Article

Predicting Microbial Species in a River Based on Physicochemical Properties by Bio-Inspired Metaheuristic Optimized Machine Learning

Jui-Sheng Chou ^{1,*} , Chang-Ping Yu ², Dinh-Nhat Truong ¹ , Billy Susilo ¹, Anyi Hu ³ and Qian Sun ³

¹ Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan; d10605806@mail.ntust.edu.tw (D.-N.T.); susilobilly@yahoo.com (B.S.)

² Graduate Institute of Environmental Engineering, National Taiwan University, Taipei 10617, Taiwan; cpyu@ntu.edu.tw

³ CAS Key Laboratory of Urban Pollutant Conversion, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; ayhu@iue.ac.cn (A.H.); qsun@iue.ac.cn (Q.S.)

* Correspondence: jschou@mail.ntust.edu.tw

Received: 23 October 2019; Accepted: 2 December 2019; Published: 4 December 2019

Abstract: The main goal of the analysis of microbial ecology is to understand the relationship between Earth's microbial community and their functions in the environment. This paper presents a proof-of-concept research to develop a bioclimatic modeling approach that leverages artificial intelligence techniques to identify the microbial species in a river as a function of physicochemical parameters. Feature reduction and selection are both utilized in the data preprocessing owing to the scarce of available data points collected and missing values of physicochemical attributes from a river in Southeast China. A bio-inspired metaheuristic optimized machine learner, which supports the adjustment to the multiple-output prediction form, is used in bioclimatic modeling. The accuracy of prediction and applicability of the model can help microbiologists and ecologists in quantifying the predicted microbial species for further experimental planning with minimal expenditure, which is become one of the most serious issues when facing dramatic changes of environmental conditions caused by global warming. This work demonstrates a neoteric approach for potential use in predicting preliminary microbial structures in the environment.

Keywords: microbial community; physicochemical properties; bioclimatic modeling; river environment; multi-output prediction; bio-inspired metaheuristics; optimization; machine learning; data mining

1. Introduction

Microorganisms play an important role in mediating global biochemical cycling. Understanding the diversity and composition of a microbial community in a particular environment and its controlling factors is a critical goal of the analysis of microbial ecology [1–3]. Microbial ecology is the study of the interactions of microorganisms with their environment, each other, and plant and animal species [4–6]. It also includes the study of biogeochemical cycles, symbioses, and the interaction of microbes with anthropogenic phenomena such as climate change and pollution.

Microorganisms are the smallest living organisms on Earth, but they are also the most abundant as they occupy the entire biosphere. They are also the most diverse; the majority of them are unknown to scientists [7]. They can be found in every macro- or microenvironment from the surface and depths of the ocean to the skin and digestive systems of humans and animals. They are found in water, soil, and the human gut [8]—they are literally everywhere.

Many conventional investigations have sought to classify microbial communities. The newest method for so doing is DNA sequencing, which is costly and time-consuming [9]. DNA sequencing is a laboratory method for determining the sequence of a DNA molecule [10]. Physicochemical properties can determine the presence of microorganisms and microbial communities. Physicochemical parameters are simple numerical values that provide insight into the state of the environment; they include pH, temperature, Dissolved Oxygen (DO), and others. Indicators are developed based on quantitative measurements or statistical data of environmental conditions that are tracked over time.

Recently, artificial intelligence (AI) is being used to predict microbial communities in the environment [2]. Artificial intelligence (AI) enables machines to learn from experience, adjust to new inputs, and perform human-like tasks [11]. It is capable of advanced learning using large complex datasets, including microbial datasets. AI-based approaches have advantages over traditional deterministic methods when applied to microbial datasets as they eliminate the complexity that is associated with many factors with the DNA sequencing process. One AI approach has been confirmed to be useful in predicting microbial community assemblages based on physicochemical parameters [2].

This paper develops a hybrid multi-output model for predicting various microbial communities, which integrates an optimization algorithm, called “adjusted particle swarm optimization”, to least squares support vector regression, which is a form of support vector machine; it is developed in MATLAB. This predictive system can solve a multiple-output problem, since most of the predictions about a microbial community involve a single output, with physicochemical parameters as inputs, rather than DNA sequencing data as inputs, because DNA sequencing is expensive and time-consuming. The model enables researchers and environmental scientists who use AI for environmental purposes to predict future responses of microbial communities to various environmental scenarios [12].

Most multi-input single-output models are trained independently to get the best prediction model without considering the relations between the outputs. In reality, outputs may well be mutually related. The hybrid model can be advantageous since this hybrid model can predict all outputs simultaneously, without neglecting those relationships among outputs. A multi-output prediction model learns a mapping from a multivariate input feature space to a multivariate output space [13].

Based on physicochemical properties of a river, this work proposes a predictive model that has various microbial types in a river as outputs. The model runs the outputs simultaneously, so the relationships between pairs of outputs cannot be neglected. Physicochemical parameters can easily be predicted from baseline environmental conditions, but only modeling by artificial intelligence can predict microbial communities based on present environmental conditions or even future environmental conditions. The presented model can help microbiologists and ecological researchers to plan future functions of microbial ecology in a river for sustainable watershed management when facing dramatic changes of environmental conditions caused by global warming scenario. This work provides a neoteric approach for use in predicting the microbial structure in environment.

2. Literature Review

2.1. Microbial Community in a River

Rivers are main components of the hydrological cycle and have a fundamental function in the ecosystem. They support human health, agricultural production, and industry because they are distributed widely throughout the landscape and provide large volumes of water [14]. For a long time, rivers have been recognized as important for transferring nutrients from the land to coastal areas, and recently ecologists have recognized that the ecosystem in the river has an essential role in both regional and global biogeochemical cycles.

A river can be an ideal caretaker of environmental changes in terrestrial and atmospheric processes. Most obviously, the water in a river is an important resource, supporting biological processes and as a habitat for aquatic species [15]. Microbes in a river have significant roles in intervening in, and managing, carbon and nutrient fluxes, and in removing contaminants [16,17]. Microbes are the

predominant, and most varied, organisms in river ecosystems. Accordingly, microbial communities in water have diverse and dynamic compositions under the influence of various environmental factors [18]. Recent studies have demonstrated that, as a biological index of the state of an ecosystem, bacterial population might be a better choice than conventional biological indices, such as the populations of macroinvertebrates, fish, and birds [19].

The composition of microbial communities may have significant impact on the functioning of an ecosystem. Bacteria do not always negatively affect the environment; on the contrary, bacteria, especially in rivers, are advantageous to many ecosystem-related processes. Bacteria are the only living things that can fix nitrogen [20], which is a principal nutrient for all living organisms [21]. Nitrogen is a necessary component of many biomolecules, including proteins, DNA, and chlorophyll.

Bacteria have enzyme systems that fix nitrogen from the atmosphere, converting it directly to nitrogen compounds that can be used by plants. Furthermore, bacteria are fundamental to the decomposition of dead organic material into forms that can be used by other organisms [22]; this is the main reason the involved bacteria enzyme systems are regarded as key drivers of the Earth's biogeochemical cycles. The cycles of nitrogen and sulfur in the atmosphere are also associated with the activities of bacteria [23], which exclusively transform under anaerobic conditions. The utilization of inorganic nitrogen and sulfur for biosynthesis is limited to microorganisms and plants [24].

In addition to performing numerous functions, bacteria have major disadvantages for our environment. The greatest risks posed by bacteria are associated with the ingestion of water that is contaminated with human or animal feces. Wastewater discharges in fresh water and coastal seawater are major sources of fecal microorganisms, including pathogens [25]. Some bacteria can make water a vehicle of disease [26]. Water-borne diseases are the target of the greatest concern about the quality of water, and relevant pathogens include a wide range of viruses, bacteria, and protozoan parasites. A pathogen is defined as a microorganism that causes, or can cause, disease damage in a host [27]. In short, some bacteria can harm living creatures in water.

2.2. Physicochemical and Deoxyribose Nucleic acid Sequencing Factors

Physicochemical factors (physical and chemical conditions) are abiotic factors that influence the environment. Hydrogen ion activity (pH), temperature, dissolved oxygen (DO), and ammonium ($\text{NH}_4\text{-N}$) are examples of physicochemical factors [28]. Numerous previous investigations have established that the bacterial communities in freshwater is commonly strongly correlated with physicochemical factors, including temperature, pH, DO, and nutrients [24]. Physicochemical factors can be used to predict the microorganisms, and especially bacteria. In short, physicochemical factors are simple numerical values that provide insight into the state of the environment and related indicators are developed based on quantitative or statistical measurements of environmental conditions that are tracked over time.

Specifying the order of nucleic acid sequences in biological samples is an initial, critical process in a wide variety of research applications [10]. Deoxyribose nucleic acid (DNA) sequencing is one of the reliable methods, which is a process that is carried out in the laboratory to determine the precise order of nucleotides within a DNA molecule [29], which consists of the nitrogen-containing nucleobases cytosine (C), guanine (G), adenine (A), and thymine (T) [30]. The type and order of these nucleobases in a DNA molecule carry a large amount of genetic and biological information with respect to “who we are” or “what species we are” [31]. In short, the DNA sequencing process is crucial because knowledge of DNA sequences has become indispensable for basic biological research, and several applied fields, such as biotechnology, biological systematics, etc.

Next-generation high-throughput DNA sequencing techniques are opening up fascinating opportunities in the life sciences, and Illumina DNA sequencing is one of the most popular. The Illumina sequencing platform was introduced in 2006, and Illumina acquired Solexa in early 2007 [32]. Illumina sequencers have huge advantages over other DNA sequencing methods since they provide a widely used platform for the parallel readout of several hundred million immobilized sequences, using fluorescent-dye reversible-terminator chemistry [33]. In many sequencing applications, Illumina DNA sequencing would be the method of choice if not technically unfeasible or feasible but prohibitively expensive [34].

2.3. Current Use of Artificial Intelligence to Predict Microbial Community

In ecological modeling, methods that are based on ecological informatics should be used when the modeled system has many variables, when some of those variables are not precisely accounted for, when they are categorical or nominal, or when nonlinear effects and/or interactions between variables are suspected. Generally, ecological informatics can have a relevant role when theory does not suffice to explain the dynamics of a system or the relationships among its components.

Ecological informatics includes several methods that are particularly useful in empirical modeling applications, which are used when one or more variables are expensive and time-consuming to measure or when information can be accurately estimated using other variables that are cheaper and easier to measure. Artificial intelligence methods have advantages over traditional statistical techniques such as those that involve linear and nonlinear patterns since AI can deal with most of the characteristics that are typical of ecological data such as unusual distributions, non-linearity, multiple missing values, complex data interactions, and dependence of the observations.

Numerous physicochemical parameters of water, such as temperature, DO, nitrate, ammonium, silicate, chlorophyll, labile dissolved organic carbon (LDOC), and others, are already used to predict microbial communities using multiple methods that are based on artificial neural networks (ANNs), which are commonly used in AI techniques. Larsen et al. predicted bacterial community assemblages using an artificial neural network approach [2].

Wu et al. [35] rapidly predicted bacterial heterotrophic fluxomics by machine learning using methods that are based on support vector regression (SVM), k-nearest neighbors (k-NN), and decision trees. They discovered the relationship between environmental and genetic factors and metabolic fluxes that are hidden in fluxomic data, to generate predictive models that can significantly accelerate flux quantification. These achievements raise the expectation of artificial intelligence (AI) as a powerful but simple alternative tool for predicting bacterial communities, since DNA sequencing, even though it is the ideal tool for obtaining information about a bacterial community, is time-consuming and expensive.

3. Methodology

In this investigation, dimensionality reduction and the handling of missing values are performed in the preprocessing of data before a multiple-output prediction model is constructed. Dimensionality reduction is conducted by combining three techniques: feature subset selection, feature creation, and expert engineering judgement. Various forms of machine learning are utilized to predict missing values. After the data are preprocessed, the multiple-output hybrid predictive model is built. Figure 1 presents the development of the hybrid multiple-output prediction model.

Figure 1. Development of the hybrid multi-output prediction model.

3.1. Dimensionality Reduction and Handling of Missing Data

The first operation is the preprocessing of data, which is critical in machine learning. The main goal of preprocessing data is to improve the quality of data to make the data ready for processed in the next step, substantially improving the overall quality of the identified patterns and/or the time that is required to perform the data mining [36]. In short, the two main reasons for preprocessing data are: (i) to solve problems inside the data; and (ii) to prepare the data for analysis [37]. Dimensionality reduction and the handling of missing value are two common data preprocesses.

Dimensionality reduction is an important data preprocessing technique, which is used to simplify the structure of data before the fitting of meaningful empirical models; it can make the resulting model more understandable by reducing the number of data attributes. The key is to reduce the dimensionality of predictors while preserving their regression relationship with a response, which is essential for data pattern recognition [13,38].

Most commonly used dimensionality reduction methods fall into one of two categories: feature creation or feature subset selection. Expert engineering judgement is another dimensionality reduction strategy for reducing the number of attributes in a qualitative manner that differs from statistical and mathematical dimensionality reduction [39].

Feature creation generates new attributes that capture important information in a dataset much more efficiently than the original attributes, and have been established to be very effective in real-world dimensionality reduction problems [13]. Feature subset selection is another way to reduce the dimensionality of data that concerns the selection of a subset of features and has been shown to have a positive effect on the performance of machine learning algorithms.

The most well-known feature subset selection method is brute force approach, which tries inputting all possible feature subsets to a data mining algorithm. In summary, feature creation generates more efficient attributes from the original attributes, while feature subset selection picks the most important attributes from the original attributes. This method utilizes stepwise regression in IBM-SPSS Statistics 19 to implement feature subset selection for dimensionality reduction [13].

Values are frequently missing in data that are collected in the natural and social sciences, and this problem is frequently addressed in preprocessing [40]. Various investigations have discussed the analysis and handling of missing values [41,42]. In general, missing data are either ignored in favor of simplicity or replaced with substituted values that are estimated using a statistical method, such as mean values. However, this method uses machine learning to predict missing values, because numerous studies have already proved that artificial intelligence is an effective and powerful tool for predicting the missing value [12]. Therefore, this investigation utilized a baseline and ensemble model with ANN, SVR, LR, and CART as artificial intelligence techniques, implemented in WEKA software, to predict 20 data points whose values are missing based on 40 other data points with values.

For the baseline model, this study uses Li [43] as the corresponding literature on ANN, the literature on SVR [44], literature about LR [45], and CART [46]. The ensemble model in this work combines four AI techniques, ANN + SVR + LR + CART. Chou et al. [47] applied those four techniques into various ensemble methods: voting, bagging, and stacking. Breiman [48] described the development of voting, bagging, and stacking ensemble models.

This work proposes four homogenous ensembles, as depicted in Figure 2, that are generated using the individual learning technique: an ANN ensemble, an SVR ensemble, an LR ensemble, and a CART ensemble. In particular, the bagging ensemble method generates multiple versions of a predictor and uses them to obtain an aggregated predictor [48]. The bagging ensemble method utilizes bootstrapping to train several models independently of each other and with different training sets. Bootstrapping is a statistical estimation technique by which a statistical quantity such as a mean is estimated from multiple random samples of data. It is a useful technique when data are limited and a highly robust statistical quantity is to be estimated. Notably, bagging SVR outperformed all other prediction models, whether baseline or ensemble.

Figure 2. Proposed homogenous ensembles in bagging method.

3.2. Hybrid of Multi-Output Model and Bio-Inspired Metaheuristic Optimization Algorithm

3.2.1. Multi-Output Least Squares Support Vector Regression

The artificial intelligence technique that has recently gained the most attention is least squares support vector regression (LSSVR). LSSVR, which is an alternative to SVM for regression, modifies Vapnik's original SVR formulation for estimating nonlinear functions by solving a linear set of equations rather than solving a time-consuming quadratic programming problem [49]. LSSVR, developed in 1998, has been proven to solve nonlinear function estimation problems. However, LSSVR only works in single-output cases.

Xu et al. [50] developed a multi-output regression approach to perform input-output mapping in multivariate space. The multi-output least square support vector regression (MLSSVR) solves problems with multi-output settings, since the standard formulation of the LSSVR, despite its potential usefulness, cannot cope with multi-output cases. In such cases, LSSVR considers learning and treats each parameter individually and separately, and disregards any relationships among outputs. MLSSVR represents a major breakthrough in multi-output cases because it can learn all of the parameters simultaneously and consider the relationships among outputs. Given a sample $\{(x_i, y^i)\}_{i=1}^l$ with $x_i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}^m$, the multi-output regression is to predict an output vector $y \in \mathbb{R}^m$ from an input vector $x \in \mathbb{R}^d$. The MLS-SVR is briefly presented as below:

Find $w_0 \in \mathbb{R}^{n_h}$, $V = (v_1, v_2, \dots, v_m) \in \mathbb{R}^{n_h \times m}$, and $b = (b_1, b_2, \dots, b_m)^T \in \mathbb{R}^m$ by minimizing the following objective function:

$$\min_{w_0 \in \mathbb{R}^{n_h}, V \in \mathbb{R}^{n_h \times m}, b \in \mathbb{R}^m} \mathcal{F}(w_0, V, \Xi) = \frac{1}{2} (w_0^T w_0) + \frac{1}{2} \frac{\lambda}{m} \text{trace}(V^T V) + \eta \frac{1}{2} \text{trace}(\Xi^T \Xi) \quad (1)$$

$$\text{s.t. } Y = Z^T W + \text{repmat}(b^T, l, 1) + \Xi$$

where $\Xi = (\xi_1, \xi_2, \dots, \xi_c) \in \mathbb{R}^{l \times m}$, $W = (w_0 + v_1, w_0 + v_2, \dots, w_0 + v_m) \in \mathbb{R}^{n_h \times m}$, $Z = (\varphi(x_1), \varphi(x_2), \dots, \varphi(x_l)) \in \mathbb{R}^{n_h \times l}$, and $\lambda, \eta \in \mathbb{R}_+$ are two positive real parameters.

The decision function for the multiple outputs of MLSSVR is as follows.

$$f(x) = \varphi(x)^T W^* + b^{*T} \quad (2)$$

where $\varphi(\mathbf{x})^T$ is the transpose matrix of the mapping $\mathbb{R}^d \leadsto \mathbb{R}^{n_h}$ from d -dimensional to some higher (maybe infinite)-dimensional space with n_h dimensions, called a Hilbert space, while $\mathbb{R}$ denotes the set of real numbers, $\mathbf{W}^*$ is a decision factor that carries information of commonality and specialty, and $\mathbf{b}^{*T}$ denotes the transpose matrix of Bayes, which is consistent with the number of outputs. This concept concerns the underlying relationship in multi-output least squares support vector regression.

The major differences between multi-input multi-output (MIMO) and multi-input single-output (MISO) algorithms are the input–output mapping system, the relations among outputs, the type of parameters, and the generated prediction model. However, whether MIMO or MISO method is used, the hyperparameters involved in the LSSVR must be fine-tuned to obtain optimal predictions. Thus, an effective optimization algorithm is needed to find the efficient solutions.

3.2.2. Accelerated Particle Swarm Optimization

Many researchers have used bio-inspired metaheuristic optimization algorithms, such as genetic algorithms (GA) and particle swarm optimization (PSO), to improve predictive AI models with good results. In the recently developed fields of metaheuristics and soft computing, particle swarm intelligence algorithms are extensively used for optimization and computational intelligence [51]. Particle swarm optimization is a computational method that optimizes a problem by iteratively improving a candidate solution in terms of a given measure of quality. It solves a problem by moving a population of candidate solutions, here dubbed particles, around a search space according to a simple mathematical formula that governs each particle's position and velocity. Yang et al. [52] proved that particle swarm optimization (PSO) algorithm is highly effective for solving various optimization problems.

However, standard particle swarm optimization has some drawbacks for finding global solution capabilities and in its convergence rate; therefore, Yang et al. [52] developed the accelerated particle swarm optimization as a simpler version of PSO with much faster convergence. In the accelerated particle swarm optimization (A-PSO), the velocity vector is generated by the following formula.

$$\mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \beta(\mathbf{g}^* - \mathbf{x}_j^t) + \alpha\epsilon_t \quad (3)$$

where ϵ_t is a random number (0,1) that modifies the second term, which represents individual particle velocity. Particle position is updated as,

$$\mathbf{x}_j^{t+1} = \mathbf{x}_j^t + \mathbf{v}_j^{t+1}. \quad (4)$$

A further development of the accelerated PSO is the reduction of randomness as iterations proceed, using the following decreasing function.

$$\alpha = \alpha_0\gamma^t, \quad (0 < \gamma < 1) \quad (5)$$

Yang et al. [52] recommended $\alpha_0 \approx 0.5 \sim 1$ and $\gamma = 0.9$ to 0.97 as the initial randomness parameter and control parameter, respectively.

In this study, a multi-output least squares support vector regression model with an accelerated particle swarm algorithm to optimize the hyperparameters is implemented in MATLAB to construct a multi-output prediction model, called Multi Input Multi Output Particle Swarm Optimization–Least Squares Support Vector Regression (MIMOPSO-LSSVR). This model is used to predict microbial community, with the physicochemical parameters as inputs. Those hyperparameters are: (1) two regularization parameters (λ , η), which specify the cost trade-off between minimizing the training error and minimizing model complexity; and (2) the kernel parameter (ρ) of the RBF kernel function, which specifies the non-linear mapping from the input space to the high-dimensional feature space. This investigation represents a breakthrough in the field since it is the first to implement multi-output setting to predict microbial community, such that it considers the relationships among outputs.

3.3. Performance Evaluation

The accuracy of the hybrid model for predicting microbial community can only be evaluated by considering how well it performs on new data (as a testing data) that were not used when the model was fitted. Therefore, four performance indices were used to evaluate the performance of the prediction model: are MAAPE (mean absolute arctangent percentage error), MAE (mean average error), R^2 , and SI (synthesis index).

In this hybrid model, MAAPE is preferred to MAPE (mean absolute percentage error) as a performance index for several reasons. Even though MAAPE is the most popular measure of accuracy, it has infinite or undefined values when the actual values of data are zero or close to zero, as commonly arises in some fields, especially for ecological or microbial data. If the actual values are very small (usually less than one), then MAAPE yields extremely large percentage errors (outliers) and zero actual values result in infinite MAAPE; this problem is one of its greatest disadvantages. MAAPE represents a perfect solution to this problem. The equation for MAAPE is as follows.

$$\text{MAAPE} = \frac{1}{n} \sum_{i=1}^n \arctan \left| \frac{y - y'}{y} \right| \quad (6)$$

Equations (6)–(8) are equations for MAE, R^2 , and SI, respectively, as other performance indices of the hybrid prediction model.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y - y'| \quad (7)$$

$$R^2 = \left(\frac{n \sum y \cdot y' - (\sum y)(\sum y')}{\sqrt{n(\sum y^2) - (\sum y)^2} \sqrt{n(\sum y'^2) - (\sum y')^2}} \right)^2 \quad (8)$$

$$\text{SI} = \frac{1}{k} \sum_{i=1}^k \left(\frac{P_i - P_{\min i}}{P_{\max i} - P_{\min i}} \right) \quad (9)$$

where y' is the predicted value; y is the actual value; and n is the number of data samples. In SI, k denotes the number of performance measures, and $P_i = i^{\text{th}}$ performance measure. The range of SI is from zero to one; an SI value close to zero indicates a highly accurate predictive model.

4. Model Development

4.1. Data Collection

This study involved 60 data points that were directly obtained from Jiu Long River (Figure 3), Southeast China, during 5–6 September 2012, 15–17 January 2012, and 7–9 June 2013, during the normal season (transition between dry and wet season), dry, and wet seasons, respectively. Table 1 presents the environmental conditions under which the data were collected. Jiu Long River is divided into two main parts, the North River and the West River; it is the largest river in Southern Fujian and the second largest in Fujian province. This 258-km-long river flows into the Taiwan Strait. Figure 3 displays the geographical location of data collection in Jiu Long River.

Figure 3. Geographical locations of data collecting sites.

Table 1. Environmental Condition of Data Collecting Period.

Data Collecting Period	Season	Average Temperature (°C)	Average Precipitation (mm)
5–6 September 2012	Normal	26	141
15–17 January 2012	Dry	13	34
7–9 June 2013	Wet	26	187

N1–N11 represent the data collecting sites in the North River and W1–W9 denote the data collecting sites in the West River. Firstly, 60 data points were obtained, with 54 physicochemical parameters inputs (Table 2) and 81 microbial communities' outputs (Table 3), respectively. The data were preprocessed to reduce the numbers of input and output attributes to eight and seven, respectively. Table 4 presents the statistical parameters of the microbial datasets.

Table 2. Original input attributes.

No.	Attribute	No.	Attribute	No.	Attribute	No.	Attribute
1	Temperature	15	ATP	29	GBL	43	OMC
2	pH	16	BP3	30	IDT	44	PPB
3	DO	17	BPA	31	IPF	45	PPL
4	DO%	18	BPB	32	KPF	46	PPZ
5	Spc	19	CAF	33	LOR	47	SDZ
6	NH4-N	20	CDI	34	LOS	48	SFZ
7	NO2-N	21	CPP	35	MBB	49	SIL
8	NO3-N	22	CRO	36	MEF	50	SMT
9	DIN	23	CZP	37	MNZ	51	SMZ
10	TP	24	DAN	38	MPB	52	TBD
11	PO4-P	25	DFA	39	MPL	53	TCC
12	DIN/SRP	26	DZP	40	NPF	54	TCS
13	APR	27	ENX	41	OCT	-	
14	ASP	28	FPF	42	OF	-	

Table 3. Original output attributes.

No.	Attribute	No.	Attribute	No.	Attribute
1	Unclassified	28	Firmicutes	55	Betaproteobacteria
2	Other Archaea	29	Fusobacteria	56	Deltaproteobacteria
3	Crenarchaeota	30	GAL15	57	Epsilonproteobacteria
4	Euryarchaeota	31	GN02	58	Gammaproteobacteria
5	Parvarchaeota	32	GN04	59	TA18
6	Other Bacteria	33	GOUTA4	60	Other Proteobacteria
7	Unclassified Bacteria	34	Gemmatimonadetes	61	SBR1093
8	AC1	35	H-178	62	SC4
9	AD3	36	Hyd24-12	63	SR1
10	Acidobacteria	37	KSB3	64	Spirochaetes
11	Actinobacteria	38	Kazan 3B-28	65	Synergistetes
12	Aquificae	39	LCP-89	66	TM6
13	Armatimonadetes	40	LD1	67	TM7
14	BHI80-139	41	Lentisphaerae	68	TPD-58
15	BRC1	42	MVP-21	69	Tenericutes
16	Bacteroidetes	43	MVS-104	70	Thermotogae
17	Caldiserica	44	NC10	71	Verrucomicrobia
18	Caldithrix	45	NKB19	72	WPS-2
19	Chlamydiae	46	Nitrospirae	73	WS1
20	Chlorobi	47	OC31	74	WS2
21	Chloroflexi	48	OD1	75	WS3
22	Cyanobacteria	49	OP11	76	WS4
23	Deferribacteres	50	OP3	77	WS5
24	Elusimicrobia	51	OP8	78	WWE1
25	FBP	52	OP9	79	ZB3
26	FCPU426	53	Planctomycetes	80	Caldithrix
27	Fibrobacteres	54	Alphaproteobacteria	81	Thermi

Table 4. Statistical parameters of the microbial dataset.

Variable	Unit	Parameter (Abbreviation)	Direction	Min.	Max.	Mean
X ₁	°C	Temperature (Temp)	Input	14.00	30.67	22.75
X ₂	-	pH	Input	6.58	8.40	7.39
X ₃	mg/L	Dissolved Oxygen (DO)	Input	4.37	11.11	7.42
X ₄	mg/L	Ammonium Nitrogen (NH ₄ -N)	Input	0.00	9.24	1.38
X ₅	mg/L	Nitrate Nitrogen (NO ₃ -N)	Input	0.27	18.65	6.03
X ₆	mg/L	Total Phosphorus (TP)	Input	0.08	1.82	0.41
X ₇	mg/L	Orthophosphate (PO ₄ -P)	Input	0.02	0.68	0.20
X ₈	mg/L	Inorganic Nitrogen (DIN)	Input	1.99	24.92	7.74
Y ₁	-	Actinobacteria	Output	0.00	0.54	0.11
Y ₂	-	Bacteroidetes	Output	0.01	0.30	0.10
Y ₃	-	Cyanobacteria	Output	0.00	0.24	0.02
Y ₄	-	Firmicutes	Output	0.00	0.75	0.12
Y ₅	-	Planctomycetes	Output	0.00	0.20	0.03
Y ₆	-	Proteobacteria	Output	0.19	0.90	0.57
Y ₇	-	Others	Output	0.01	0.13	0.05

Water-related parameters, such as surface water temperature, pH, DO, and ammonium (NH₄-N), are examples of physicochemical factors that may be strongly correlated with the freshwater bacterial community. Water-related parameters, such as DIN, TP, and others, were also collected as input attributes. Common dominant microbial species in freshwater are Proteobacteria, Bacteroidetes, Actinobacteria and others. The output attribute “Others” constitutes the percentage of any other microbial community, which has very small value, beside the main six output attributes and is given by the following equation.

$$Y_7 = 1 - \sum_{i=1}^6 Y_i \quad (10)$$

4.2. Determining Critical Factors Related to Microbial Community in a River

To determine the critical factor that is related to microbial community in the river, four stages of dimensionality reduction were applied using three methods, namely expert engineering judgment, feature creation, and feature subset selection, all of which were implemented in IBM SPSS software. Table 5 shows all information relevant to the four stages of dimensionality reduction. In feature creation in Stage II, all categories of pharmaceutical and personal care products (PPCP), such as antipyrine (APR), aspartame (ASP), gemfibrozil (GBL), etc., were combined into one input attribute, PPCPs (ng/L), by summing their concentrations. In Stage IV, feature creation was carried out by combining classes of proteobacteria into a single output attribute, proteobacteria. Table 6 presents the information that was required for feature creation in Stages II and IV.

Table 5. Dimensionality reduction approach at various stages.

Stage	Amount of Inputs and Outputs	Dimensionality Reduction Method
I	54 and 81 38 and 17	Expert engineering judgement
II	38 and 17 10 and 17	Expert engineering judgement, feature creation, feature subset selection
III	10 and 17 10 and 7	Expert engineering judgement, feature creation
IV	10 and 7 8 and 7	Feature subset selection

Table 6. Feature creation at stages II and IV.

Stage	Parameter (Abbreviation)	Feature Creation
II	Antipyrine (APR)	Pharmaceuticals and Personal Care Products (PPCP)
	Paracetamol (ATP)	
	BP-3 (BP3)	
	BPA	
	Caffeine (CAF)	
	Carbamazepine (CZP)	
	Diclofenac (DFA)	
	Diazepam (DZP)	
	Fenoprofen (FPF)	
	Gemfibrozil (GBL)	
	Indomethacine (IDT)	
	Ibuprofen (IPF)	
	Ketoprofen (KPF)	
	Mefenamic (MEF)	
	Methylparaben (MPB)	
	Metoprolol (MPL)	
	Naprofen (NPF)	
	Octocrylene (OCT)	
	OMC	
	Propyl Paraben (PPB)	
	Sulfadiazine (SDZ)	
	Sulfamethoxazole (SFZ)	
	Sulfamethazine (SMT)	
	Thiabendazole (TBD)	
IV	TCC	Proteobacteria
	TCS	
	Alphaproteobacteria	
	Betaproteobacteria	
	Deltaproteobacteria	
	Gammaproteobacteria	

A score box that is related to the feature subset selection method in Stages II and IV is presented.

Table 7 presents the score box in Stage IV for determining the critical factor that is related to the microbial community, based on each attribute's statistical analysis result. Hence, R^2 and analysis of variance (ANOVA) were used as statistical tools to determine the relationships between input and output variables. Colin Cameron and Windmeijer [53] interpreted R^2 , also known as the coefficient of determination, as a contribution of the independent variable (X) to the dependent variable (Y). Analysis of variance (ANOVA) is an extremely important method in exploratory and confirmatory data analysis [54], in which the F value specifies the significance of the results of the statistical analysis, with <0.05 specifying a significance threshold.

Table 7. Score box of dimensionality reduction at stage IV.

	R^2	Sig.	Temp	pH	DIN	DO	NH ₄ -N	NO ₃ -N	TP	PO ₄ -P	NO ₂ -N	PPCP
Bacteroidetes	0.309	0.000		√	√							
Firmicutes	0.474	0.000						√	√	√		
Proteobacteria	0.315	0.000	√	√	√							
Actinobacteria	0.140	0.003				√						
Cyanobacteria	0.335	0.000		√			√					
Planctomycetes	0.237	0.000	√									
Others	0.266	0.000			√							
Score			2	3	3	1	1	1	1	1	0	0

Missing values were predicted using baseline and ensemble models with ANN, SVR, LR, and CART as machine learning techniques, implemented in WEKA. One third of the 60 data points had missing values of the DO attribute, thus the remaining 40 complete data points were used to build a reliable and accurate model to predict those missing values. Table 8 shows the information and attributes that were used to predict the 20 missing values of DO. Park and Lek [55] noted that missing values are often replaced with mean values of the variables or values even better estimated using predictive models.

Table 8. Information of DO Prediction.

Field	Value
Dataset characteristics	Multivariate
Attributes characteristics	Real
Number of instances	40
	Temperature (Temp)
	pH
	Ammonium Nitrogen (NH ₄ -N)
	Nitrate Nitrogen (NO ₃ -N)
	Total Phosphorus (TP)
	Orthophosphate (PO ₄ -P)
	Inorganic Nitrogen (DIN)
	Pharmaceuticals and Personal Care Products (PPCP)
	Crenarchaeota
	Euryarchaeota
	Acidobacteria
	Actinobacteria
	Bacteroidetes
	Chlorobi
	Chloroflexi
	Cyanobacteria
	Firmicutes
	Nitrospirae
	Planctomycetes
	Alphaproteobacteria
	Betaproteobacteria
	Gammaproteobacteria
	Verrucomicrobia
	Others
Input attributes (26)	
Output attribute (1)	Dissolved Oxygen (DO)

Artificial intelligence was used in this study to predict missing values based on the 40 complete data points using ten-fold cross-validation with ten original input attributes and 17 output attributes in Stages II and III. The original output attributes were made input attributes, along with the original input attributes, with the consideration that those input and output attributes may be related. The above concepts were combined to predict the missing values of DO attributes as reliably as possible. WEKA software provides a user-friendly interface for developing AI techniques. The baseline models contain various parameter fields with various settings. Accordingly, for efficiency and ease of use, the default settings of the main parameters in each model in the software were selected.

A comprehensive performance comparison between baseline and ensemble models is crucial to identify the most effective, accurate, and suitable model for predicting microbial species. The simulation result indicates that the ensemble bagging support vector regression model is the best. Thus, the bagging support vector regression ensemble model was used to complement microbial community data.

4.3. Hybrid Model Development

A hybrid multi-input multi-output model was constructed and implemented in MATLAB. Since this hybrid model utilizes accelerated particle swarm optimization (A-PSO) to optimize the hybrid parameters in MLSSVR, which are λ , γ , and ρ , the initial parameters used in study of A-PSO were used. According to Yang, Deb, and Fong [52], the typical parameters for the accelerated PSO are $\alpha \approx 0.1$ – 0.4 and $\beta \approx 0.1$ – 0.7 . The values $\alpha \approx 0.2$ and $\beta \approx 0.5$ can be considered as default for most unimodal objective functions. Table 9 presents the controlling parameters of the optimization tool. The optimization yields the best hyperparameters in MLSSVR, which were then used in the hybrid multi-input multi-output model for predicting the microbial community in a river.

Table 9. Controlling parameters in the optimization tool.

Parameters	Value
Number of Iteration	50
Number of Swarm Population	20
α	0.2
β	0.5
Number of Folds	10
Upper Bound	[2 2 2]
Lower Bound	[−2 −2 −2]

5. Experimental Results

5.1. Metaheuristic Optimized Multiple-Input Multiple-Output Machine Learning

Table 10 presents the optimized parameters of the hybrid multi-input multi-output model and information related to the model herein. Despite the limited data available, the hybrid multi-input multi-output model is a satisfactory model of the microbial community in the river of interest, based on psychochemical factors, as evidenced by the performance measures (MAAPE, MAE, and R^2), which show the acceptability of the multi-output prediction.

Table 10. Hybrid multi-input multi-output prediction model settings.

Field	Value
Number of data points	60
Training/testing data proportions	80/20 (%)
Number of input attributes	8
Number of output attributes	7
Hyperparameters:	
λ	0.1553
γ	0.7752
ρ	1.3620

The hybrid model yields MAAPE, MAE, and R^2 values of 35%, 0.036 and 0.357, respectively, confirming the favorable performance of the MIMOPSO-LSSVR in solving prediction problems in ecological informatics, and, in this case, in the prediction of the microbial community in a river, based on physicochemical factors. This hybrid prediction model can help microbiologists and ecologists to plan for future environmental conditions when working with the microbial community in the river, and provide a new point of view on tools for making predictions about microbial communities without using DNA sequencing.

This investigation also adds a prediction model using two input variables and three input variables based on score box above in critical factor, indicating that the hybrid model performs better if the number of output attributes is less than the number of input attributes. Table 11 presents a performance comparison of the three hybrid predictions, while Figure 4 presents a graphical comparison in terms of MAAPE, MAE, and R^2 . The graphs clearly reveal that the performance of the hybrid model tends to be worse when the number of input attributes is less than the number of output attributes.

Table 11. Performance comparison of hybrid prediction model.

Number of Inputs	Performance Measure						Overall SI
	MAAPE (%)		MAE		R ²		
	Training	Testing	Training	Testing	Training	Testing	
8	53	35	0.041	0.036	0.660	0.357	1 (1)
3	58	37	0.053	0.036	0.442	0.223	0.415 (2)
2	62	39	0.057	0.039	0.351	0.239	0.012 (3)

Note: Bold value denotes the best performance; (.) indicates the overall ranking of the model.

(a)

Figure 4. Cont.

(b)

(c)

Figure 4. Multi-output hybrid prediction model performance comparison of various number of input attributes: (a) MAAPE; (b) MAE; and (c) R-square.

5.2. General Discussion

Several studies applying machine learning related to microbial or ecological modeling are listed in Table 12. Prior works merely used single-output models, in which the parameters were set via personal

experiments. In this investigation, a multiple relation between eight inputs and seven outputs was built, thus the model would save time for the researchers. Additionally, the parameters in the proposed model were determined automatically by a metaheuristic optimization algorithm (i.e., A-PSO) that helps the microbiologists and ecologists who possess little background of machine learning to easily use this model.

Table 12. Machine learning techniques applied in previous studies.

Literature	Machine Learning Technique	Parameter Setting	Model Type	Multiple Outputs
Larsen et al. [2]	Artificial neural network (ANN)	Manual experiment	Single	No
Jerves-Cobo et al. [56]	Decision tree models (DTMs)	Manual experiment	Single	No
Jerves-Cobo et al. [57]	Generalized linear model (GLM)	Manual experiment	Single	No
M.N. Damanik-Ambarita et al. [58]	Generalized linear model (GLM)	Manual experiment	Single	No
Aazami et al. [59]	Multimetric Macroinvertebrate Index (MMI)	Manual experiment	Single	No
Forio et al. [60]	Negative binomial regression (NBM)	Manual experiment	Single	No
This study	Multiple-input multiple-output machine learning	Automatically identified by bio-inspired metaheuristic optimization algorithms	Hybrid	Yes

With optimized values of hyperparameters, i.e. λ , γ and ρ of 0.1553, 0.7752, and 1.3620, respectively, the hybrid multi-output prediction model satisfactorily predicts microbial community in a river based on the physicochemical parameters. This model, which has a MAAPE value of 35%, a MAE value of 0.036, and an R^2 of 0.357, respectively, is categorized as satisfactory because it can solve multi-output problems simultaneously, without neglecting relationships among outputs. Therefore, this model can help microbiologists and ecologists to plan work under future environmental conditions, which has become one of the most critical issues for the world considering climate change. This investigation also contributes to a new point of view regarding techniques other than DNA sequencing for making predictions about microbial communities.

Moreover, as the accuracy of hybrid prediction model based on MAAPE increases by 2% every experiment from two, three, to eight input attributes in the testing phases, the hybrid model with eight inputs outperforms the other two models. This experiment indicates the limitations of MIMOPSO-LSSVR, with the hope that they will be overcome in the future by extending the scope of data points.

6. Conclusions

Understanding the composition and diversity of a microbial community in a given environment and the factors that control them is a critical issue in microbial ecology. Microbial communities can be found in every macro- and microenvironment, including in rivers. This paper develops a bioclimatic model that leverages hybrid artificial intelligence techniques for implementing a multi-input multi-output approach to predict the microbial community in a river, as a function of its physicochemical properties. Unlike numerous studies of predictions about microbial species, this work constructs a hybrid prediction model with multi-inputs and multi-outputs and runs it simultaneously, enabling the model to account for relationships among outputs. Sixty data points concerning the Jiu Long River, South East China, were collected directly in various seasons under various environmental conditions.

This study utilized hybrid multi-output prediction by MLSSVR and optimized its hyperparameters using A-PSO to build a prediction model. Three dimensionality-reducing techniques—expert engineering judgement, feature creation, and feature subset selection by the brute force approach—were combined to identify critical physicochemical factors in microbial data. The simulation results demonstrate that bagging support vector regression (SVR) is the best prediction model of the single/ensemble models for predicting 20 missing values of the DO attribute, with the purpose of completing the microbial community data that were used in this investigation.

The addition of two input and three input variables based on score box revealed the limitations of the proposed MIMOPSO-LSSVR hybrid model. This hybrid model tends to perform well when the number of input attributes exceeds the number of output attributes under the required number of data points. In future work, ecological researchers should collect more data about microbial communities. Since such data are very difficult to obtain for reasons of time and expense, 60 data points—a very limited amount of data—were used herein to build a prediction model. Thus, the accuracy of the developed prediction model would be increased. Finally, this study presents a proof-of-concept approach that is potentially useful for predicting microbial structures in environment by using a hybrid multi-output prediction model, which will likely be effective in the near future in expanding knowledge of ecological informatics.

Author Contributions: The authors' contributions are provided as below: Conceptualization, J.-S.C. and C.-P.Y.; methodology, J.-S.C.; software, D.-N.T. and B.S.; validation, D.-N.T. and B.S.; formal analysis, B.S.; investigation, B.S.; resources, J.-S.C. and C.-P.Y.; data curation, A.H., Q.S. and B.S.; writing—original draft preparation, B.S. and J.-S.C.; writing—review and editing, J.-S.C. and C.-P.Y.; visualization, B.S. and D.-N.T.; supervision, J.-S.C. and C.-P.Y.; project administration, J.-S.C. and C.-P.Y.; funding acquisition, J.-S.C. and C.-P.Y.

Funding: This research and APC was funded by the Ministry of Science and Technology, grant number 107-2221-E-011-035-MY3.

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

References

1. Bier, R.L.; Bernhardt, E.S.; Boot, C.M.; Graham, E.B.; Hall, E.K.; Lennon, J.T.; Nemergut, D.R.; Osborne, B.B.; Ruiz-González, C.; Schimel, J.P. Linking Microbial Community Structure and Microbial Processes: An Empirical and Conceptual Overview. *FEMS Microbiol. Ecol.* **2015**, *91*, 1–11. [[CrossRef](#)] [[PubMed](#)]
2. Larsen, P.E.; Field, D.; Gilbert, J.A. Predicting Bacterial Community Assemblages using an Artificial Neural Network Approach. *Nat. Methods* **2012**, *9*, 621–625. [[CrossRef](#)] [[PubMed](#)]
3. Freguia, S.; Logrieco, E.M.; Monetti, J.; Ledezma, P.; Virdis, B.; Tsujimura, S. Self-Powered Bioelectrochemical Nutrient Recovery for Fertilizer Generation from Human Urine. *Sustainability* **2019**, *11*, 5940. [[CrossRef](#)]
4. Konopka, A. What is Microbial Community Ecology? *ISME J.* **2009**, *3*, 1223–1230. [[CrossRef](#)] [[PubMed](#)]
5. Qian, J.; Yang, T.; Zhang, W.; Lei, Y.; Zhang, C.; Ma, J.; Zhang, C. Preparation of NH₂-Functionalized Fe₂O₃ and Its Chitosan Composites for the Removal of Heavy Metal Ions. *Sustainability* **2019**, *11*, 5186. [[CrossRef](#)]
6. Baek, S.; Kim, S. Optimum Design and Energy Performance of Hybrid Triple Glazing System with Vacuum and Carbon Dioxide Filled Gap. *Sustainability* **2019**, *11*, 5543. [[CrossRef](#)]
7. Wang, S.; Zhang, Z.; Yin, X.; Wang, N.; Chen, D. Influences of Nitrogen Application Levels on Properties of Humic Acids in Chernozem Amended with Different Types of Organic Materials. *Sustainability* **2019**, *11*, 5405. [[CrossRef](#)]
8. Heintz-Buschart, A.; Wilmes, P. Human Gut Microbiome: Function Matters. *Trends Microbiol.* **2017**. [[CrossRef](#)]
9. Sboner, A.; Mu, X.J.; Greenbaum, D.; Auerbach, R.K.; Gerstein, M.B. The Real Cost of Sequencing: Higher Than You Think! *Genome Biol.* **2011**, *12*, 1–10. [[CrossRef](#)]
10. Heather, J.M.; Chain, B. The Sequence of Sequencers: The History of Sequencing DNA. *Genomics* **2016**, *107*, 1–8. [[CrossRef](#)]
11. Janizadeh, S.; Avand, M.; Jaafari, A.; Phong, V.T.; Bayat, M.; Ahmadisharaf, E.; Prakash, I.; Pham, T.B.; Lee, S. Prediction Success of Machine Learning Methods for Flash Flood Susceptibility Mapping in the Tafresh Watershed, Iran. *Sustainability* **2019**, *11*, 5426. [[CrossRef](#)]

12. Sivapriya, T.; Kamal, A.N.B.; Thavavel, V. Imputation and Classification of Missing Data using Least Square Support Vector Machines—A New Approach in Dementia Diagnosis. *Int. J. Adv. Res. Artif. Intell.* **2012**, *1*, 29–33. [[CrossRef](#)]
13. Zhang, D.; Zhu, Q.; Zhang, D. Multi-Modal Dimensionality Reduction using Effective Distance. *Neurocomputing* **2017**, *259*, 130–139. [[CrossRef](#)]
14. Shiklomanov, I.A. *World Water Resources: A New Appraisal and Assessment for the 21st Century: A Summary of the Monograph World Water Resources*; UNESCO International Hydrological Programme, UNESCO-IHP: Paris, France, 1998.
15. Stanley, E.H.; Fisher, S.G.; Grimm, N.B. Ecosystem Expansion and Contraction in Streams. *BioScience* **1997**, *47*, 427–435. [[CrossRef](#)]
16. Ghai, R.; Rodríguez-Valera, F.; McMahon, K.D.; Toyama, D.; Rinke, R.; de Oliveira, T.C.S.; Garcia, J.W.; de Miranda, F.P.; Henrique-Silva, F. Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. *PLoS ONE* **2011**, *6*, e23785. [[CrossRef](#)] [[PubMed](#)]
17. Newton, R.J.; Bootsma, M.J.; Morrison, H.G.; Sogin, M.L.; McLellan, S.L. A Microbial Signature Approach to Identify Fecal Pollution in the Waters Off an Urbanized Coast of Lake Michigan. *Microb. Ecol.* **2013**, *65*, 1011–1023. [[CrossRef](#)]
18. Zhang, K.; Zhang, Y.; Zhou, C.; Meng, J.; Sun, J.; Zhou, T.; Tao, J. Impact of climate factors on future distributions of *Paenonia ostii* across China estimated by MaxEnt. *Ecol. Inform.* **2019**, *50*, 62–67. [[CrossRef](#)]
19. Ager, D.; Evans, S.; Li, H.; Lilley, A.K.; Van Der Gast, C.J. Anthropogenic Disturbance Affects the Structure of Bacterial Communities. *Environ. Microbiol.* **2010**, *12*, 670–678. [[CrossRef](#)]
20. Kneip, C.; Lockhart, P.; Voß, C.; Maier, U.-G. Nitrogen Fixation in Eukaryotes—New Models for Symbiosis. *BMC Evol. Biol.* **2007**, *7*, 55. [[CrossRef](#)]
21. Bernhard, A. The Nitrogen Cycle: Processes. Available online: <https://www.nature.com/scitable/knowledge/library/the-nitrogen-cycle-processes-players-and-human-15644632> (accessed on 18 December 2018).
22. Gougoulas, C.; Clark, J.M.; Shaw, L.J. The Role of Soil Microbes in the Global Carbon Cycle: Tracking the Below-Ground Microbial Processing of Plant-Derived Carbon for Manipulating Carbon Dynamics in Agricultural Systems. *J. Sci. Food Agric.* **2014**, *94*, 2362–2371. [[CrossRef](#)]
23. Schlegel, H. Microorganisms Involved in the Nitrogen and Sulfur Cycles. In *Biology of Inorganic Nitrogen and Sulfur*; Springer: Berlin/Heidelberg, Germany, 1981; pp. 3–12. [[CrossRef](#)]
24. Hu, A.; Yao, T.; Jiao, N.; Liu, Y.; Yang, Z.; Liu, X. Community Structures of Ammonia-Oxidising Archaea and Bacteria in High-Altitude Lakes on the Tibetan Plateau. *Freshw. Biol.* **2010**, *55*, 2375–2390. [[CrossRef](#)]
25. Fenwick, A. Waterborne Infectious Diseases—Could They be Consigned to History? *Science* **2006**, *313*, 1077–1081. [[CrossRef](#)] [[PubMed](#)]
26. Cabral, J.P. Water Microbiology. Bacterial Pathogens and Water. *Int. J. Environ. Res. Public Health* **2010**, *7*, 3657–3703. [[CrossRef](#)] [[PubMed](#)]
27. Pirofski, L.-A.; Casadevall, A. Q and A What is a Pathogen? *BMC Biol.* **2012**, *10*, 6. [[CrossRef](#)]
28. Breznak, J.A.; Costilow, R.N. Physicochemical Factors in Growth. In *Methods for General and Molecular Microbiology*, 3rd ed.; American Society of Microbiology: Washington DC, USA, 2007; pp. 309–329. [[CrossRef](#)]
29. Alesheikh, S.; Shahtahmassebi, N.; Roknabadi, M.R.; Pilevar Shahri, R. Silicene Nanoribbon as a New DNA Sequencing Device. *Phys. Lett. A* **2018**, *382*, 595–600. [[CrossRef](#)]
30. Yang, N.; Jiang, X. Nanocarbons for DNA Sequencing: A Review. *Carbon* **2017**, *115*, 293–311. [[CrossRef](#)]
31. Feng, Y.; Zhang, Y.; Ying, C.; Wang, D.; Du, C. Nanopore-Based Fourth-Generation DNA Sequencing Technology. *Genom. Proteom. Bioinform.* **2015**, *13*, 4–16. [[CrossRef](#)]
32. Ansorge, W.J. Next-Generation DNA Sequencing Techniques. *New Biotechnol.* **2009**, *25*, 195–203. [[CrossRef](#)]
33. Kircher, M.; Heyn, P.; Kelso, J. Addressing Challenges in the Production and Analysis of Illumina Sequencing Data. *BMC Genom.* **2011**, *12*, 382. [[CrossRef](#)]
34. Buermans, H.P.J.; den Dunnen, J.T. Next Generation Sequencing Technology: Advances and Applications. *Biochim. Et Biophys. Acta (Bba) Mol. Basis Dis.* **2014**, *1842*, 1932–1941. [[CrossRef](#)]
35. Wu, S.G.; Wang, Y.; Jiang, W.; Oyetunde, T.; Yao, R.; Zhang, X.; Shimizu, K.; Tang, Y.J.; Bao, F.S. Rapid Prediction of Bacterial Heterotrophic Fluxomics using Machine Learning and Constraint Programming. *PLoS Comput. Biol.* **2016**, *12*, e1004838. [[CrossRef](#)] [[PubMed](#)]

36. Han, J.; Kamber, M.; Pei, J. Data Preprocessing. In *Data Mining*, 3rd ed.; Morgan Kaufmann: Boston, MA, USA, 2012.
37. Famili, A.; Shen, W.-M.; Weber, R.; Simoudis, E. Data Preprocessing and Intelligent Data Analysis. *Intell. Data Anal.* **1997**, *1*, 3–23. [[CrossRef](#)]
38. Ngoc Thach, N.; Bao-Toan Ngo, D.; Xuan-Canh, P.; Hong-Thi, N.; Hang Thi, B.; Nhat-Duc, H.; Dieu, T.B. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecol. Inform.* **2018**, *46*, 74–85. [[CrossRef](#)]
39. Rajeswari, K.; Vaithiyanathan, V.; Pede, S.V. Feature Selection for Classification in Medical Data Mining. *Int. J. Emerg. Trends Technol. Comput. Sci. (Ijettcs)* **2013**, *2*, 492–497. [[CrossRef](#)]
40. Kwak, S.K.; Kim, J.H. Statistical Data Preparation: Management of Missing Values and Outliers. *Korean J. Anesthesiol.* **2017**, *70*, 407–411. [[CrossRef](#)]
41. Qi, Z.; Wang, H.; Li, J.; Gao, H. FROG: Inference from Knowledge Base for Missing Value Imputation. *Knowl. Based Syst.* **2018**, *145*, 77–90. [[CrossRef](#)]
42. Tsai, C.-F.; Li, M.-L.; Lin, W.-C. A Class Center based Approach for Missing Value Imputation. *Knowl. -Based Syst.* **2018**, *151*, 124–135. [[CrossRef](#)]
43. Li, E.Y. Artificial Neural Networks and Their Business Applications. *Inf. Manag.* **1994**, *27*, 303–313. [[CrossRef](#)]
44. Chou, J.-S.; Yang, K.-H.; Lin, J.-Y. Peak Shear Strength of Discrete Fiber-Reinforced Soils Computed by Machine Learning and Metaensemble Methods. *J. Comput. Civ. Eng.* **2016**, *30*, 04016036. [[CrossRef](#)]
45. Zou, K.H.; Tuncali, K.; Silverman, S.G. Correlation and Simple Linear Regression. *Radiology* **2003**, *227*, 617–628. [[CrossRef](#)]
46. Kumar, R. Decision Tree for the Weather Forecasting. *Int. J. Comput. Appl.* **2013**, *76*, 31–34. [[CrossRef](#)]
47. Chou, J.-S.; Ho, C.-C.; Hoang, H.-S. Determining quality of water in reservoir using machine learning. *Ecol. Inform.* **2018**, *44*, 57–75. [[CrossRef](#)]
48. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [[CrossRef](#)]
49. Jiang, T.; Zhou, X. Gradient/Hessian-Enhanced Least Square Support Vector Regression. *Inf. Process. Lett.* **2018**, *134*, 1–8. [[CrossRef](#)]
50. Xu, S.; An, X.; Qiao, X.; Zhu, L.; Li, L. Multi-Output Least-Squares Support Vector Regression Machines. *Pattern Recognit. Lett.* **2013**, *34*, 1078–1084. [[CrossRef](#)]
51. Khennak, I.; Drias, H. An Accelerated PSO for Query Expansion in Web Information Retrieval: Application to Medical Dataset. *Appl. Intell.* **2017**, *47*, 793–808. [[CrossRef](#)]
52. Yang, X.-S.; Deb, S.; Fong, S. Accelerated Particle Swarm Optimization and Support Vector Machine for Business Optimization and Applications. In *Proceedings of the International Conference on Networked Digital Technologies*, Macau, 11–13 July 2011; pp. 53–66.
53. Colin Cameron, A.; Windmeijer, F.A.G. An R-Squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *J. Econom.* **1997**, *77*, 329–342. [[CrossRef](#)]
54. Huang, L.-S.; Chen, J. Analysis of Variance, Coefficient of Determination and F-Test for Local Polynomial Regression. *Ann. Stat.* **2008**, *36*, 2085–2109. [[CrossRef](#)]
55. Park, Y.-S.; Lek, S. Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. In *Developments in Environmental Modelling*; Elsevier: Amsterdam, The Netherlands, 2016; Volume 28, pp. 123–140.
56. Jerves-Cobo, R.; Córdova-Vela, G.; Iñiguez-Vela, X.; Díaz-Granda, C.; Van Echelpoel, W.; Cisneros, F.; Nopens, I.; Goethals, P. Model-based analysis of the potential of macroinvertebrates as indicators for microbial pathogens in rivers. *Water* **2018**, *10*, 375. [[CrossRef](#)]
57. Jerves-Cobo, R.; Forio, M.A.E.; Lock, K.; Van Butsel, J.; Pauta, G.; Cisneros, F.; Nopens, I.; Goethals, P.L.M. Biological water quality in tropical rivers during dry and rainy seasons: A model-based analysis. *Ecol. Indic.* **2020**, *108*, 105769. [[CrossRef](#)]
58. Damanik-Ambarita, M.N.; Everaert, G.; Forio, M.A.E.; Nguyen, T.H.T.; Lock, K.; Musonge, P.L.S.; Suhareva, N.; Dominguez-Granda, L.; Bennetsen, E.; Boets, P.; et al. Generalized linear models to identify key hydromorphological and chemical variables determining the occurrence of macroinvertebrates in the guayas river basin (ecuador). *Water* **2016**, *8*, 297. [[CrossRef](#)]

59. Aazami, J.; Esmaili Sari, A.; Abdoli, A.; Sohrabi, H.; Van den Brink, P.J. Assessment of ecological quality of the tajan river in iran using a multimetric macroinvertebrate index and species traits. *Environ. Manag.* **2015**, *56*, 260–269. [[CrossRef](#)] [[PubMed](#)]
60. Forio, M.A.E.; Goethals, P.L.M.; Lock, K.; Asio, V.; Bande, M.; Thas, O. Model-based analysis of the relationship between macroinvertebrate traits and environmental river conditions. *Environ. Model. Softw.* **2018**, *106*, 57–67. [[CrossRef](#)]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).