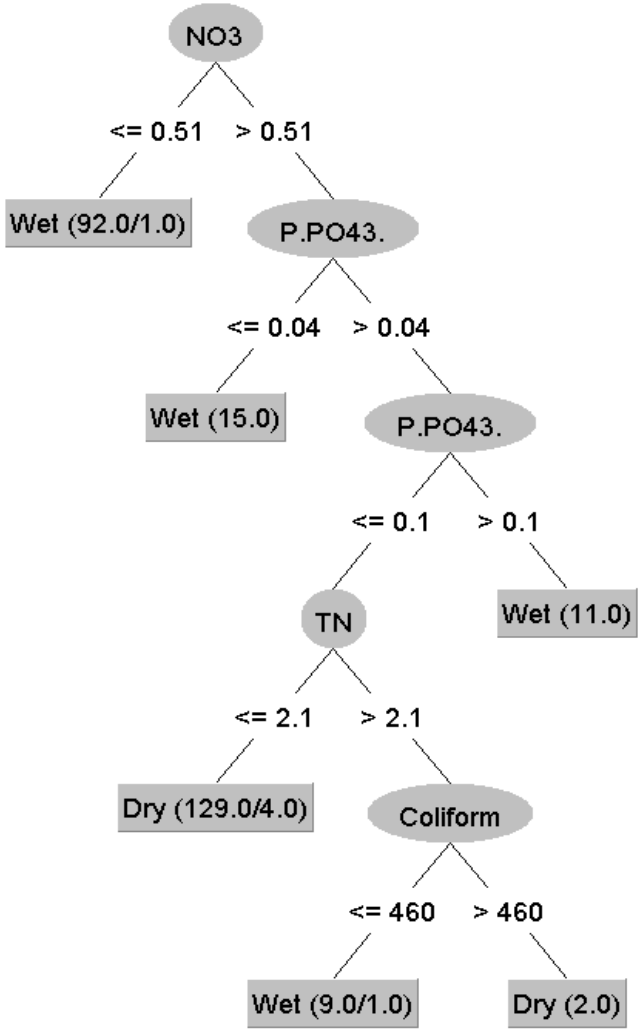

Figure 7. Visualization of the decision tree for temporal variation.

Table 6. The percentage of accurate and inaccurate classification for temporal variation.

Model	Correctly Classified Samples	Incorrectly Classified Samples
DT (J48)	244 (94.57%)	14 (5.43%)
MLP	244 (94.57%)	14 (5.43%)
Naïve Bayes	244 (94.57%)	14 (5.43%)
RBF	243 (94.19%)	15 (5.81%)
SVM	198 (76.74%)	60 (23.25%)

Table 7. Performance statistics for temporal variation.

Model	MAE	RMSE	RAE	RRSE
DT (J48)	0.069	0.23	13.68%	45.69%
MLP	0.062	0.2	12.33%	41.41%
Naïve Bayes	0.085	0.22	16.98%	44.49%
RBF	0.133	0.22	26.49%	43.42%
SVM	0.233	0.48	46.51%	96.44%

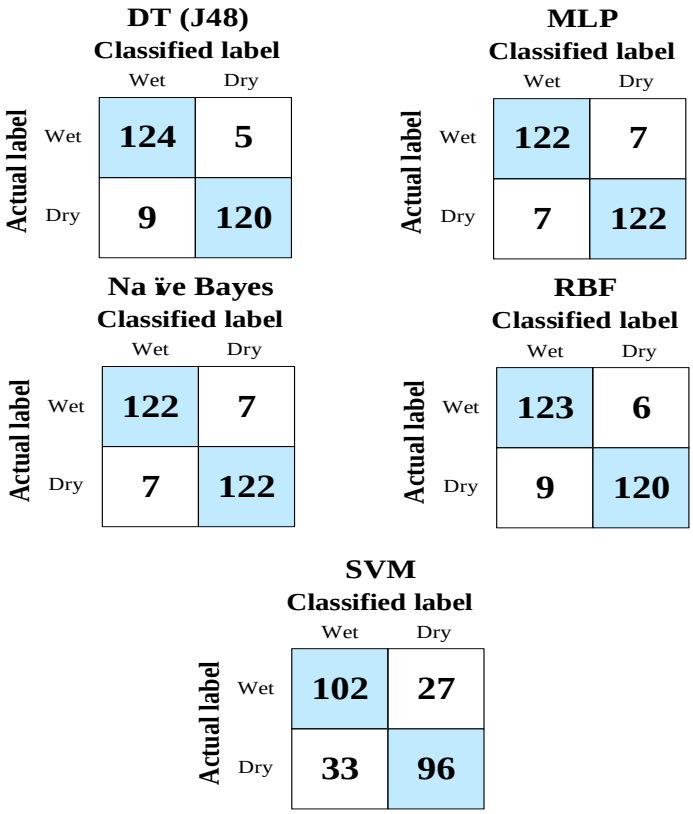

Figure 8. Obtained confusion matrixes for temporal variation.

3.4. Methodological Implications

The positive implications of AI techniques in environmental research are manifold and have been explored by scientists in various disciplines. Our work further expands relevant insights from this literature with the applicability of five of the most commonly used techniques for a relatively small data set.

More specifically, in comparison with the multiple years of automatic monitoring data from other studies, our data set is relatively limited. It is, however, a meaningful contribution to the empirical reference of the water quality of the Song Quao catchment, which is only measured four times annually. Against this limit, the presented methods nonetheless proved to be useful in revealing the structural bundles of not only the observed variables, but also the sampling locations. Our findings have also captured and visualized the seasonal variations of relevant water quality parameters.

3.5. Water Quality Management

Since the first Vietnamese environmental Law was released in 1993, an environmental monitoring network has been established over Vietnam at both national and local levels. However, the allocated budget for most of the monitoring program is still below demand [36]. The monitoring program is usually limited in locations, frequency, and parameters to be measured. The existing river monitoring program at Quao river only collects data four times per year at two locations, which is much less than the data obtained by this study. Results from this study provide a baseline of water quality status for the river basin. This study suggests that, in order to effectively manage water quality in the catchment for a safe domestic water supply, the province should consider three main solutions. First, adding at least one water quality monitoring station (monthly frequency) in the upper catchment of the Quao reservoir, especially during rainy the season, e.g., at SA1, is urgently needed. Second, controlling pollution sources in the downstream area of the Quao canal (near C14) is also of particular

importance. Finally, an automatic water quality monitoring station should be installed near the intake of the domestic water supply system as part of developing a water safety warning system.

4. Conclusions

The protection of the Song Quao reservoir and channel water environment is challenging yet urgent, with a high multidisciplinary and interregional dimension. Therefore, the temporal and spatial analysis and assessments of water quality as shown in this paper essentially facilitate a clear view of the current state of water quality of the SQ-CG water supply system. More specifically, the water resource is likely prone to heavy metals related to pollution, especially Arsenic. Future intensive investigations are an important research need in terms of understanding the water quality status. These findings constitute important baseline knowledge to support the implementation of water management initiatives to protect the reservoir and channel water quality for domestic water supply.

As a methodological contribution, this paper has presented a practical AI-based workflow to explore the temporal and spatial variations of water quality. The incorporated AI techniques include the decision tree (DT), multilayer perceptron (MLP) network, Naïve Bayes, radial basis function (RBF) network, and support vector machine (SVM). With the relatively limited data set, these techniques have successfully uncovered and visualized the data structure and facilitated meaningful references on temporal and spatial variations of river water quality across the study area. This is of particular importance for future studies on river quality monitoring and assessment at sparsely-gauged catchments.

Author Contributions: Investigation, N.H.Q., L.V.T.; Data collection, L.V.T., N.H.Q., N.T.T.D; methodology, Q.H.D., H.H.L, N.H.Q., N.T.T.D.; writing, N.H.Q., H.H.L, Q.H.D., N.T.T.D., T.D.D., H.D.N.; Review and editing, N.D.H, H.H.L, N.H.Q.

Funding: This research was initially funded by the Department of Science and Technology of Binh Thuan province; and finally funded by Nguyen Tat Thanh University.

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

References

1. Goncharuk, V.V.; Pshinko, G.N.; Rudenko, A.V.; Pleteneva, T.V.; Syroeshkin, A.V.; Uspenskaya, E.V.; Zlatskiy, I.A. Genetically Safe Drinking Water. Requirements and Methods of Its Quality Control. *J. Water Chem. Technol.* **2018**, *40*, 16–20. [CrossRef]
2. Tabari, H.; Hosseinzadeh, T.P. Reconstruction of river water quality missing data using artificial neural networks. *Water Qual. Res. J. Can.* **2015**, *5*, 326–335. [CrossRef]
3. Palani, S.; Liong, S.-Y.; Tkalich, P. An ANN application for water quality forecasting. *Mar. Pollut. Bull.* **2008**, *56*, 1586–1597. [CrossRef] [PubMed]
4. Swain, R.; Sahoo, B. Improving river water quality monitoring using satellite data products and a genetic algorithm processing approach. *Sustainabil. Water Qual. Ecol.* **2007**, *9–10*, 88–114. [CrossRef]
5. Jolk, C.; Greassidis, S.; Jaschinski, S.; Stolpe, H.; Zindler, B. Planning and Decision Support Tools for the Integrated Water Resources Management in Vietnam. *Water* **2011**, 711–725. [CrossRef]
6. Quan, N.H.; Meon, G. Nutrient Dynamics during Flood Events in Tropical Catchments: A Case Study in Southern Vietnam. *Clean Soil Air Water* **2014**, *43*, 652–661. [CrossRef]
7. Meon, G.; Pätsch, M.; Phuoc, N.V.; Quan, N.H. EWATEC-COAST: Technologies for Environmental and Water Protection of Coastal Zones in Vietnam. In Proceedings of the 4th International Conference for Environment and Natural Resources—ICENR, Göttingen, Germany, 17–18 June 2014.
8. Keen, B.; Chu, T.H.; Slavich, P.; Bell, R.; Hoang, M.T. *Opportunities to Improve the Sustainable Utilisation and Management of Water and Soil Resources for Coastal Agriculture in Vietnam and Australia*; Report FR2013-12; Australian Centre for International Agricultural Research (ACIAR): Canberra, Australia, 2013; ISBN 978-1-922137-60-9.

9. Binh Thuan Department of Agriculture and Rural Development. *Irrigation Development Plan of Binh Thuan Province 2011–2020*; Binh Thuan Department of Agriculture and Rural Development: Bình Thuận, Vietnam, 2011.
10. Mirauda, D.; Ostoich, M. Assessment of Pressure Sources and Water Body Resilience: An Integrated Approach for Action Planning in a Polluted River Basin. *Int. J. Environ. Res. Public Health* **2018**, *15*, 390. [CrossRef]
11. Mirauda, D.; Ostoich, M.; Di Maria, F.; Benacchio, S.; Saccardo, I. Integrity Model Application: A Quality Support System for Decision-makers on Water Quality Assessment and Improvement. *IOP Con. Ser. Earth Environ. Sci.* **2018**, *120*. [CrossRef]
12. Kulishenko, A.E.; Ostapenko, V.T.; Kravchenko, T.B.; Kvasnitsa, E.A.; Ostapenko, R.V. The statistical analysis of quality indicators of the Dnieper river water and directions for reconstruction of water treatment facilities of the Dnieper waterworks in Kiev. *J Water Chem. Technol.* **2011**, *3*, 117. [CrossRef]
13. Mishra, A. Assessment of water quality using principal component analysis: A case study of the river Ganges. *J Water Chem. Technol.* **2010**, *32*, 227–234. [CrossRef]
14. Nnaji, C.C.; Agunwamba, J.C. The environmental impact of crude oil formation water: A multivariate approach. *J Water Chem. Technol.* **2013**, *35*, 222–232. [CrossRef]
15. Wu, W.; Dandy, G.; Maier, H. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environ. Modell. Softw.* **2014**, *54*, 108–127. [CrossRef]
16. Arain, M.B.; Ullah, I.; Niaz, A.; Shah, N.; Shah, A.; Hussain, Z.; Muhammad, T.; Hassan, I.A.; Jameel, A.B.; Tasneem, K. Evaluation of water quality parameters in drinking water of district Bannu, Pakistan: Multivariate study. *Sustainabil. Water Qual. Ecol.* **2014**, *3–4*, 114–123. [CrossRef]
17. Kikuchi, T.; Furuichi, T.; Hai, H.T.; Tanaka, S. Assessment of Heavy Metal Pollution in River Water of Hanoi Using Multivariate Analyses. *Bull. Environ. Contam. Toxicol.* **2009**, *85*, 575–582. [CrossRef] [PubMed]
18. Loc, H.H.; Hong Diep, N.T.; Can, N.T.; Irvine, K.N.; Shimizu, Y. Integrated evaluation of Ecosystem Services in Prawn-Rice rotational crops, Vietnam. *Ecosyst. Serv.* **2017**, *26*, 377–387. [CrossRef]
19. Shrestha, S.; Fazama, F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environ. Modell. Softw.* **2007**, *22*, 464–475. [CrossRef]
20. Funahashi, K. On the approximate realization of continuous mappings by neural networks. *Neural Netw.* **1989**, *2*, 183–192. [CrossRef]
21. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feed-forward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
22. Norgaard, M.R.; Poulsen, N.K.; Hansen, L.K. *Neural Networks for Modelling and Control of Dynamic Systems*; A Practitioner's Handbook; Springer: London, UK, 2000.
23. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. 2001. Available online: https://papers.nips.cc/paper/1895-overfitting-in-neural-nets-backpropagation-conjugate-gradient-and-early-stopping.pdf (accessed on 25 October 2018).
24. Jiang, J.; Chao, D.; Chen, H. Boundary value problems for fractional differential equation with causal operators. *Appl. Math. Nonlinear Sci.* **2016**, *1*, 11–22. [CrossRef]
25. Batool, F.; Adeel, S.; Azeem, M.; Khan, A.A.; Bhatti, I.A.; Ghaffar, A.; Iqbal, N. Gamma radiations induced improvement in dyeing properties and colorfastness of cotton fabrics dyed with chicken gizzard leaves extracts. *Radiat. Phys. Chem.* **2013**, *89*, 33–37. [CrossRef]
26. Guan, X.; Zhu, Y.; Song, W. Application of RBF neural network improved by peak density function in intelligent color matching of wood dyeing. *Chaos, Solitons Fractals* **2016**, *89*, 485–490. [CrossRef]
27. Myles, A.J.; Feudale, R.N.; Liu, Y.; Woody, N.A.; Brown, S.D. An introduction to decision tree modeling. *J. Chemomet. Soc.* **2004**, *18*, 275–285. [CrossRef]
28. Debeljak, M.; Džeroski, S. Decision trees in ecological modelling. In *Modelling Complex Ecological Dynamics*; Springer: Heidelberg, Germany, 2011; pp. 197–209.
29. Abe, S. *Support Vector Machines for Pattern Classification*; Springer: London, UK, 2005.
30. Han, J.; Jian, P.; Micheline, K. *Data Mining: Concepts and Techniques*; Elsevier: New York, NY, USA, 2011.
31. International Standard ISO Document 5667—1: Guidance on the Design of Sampling Programmes and Sampling Techniques. Available online: https://www.iso.org/obp/ui/#iso:std:iso:5667:-1:ed-2:v1:en (accessed on 15 December 2018).

32. International Standard ISO Document 5667—3: Preservation and Handling of Water Samples. Available online: https://www.iso.org/obp/ui/#iso:std:iso:5667:-1:ed-2:v1:en (accessed on 15 December 2018).
33. International Standard ISO Document 5667—4: Guidance on Sampling from Lakes, Natural and Man-made. Available online: https://www.iso.org/obp/ui/#iso:std:iso:5667:-1:ed-2:v1:en (accessed on 15 December 2018).
34. International Standard ISO Document 5667—1: Guidance on Sampling of Rivers and Streams. Available online: https://www.iso.org/obp/ui/#iso:std:iso:5667:-1:ed-2:v1:en (accessed on 15 December 2018).
35. Ministry of Natural Resources and Environment, QCVN 08 –MT: 2015. National Technical Regulation on Surface Water Quality. Available online: http://moitruong.com.vn/Upload/48/Nam_2017/Thang_3/Ngay_17/QCVN08-2015_Quy_chuan_ky_thuat_quoc_gia_ve_chat_luong_nuoc_mat.pdf (accessed on 15 December 2018).
36. Hoang, D.T. *Environmental Monitoring in Vietnam: Current Status and Perspective*; Vietnam Environment Administration: Nanoi, Vietnam, 2011.