water

MDPI

Article

Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model

Illa Iza Suhana Shamsuddin, Zalinda Othman * and Nor Samsiah Sani *

Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
* Correspondence: zalinda@ukm.edu.my (Z.O.); norsamsiahsani@ukm.edu.my (N.S.S.)

Abstract: Traditionally, water quality is evaluated using expensive laboratory and statistical procedures, making real-time monitoring ineffective. Poor water quality requires a more practical and cost-effective solution. Water pollution has been a severe issue, hurting water quality in recent years. Therefore, it is crucial to create a model that forecasts water quality to control water pollution and inform consumers in the event of the detection of poor water quality. For effective water quality management, it is essential to accurately estimate the water quality class. Motivated by these considerations, we utilize the benefits of machine learning methods to construct a model capable of predicting the water quality index and water quality class. This study aims to investigate the performance of machine learning models for multiclass classification in the Langat River Basin water quality assessment. Three machine learning models were developed using Artificial Neural Networks (ANN), Decision Trees (DT), and Support Vector Machines (SVM) to classify river water quality. Comparative performance analysis between the three models indicates that the SVM is the best model for predicting river water quality in this study. In addition, there is a statistically significant difference in performance between the SVM, DT, and ANN models at the 0.05 level of confidence. The use of the kernel function, the grid search method, and the multiclass classification technique used in this study significantly impacts the effectiveness of the SVM model. The findings bolster the idea that machine learning models, particularly SVM, can be used to forecast WQI with a high degree of accuracy, hence enhancing water quality management. Consequently, the model based on machine learning lowered the cost and complexity of calculating sub-indices of six water quality parameters and classifying water quality compared to the standard IKA-JAS formula.

Keywords: classification; machine learning; water quality index (WQI); Langat River Basin

Citation: Shamsuddin, I.I.S.; Othman, Z.; Sani, N.S. Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. *Water* **2022**, *14*, 2939. https://doi.org/10.3390/w14192939

Academic Editors: Xing Fang, Jiangyong Hu and Suresh Sharma

Received: 6 August 2022
Accepted: 10 September 2022
Published: 20 September 2022

1. Introduction

Water quality has been monitored by the Department of Environment (DOE) since 1978, primarily to set guidelines for detecting changes in water quality and identify sources of pollution. Current water quality monitoring in Malaysia is based on the IKA-JAS. IKA-JAS is used to measure the degree of pollution and classify water quality in accordance with the National Water Quality Standard and the kind of water used. River water quality in Malaysia is categorized into five classes based on IKA-JAS. IKA-JAS is used to measure the pollution level and water use suitability as outlined by the National Water Quality Standard (SKAN). IKA-JAS considers the six water quality parameters in the formula and its calculations to produce a score value. The parameters are dissolved oxygen, biochemical oxygen requirement, chemical oxygen requirement, ammonia nitrogen, suspended solids, and pH. These parameters are obtained from water samples that have been analyzed to determine water's physical, chemical, and biological properties [1]. The score value obtained will then be compared with the IKA-JAS water quality index range to determine the water quality class. Water quality classification using the conventional IKA-JAS method will be problematic when one of the water quality parameters has a missing value. Therefore,

the calculation of the sub-index of water quality parameters cannot be performed. This affects the aggregation of six sub-indices of water quality parameters to obtain IKA-JAS score values. The WQI calculation procedure in Malaysia involves a lengthy calculation, transformation, and application of various sub-index formulas for each water quality indicator [2]. Traditional WQI computation, according to Ho et al. [3], is always accompanied by inaccuracy induced by water quality parameter assessment. In addition, obtaining and analyzing water samples takes a lot of time and work. Furthermore, determining the size of certain factors by experimentation comes at a considerable expense. According to Abba et al. [4], water quality evaluation is a multidimensional problem, and one of the problems in water quality research is to reduce expenses and develop clever computer assistance in analyzing water quality. Previous studies have suggested assessing water quality using computational methods based on data mining algorithms and machine learning techniques [5–10].

Previous research reveals that conventional water quality assessment by the mathematical model WQI has drawbacks. The WQI mathematical model requires sophisticated calculations to get a final index value. WQI computation includes parameter selection, sub-index value generation, parameter weight generation, and index aggregation technique selection, and there is no universal standard approach. The water quality class is determined by comparing this final index to the WQI range. Therefore, it can cause long calculation times, costs, and errors in water quality classification. The next constraint is the complex relationship between water quality measures and environmental elements like human, industrial, and commercial operations, anthropogenic activities, and natural processes. Previous research has proven this limitation based on WQI's parameter selection and number. Most researchers used standard approaches to model the water quality index. Creating a WQI mathematical model in the environmental domain that covers all parameter relationships is tough. Water quality has complex, non-linear interactions.

Machine learning is a data analysis tool that automates the creation of analysis models. It is a branch of artificial intelligence that focuses on machines learning from data, detecting patterns, and making choices without human intervention [11]. Machine learning dominates decision-making because it can automate complex tasks [12]. Artificial neural network methods are increasingly used in water resource studies and environmental science [13–15]. Khoi et al. [16] analyzed 12 machine learning algorithms. XGBoost accurately predicts WQI, improving water quality management. Ma et al. [17] used machine learning to estimate total suspended solids (TSS) and chlorophyll-a (Chl-a) in the turbid Pearl River Estuary (PRE). The study revealed that the ANN-based algorithm performed well. Alqahtani et al. [18] compared individual supervised machine learning models with an ensemble learning model for predicting river water salinity in the Upper Indus River Basin, Pakistan. The study recommended using the RF model with specified key parameters to assess and manage water quality. Ahmed et al. [19] used the ANN model to calculate WQI for Sungai Kinta using 23 water quality parameters and a heuristic search. In [20], the authors predicted WQI using a 7-23-1 network architecture, backpropagation training algorithm, and a learning rate of 0.02 to produce the most accurate WQI predictions. The findings show that ANN is a reliable method of relating water quality to land use, thus integrating land use development with river quality management. In ref. [21], the authors predicted Perak River Basin WQI in real-time using 25 water quality parameters without BOD and COD. The study found that combining multiple neural networks improves WQI prediction performance. A study by Chen et al. [11] used a linear regression model (LRM), multi-layered perceptron neural network (MLP), and radial basis function neural network (RBF-NN) for water quality prediction in the Johor River Basin. The results showed that using the RBF-NN model can describe the behavior of water quality parameters more accurately than the linear regression model. The effectiveness and performance of the RBF method were also demonstrated in a study by Hameed et al. [22]. This study presents a flexible structure of a Radial Basis Function (RBF) neural network (FS-RBFNN) and its application for water quality prediction based on neuronal activity and reciprocal information

(MI). The experimental results show that FS-RBFNN can be used to design RBF structures that have fewer hidden neurons. Hence, the training time is also faster.

Machine learning algorithms have succeeded in the environmental domain because of their ability to model complex relationships between variables. Although there are numerous machine learning algorithms, researchers continue to face challenges, such as determining which machine learning techniques should be used or are most appropriate for a specific problem. Therefore, to fill the gap based on past studies, this study aims to propose an approach based on machine learning techniques: the multiclass classification model to classify water quality index. This methodology is expected to change the way the classification of water quality index and water quality are monitored. Water quality assessment has become an important issue in water resources management. Conventional water quality assessment methods using WQI can be time-consuming and expensive, especially for complex datasets with multiple water quality parameters. Machine learning techniques have the ability to cut down on computation time, costs, and errors in water quality classification, water quality parameter forecasting, and water quality index forecasting. Machine learning is also considered an alternate technique for calculating the water quality index, including several sub-indices [15]. In recent years, machine learning approaches have been extensively utilized for river water quality assessment, including the calculation of WQI. These techniques have proven to be effective modeling tools for complicated non-linear processes in water resource studies. Each machine learning method has advantages and disadvantages, and its behavior depends on the input factors of water quality in the various research regions. Hence, a proactive approach is desperately needed to address Malaysia's water quality classification issues. For that purpose, an effective prediction model based on machine learning can be implemented. Therefore, this study aims to evaluate the performance of three machine learning algorithms on water quality classification problems. A classification model was developed based on Artificial Neural Network (ANN), Decision Trees (DT), and Support Vector Machine (SVM) to predict the WQI of the Langat River Basin.

2. Study Area

The Langat River Basin is one of Malaysia's most important river water catchment basins, as shown in Figure 1. It is located in the state of Selangor. The Langat River Basin spans an area of approximately 2409 km^2 and is located between latitudes 2°40′152″ N and 31°60′152″ N, and longitudes 101°19′20″ E and 102°1′10″ E [23]. The Titiwangsa granite mountain range is located upstream of the Langat River Basin, where the Langat River originates at Gunang Nuang and flows approximately 190 km through the states of Selangor and Negeri Sembilan, as well as the Federal Territories of Putrajaya and Kuala Lumpur, before entering the Malacca Strait. The river's basin is drained by the Langat, Semenyih, and Labu rivers. Meanwhile, the Langat River Basin's upstream dams are the Langat Dam and the Semenyih Dam. The two dams, Semenyih and Langat, in the Langat Basin, might serve as drinking water reservoirs. Apart from these dams, there are several ex-mining ponds scattered across the basin, particularly near the Paya Indah Wetlands in Kuala Langat. The basin's geography is described as both hilly and flat, with an average elevation of 400–1440 m. The average elevation of the central basin is less than 200 m, followed by less than 100 m in the lower basin. The igneous rock underneath the Langat River Basin is mostly granite. As a result, the basin's geology is characterized as Hawthornden Schist and Kenny Hill Formation (sandstone and phyllite). The primary soil types in the hilly upstream, flat midstream, and downstream include Tanah Curam (steep land), Rengan-Jerangan (urban land), and Tanah Gambut (peat soils). During the period 2005–2016, the average annual rainfall in the Langat River Basin ranged from 2043.68 to 2832.40 mm.

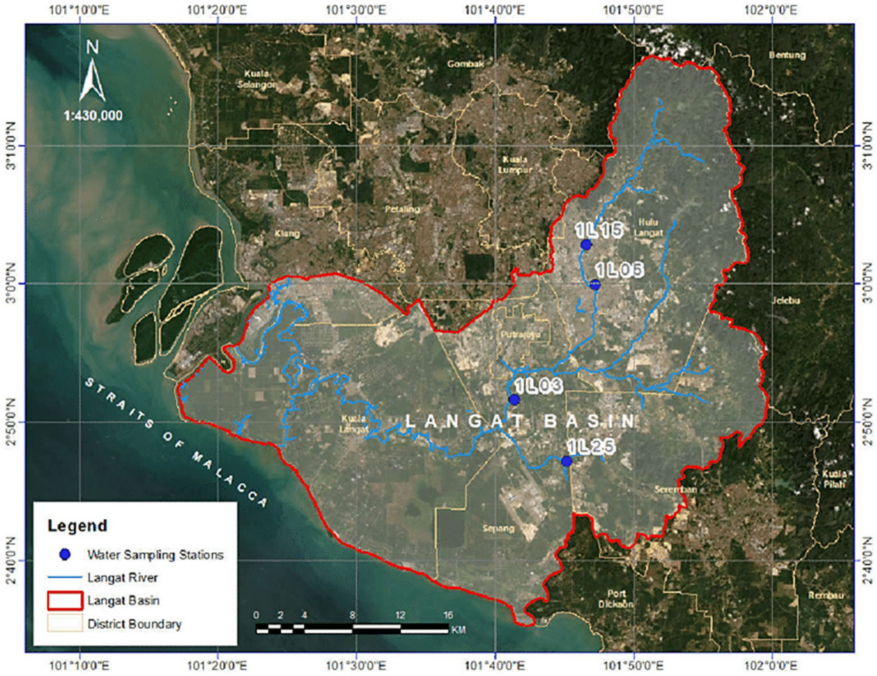

Figure 1. Langat River Basin [23].

3. Research Methodology

The methodology of this study is divided into five phases, as shown in Figure 2: (i) data collection and understanding of water quality data from the Langat River Basin is obtained from the Malaysia Department of Environment and databases for the development of the multiclass model in predicting water quality classes; (ii) data preparation involves preprocessing data for minimizing or eliminating inconsistencies of data; (iii) multiclass classification model development involves the development of multiclass machine learning models to classify the water quality; and (iv) model evaluation involves the evaluation of the water quality multiclass classification model.

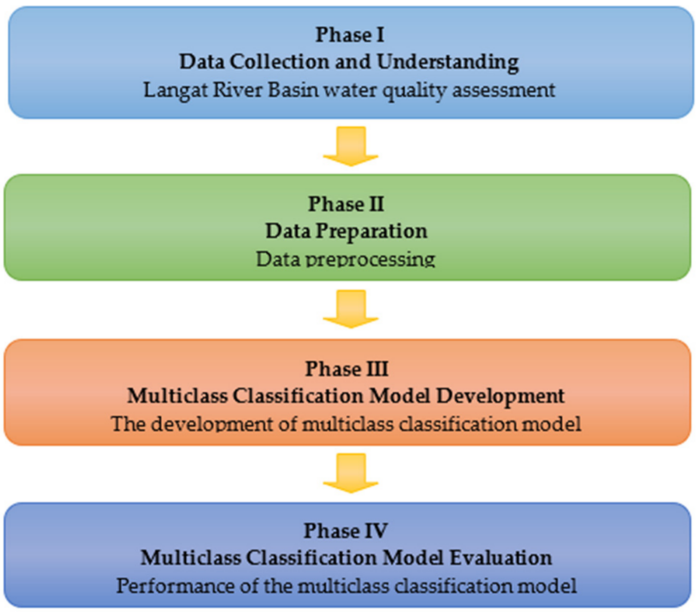

Figure 2. The main phases of the study design.

3.1. Phase I: Data Collection and Understanding

The Malaysia Department of Environment provided water quality data for the Langat River Basin for this study. The parameters used to analyze and evaluate water quality can generally be divided into physical, chemical, and biological parameters. Physical parameters include color, temperature, taste, odor, turbidity, and water solids suspended in water. Chemical parameters include dissolved oxygen, acidity, pH, and alkalinity, biochemical oxygen requirements, chemical oxygen requirements, ammonia nitrogen, electrical conductivity, total solids, and other pollutants. In contrast, biological parameters include bacteria such as fecal coliforms and algae.

The Langat River Basin raw water quality dataset is a time-series dataset. This raw dataset has a total of 560 records and 46 attributes taken from fourteen monitoring stations recorded over five years, from January 2012 to December 2016. The information related to the 46 attributes is divided and described as follows.

Table 1 shows a list of attributes that contain information related to the monitoring station where the raw water quality dataset was obtained. Table 1 also presents the attributes of the sub-index of parameters used in calculating the Water Quality Index-Department of Environment (WQI-JAS) formula for water quality assessment.

Table 1. WQI-JAS attributes.

Attributes	Description	Data Type
STATES	State	Nominal
BASIN	Basin	Nominal
LATITUDE	Latitude	Nominal
LONGTITUDE	Longitude	Nominal
WKA	Basin code	Integer
STA NO	Monitoring station number	Nominal
RIVER	The river	Nominal
SMP-DAT	Date the water sample was taken	Date
DO INDEX	Dissolved oxygen sub-index	Real
BOD INDEX	Sub-index of biochemical oxygen requirements	Integer
CODE INDEX	Sub-index of chemical oxygen requirements	Integer
AN INDEX	Ammonia nitrogen sub-index	Integer
SS INDEX	Suspended solids sub-index	Real
PH INDEX	Sub-index of pH	Real
WQI	The aggregation of sub-indices to a single value of WQI represents the water quality index	Real

Note(s): The six water quality parameter attributes used in the WQI-JAS formula are: DO, BOD, COD, SS, AN, and pH, which are chosen by an expert panel using Opinion Poll WQI to assess the water quality class of rivers in Malaysia. Meanwhile, the four attributes, STA NO, RIVER, SMP-DAT, and WQI, represent the monitoring station information and water quality index score value.

3.2. Phase II: Data Preparation

Data preparation begins with data exploration using data preprocessing techniques to better comprehend the study's datasets. There are several data preprocessing techniques used in the study. The basis of data preprocessing is a descriptive statistical analysis that helps study the general characteristics of the data and identify the presence of noise or outliers. Next, data cleaning can eliminate noise and inconsistencies in the data. Data transformations such as normalization can help improve the accuracy and efficiency of machine learning algorithms [24]. At the same time, data visualization is very helpful in visually inspecting data using plot graphs.

The Langat River Basin water quality dataset used in this study has a dimensional space of 560 rows and 10 columns. There are seven rivers in the Langat River Basin: Langat; Semenyih; Lui; Pajam; Value Bar; Jijan; and Pumpkin Stems, which are the locations where monitoring stations are located, and river water samples are collected.

Identifying the data type for each attribute is important because data are obtained in various formats and types. Table 1 shows a list of data types for each attribute in the initial

dataset that will be analyzed statistically and descriptively in the next process. Descriptive statistical analysis of the data can better describe the data and help to understand the main features of the data distribution. The characteristics of the dataset used in the study can be understood through measures of central tendency such as mean, median, and mode; and measures of data dispersion such as range, quartile, variance, and standard deviation. This descriptive statistical summary technique can be used to handle data for machine learning tasks such as identifying data properties and highlighting values identified as noise data, missing data, or outliers [25]. Table 2 shows a descriptive statistical analysis of six attributes from the initial dataset representing the water quality parameters used in calculating the WQI-JAS formula. The Missing Value column means no data value is recorded for the attribute in the observation. Missing data is a common phenomenon and can significantly impact the conclusions drawn from the data [26]. The Minimum and Maximum columns represent the minimum and maximum values in the study dataset, while the Mean (μ) columns are the average values of the attributes.

Table 2. Descriptive statistics of water quality parameters.

Parameters	Missing Value	Minimum	Maximum	Min	Standard Deviation
DO	0	11.5	154.3	75,787	21,062
BOD	2	1	43	8582	5701
COD	1	2	167	26,648	17,396
SS	3	0	821	62,864	78,208
pH	0	3.8	8.5	7124	0.568
AN	0	0.005	13.2	1.7	2126

The collection and storage of water quality data from *in situ* monitoring stations as well as those sent for analysis in the laboratory are highly vulnerable to noise data, missing data, as well as incomplete and inconsistent data. Low-quality data will result in low-quality data mining results as well. The data cleaning process cleans up data by filling in missing values, smoothing out noise data, identifying or removing external elements, and resolving inconsistent data problems.

Data recorded from laboratory work carried out by different individuals or teams can result in errors such as multiple copies, data redundancy, repetitive data, and inconsistencies. This recurring data was removed through a data cleansing process, and the current total of the initial dataset is 553 records.

Raw data is rarely cleaned and often has corrupted or missing values. Therefore, it is important to identify, mark, and handle missing data before implementing the machine learning model development phase [27]. This process is important to get the best model performance. A total of 5 records out of 553 initial dataset records will be deleted through the data cleanup process. These records were removed from the initial dataset because the number of missing value records was small, and less than 5% of the water quality samples were obtained. In addition, missing data from BOD, COD, SS, and WQI attributes will affect the water quality classification model to be developed.

In data transformation, data is transformed or consolidated into a suitable form for the data mining process [28,29]. In order to transform data beyond the initial dataset, attribute construction and normalization are required. In this study, a new attribute was derived from the initial dataset's attributes in order to facilitate data mining.

The CLASS attribute is a newly formed attribute derived from the WQI attribute. This feature represents the water quality class in the Langat River Basin at a particular location and time. Table 3 thoroughly defines the CLASS attribute as a nominal-type output variable, four class labels, and the number of records associated with each class. The definition of representation refers to the class of water and the suitability of the type of water used as outlined by the National Water Quality Standard (SKAN).

Table 3. Description of the CLASS attribute.

Data Type	Label	Number of Records	Definition of Representation
Nominal	I	28	CLASS I = Water Supply I/Fisheries I
	II	169	CLASS II = Water Supply II/Fisheries II
	III	319	CLASS III = Water Supply III/Fisheries III
	IV	32	CLASS IV = Irrigation
	Amount	548	

Table 3 is based upon the number of CLASS attribute records representing water quality in the Langat River Basin from January 2012 to December 2016. The CLASS attribute as an output or target variable referred to as a label in a classification model for water quality will be created utilizing 548 water quality dataset records. Consequently, based on the obtained label values, namely CLASS I, CLASS II, CLASS III, and CLASS IV. Since there is more than one class attribute, this study is a multiclass classification problem.

The data normalization process involves scaling the six attributes of water quality parameters into a small range of values due to the measurement of water quality parameters in different units. The normalization method introduced by Hameed et al. [22] will be used in this study by setting the minimum value of the normalized data set to 0.1 and the maximum value normalized to 0.9. This data normalization method was chosen because the environment of the study area used is almost similar to the current study and tends to give better results with high-performance accuracy.

Data normalization was performed according to the following Formula (1):

$$X_{new} = (X_{max} - X_{min}) \frac{X - X_{min}}{X_{max} - X_{min}} + X_{min} \quad (1)$$

Based on Formula (1), X_{new} is the normalized value of the original attribute, X is the original data point, and X_{min} and X_{max} are the minima and maximum values in the dataset.

Since the focus of the study was on water quality assessment based on the water quality parameters used in the calculation of the WQI-JAS formula, only the attributes related to WQI-JAS were selected, which are DO, BOD, COD, SS, pH, and AN. Four attributes have been eliminated during the attribute reduction step: STA NO, RIVER, SMP-DATE, and WQI. The initial filtered water quality dataset was reduced from ten to six attributes. In the next phase, a dataset of 548 records will be used as a training and test dataset in developing a multiclass classification model for water quality classification.

3.3. Phase III: Multiclass Classification Model Development

The algorithms that have been used in previous studies for water quality assessment have their strengths and limitations. Therefore, based on the previous studies that have been described, the ANN, DT, and SVM algorithms were selected to perform the classification task on the Langat River Basin water quality dataset used in this study. The ANN algorithm was chosen because of its ability to model non-linear and complex relationships between input and output variables or where relationships between input variables are difficult to understand [29], especially when involving water's physical, chemical, and biological parameters. The DT algorithm was chosen because it is a classification technique that is easy to understand and widely used. In addition, it is appropriate to train the dataset used, where memory usage is minimized, making it a time-saving and cost-effective approach for water quality classification [3]. Next, the SVM algorithm was chosen because it can model non-linear relationships between input variables. It uses non-linear mapping to transform the original training data to a higher dimension. It works by classifying data into different classes by finding the optimal separator hypersometric. The optimal separator hyperplane separates the training data into classes and maximizes the distance to the nearest point from one of the classes. As a result of maximizing the margin between the two classes on the training data, the classification performance is better on the test data, thus achieving

maximum generalization [30]. SVM is also the most commonly used data mining algorithm for water quality assessment. It is a powerful alternative to artificial neural networks in predicting water quality in non-point-source polluted rivers [5].

Three algorithms (i.e., ANN, DT, and SVM) have been developed to find the best classifier for the Langat River Basin water quality assessment. Figure 3 shows the experimental design for the multiclass classifier used in this study. The model development involves: (1) dividing the data into training and testing sets using the cross-validation method; (2) selecting an algorithm for the classification model; (3) the model parameters optimization; and (4) training and evaluating the model. Based on Figure 3, three multiclass classifiers will be developed and evaluated to determine the best classifier. This study will employ the search grid method to identify the best parameters for the classification model to optimize the model's performance. Whereas the performance of each classifier is measured using k-*fold cross-validation*, where $k = 10$.

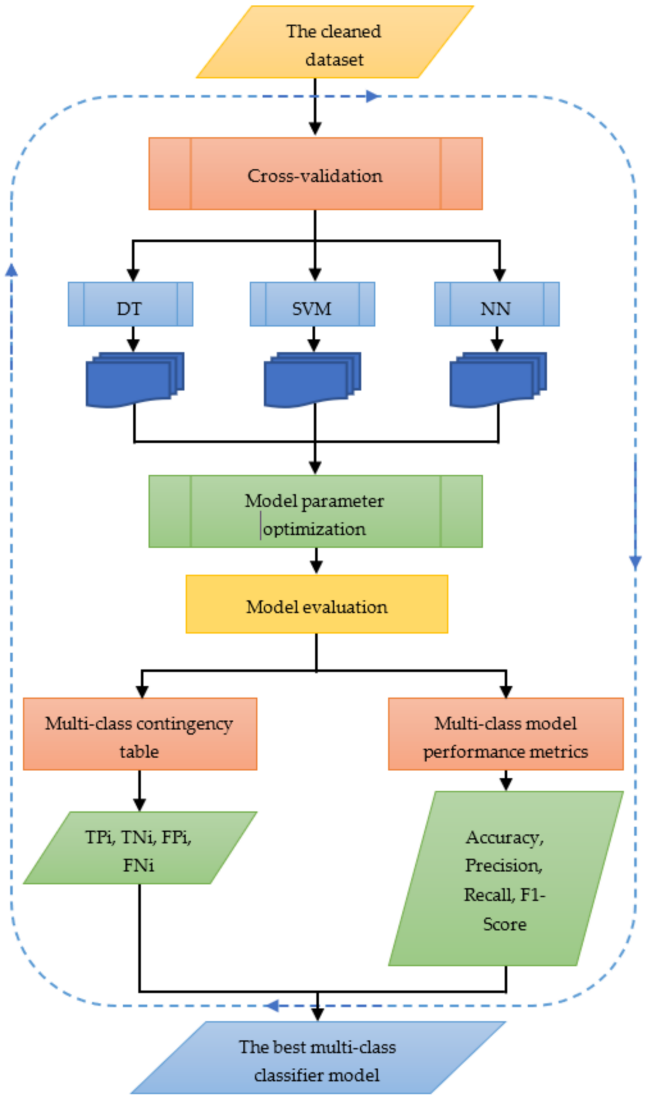

Figure 3. Experimental design of a multiclass classification model.

3.3.1. Neural Network Settings

Neural networks consist of a set of interconnected layers. The first layer is the input layer and is connected to the output layer by an acyclic graph consisting of weighted sides and nodes. Between the input and output layers, there is a hidden layer where the prediction task can be performed easily with only one or several hidden layers. In general, neural network classification requires a labeled dataset containing labeled columns. The

labeled datasets act as inputs to train the model. Values are calculated at each node in the hidden layer and at the output layer to calculate the network output for a particular input. The value is set by calculating the weighted sum of the node values from the previous layer. The activation function is then applied to that weighted amount.

According to Sani et al. [11], the choice of network type relies on the problem being solved. Backpropagation algorithms are commonly used to train neural networks. This exercise is usually performed by updating the weights iteratively based on the error signal. The error is calculated in the output layer as the difference between the true class and the actual output value multiplied by the gradient of the sigmoid activation function. Then, the error signal is propagated back to the bottom layer. Therefore, backpropagation is a gradient descent algorithm that tries to minimize each iteration's error. The learning algorithm adjusts the network weights so that the error decreases along the descending direction.

The neural network algorithm used in this study is based on a multi-layer feed-forward neural network trained with stochastic gradient decrease using backpropagation. As shown in Table 4, the developed ANN model has three parameters configured to find the optimum value. The grid search method was used to obtain the best parameter settings for the ANN model.

Table 4. ANN model parameter settings.

No.	Parameters	Description	Grid Search Range
1	*activation*	The activation function (nonlinearity) used by neurons in the hidden layer	{Tanh, Rectifier, Maxout, ExpRectifier}
2	*learning_rate*	A parameter that measures the magnitude of the weighting update to minimize the network loss function	{0.01–1.0}
3	*rate_annealing*	The learning rate of annealing reduces the learning rate to be trapped into the local minimum in the optimization space	{0.01–1.0}

3.3.2. Decision Tree Settings

The basic idea behind the decision tree is to use a divide-and-conquer approach. The decision tree (DT) algorithm can address binary or multiple class classification problems and can be represented in the form of a tree structure. Each tree node can be a leaf node or a decision node. The leaf node denotes the value of the target attribute or class. For multiclass classification problems, leaf nodes can refer to one of the relevant N classes. The result node determines the number of tests performed on one attribute from an existing observation by generating one possible branch of that test. The process of classifying specific data through a result tree begins by evaluating the test contained in the root node or result node. It moves through it up to the leaf node, which determines the classification of the data. In this study, the DT model to be developed has several parameters that need to be configured. As shown in Table 5, four parameters in the DT model will be set to find the optimal value. The grid search method was used to obtain the best parameter settings for the DT model.

Table 5. DT model parameter settings.

No.	Parameters	Description	Grid Search Range
1	*criterion*	Selects the criteria by which attributes will be selected for separation	{information_gain, gain_ratio, gini_index}
2	*apply_pruning*	The decision tree model can be pruned after generation	{true, false}
3	*confidence*	This parameter determines the level of confidence used for the calculation of the pessimistic pruning error	$\{1.0 \times 10^{-7}$–0.25$\}$

Table 5. *Cont.*

No.	Parameters	Description	Grid Search Range
4	*minimal_gain*	The minimum value that must be reached to separate nodes	{0.0–0.3}

3.3.3. Support Vector Machine Model Settings

According to Haixiang et al. [31], the quality of SVM parameter selection and kernel functions affects the performance of the SVM model. Therefore, the appropriate value of the kernel function and its parameters should be selected to obtain optimal classification performance. Once the appropriate kernel functions and their parameters have been obtained, then the prediction errors of the SVM model can be minimized. In this study, the developed SVM model has three parameters that will be configured to find the optimal value as shown in Table 6. The grid search method was used to obtain the best parameter settings for the SVM model

Table 6. SVM model parameter settings.

No.	Parameters	Description	Grid Search Range
1	*kernel_type*	SVM kernel functions	{dot, radial, polynomial, neural, anova, epachnenikov, gaussian combination}
2	*C*	SVM complexity constant that sets the tolerance for misclassification	{1.0–100.0}
3	*polynomial by binomial classification*	Build a polynomial classification model through binomial classification	{1 against all, 1 against 1}

3.4. Phase IV: Multiclass Classification Model Evaluation

The prediction model's performance was evaluated by comparing the values of accuracy, precision, recall, and F1-Score. Those values were calculated based on the confusion matrix. Prediction results and actual class were put in a matrix for comparison depending on a positive and negative value. Confusion matrices have two types of errors: Type I and Type II. A Type I error is also known as a false positive, and a Type II error is known as a false negative. For a multiclass problem with a number of classes k, the size of the confusion matrix is N^{kxk}. The confusion matrix can represent the classification results, as shown in Table 7.

Table 7. Multiclass confusion matrix.

		True			
	k	**1**	**2**	…	**n**
Predicted	1	$C_{1.1}$	$C_{1,2}$	…	$C_{1,k}$
	2	$C_{2.1}$	In, j	…	
	…	…	…	…	…
	n	$C_{k,1}$	…	…	$C_{k,k}$

Based on Table 7, where $k = \{1, 2, \ldots, n\}$ while each element C_i, j represents the number of tuples predicted to be class i but belong to class j. At the same time, the sum of all the elements in the confusion matrix is equal to the sum of the N samples given to the classifier. From the confusion matrix of the various classes, the number of true positive predictions of TP for each class k is given by Equation (2):

$$TP_k = C_{k,k} \quad (2)$$

where k is a reference to an individual class.

Next, the number of false negative predictions of FN (Type II error) for each true class k can be obtained based on the following Equation (3):

$$FN_k = \sum_{i=1, i \neq j, j=k}^{n} C_{i,j} \quad (3)$$

where n = total number of classes, i = predicted class row, and j = correct class column.

The number of true negative predictions of TN for each class k can be calculated according to the following Equation (4):

$$TN_k = \sum_{i=1,\ i \neq k}^{n} \sum_{j=1,\ j \neq k}^{n} C_{i,j} \quad (4)$$

Next, the number of false positive predictions of FP (Type I error) for each class k is given by the following Equation (5):

$$FP_k = \sum_{j=1,\ j \neq i, i=k}^{n} C_{j,i} \quad (5)$$

The experiment results of ANN, DT, and SVM are recorded and analyzed to see the performance of each multiclass classification model. These multiclass classification models were developed to classify four class labels, namely 'CLASS I', 'CLASS II', 'CLASS III', or 'CLASS IV', for water quality assessment of the Sungai Langat Basin.

Three classification algorithms are compared in this study, which are ANN, DT, and SVM. Each classifier is tuned using different tuning parameters to produce highly accurate results. This study utilized the search grid approach to determine the optimal parameters for the classification model to optimize the model's performance. A series of experiments were conducted to get the optimal values of each classifier. The performance between the three classifiers is then evaluated and compared. Table 8 shows the optimum parameter setting after the model parameter optimization experiment for the ANN, DT, and SVM. The investigation revealed that the optimal ANN parameters are the activation function = rectifier, learning rate = 0.2, and annealing rate = 0.5. Meanwhile, for the DT model settings, the experiment showed the gini_index is the criteria that determines the separation of attributes, with the minimum value to separate the node being 0:09 and the confidence value being 0.12500005. Additionally, Table 8 shows the optimum parameter of the SVM model, in which the kernel function for the model is the linear function, while the constant complexity C is set as 100.0. The parameter polynomial by binomial classification is used to expand from binary to multiclass classification, and the technique 1 against 1 has been selected.

Table 8. Parameter setting results for the ANN, DT, and SVM model.

Parameter	Parameter Optimum	Machine Learning Model
activation	*Rectifier*	ANN
learning_rate	0.2	
rate_annealing	0.5	
criterion	*gini_index*	DT
apply_pruning	*false*	
confidence	0.12500005	
minimal_gain	0.09	
kernel_type	*dot*	SVM
C	100.0	
polynomial by binomial lassification	1 against 1	

Table 9 shows the results of TP, FP (Type I error), TN, and FN (Type II error) values for each predicted class. Based on Table 9, the ANN model correctly classified the data with the true positive (TP) values represented by TPI, TPII, TPIII, and TPIV, where TPI = 14, TPII = 163, TPIII = 301, and TPIV = 22, according to Equation (2). Meanwhile, the DT model correctly classified the data at the TP rate as TPI = 16, TPII = 146, TPIII = 307, and TPIV = 21. Furthermore, the SVM model correctly classified the TP rate with TPI = 22, TPII = 158, TPIII = 305, and TPIV = 23.

Table 9. Confusion matrix for ANN, DT, and SVM.

		True			
	CLASS	**I**	**II**	**III**	**IV**
Predicted	I	14_{ANN}; 16_{DT}; **22_{SVM}**	1_{ANN}; 3_{DT}; 2_{SVM}	0_{ANN}; 0_{DT}; 0_{SVM}	0_{ANN}; 0_{DT}; 0_{SVM}
	II	14_{ANN}; 12_{DT}; 6_{SVM}	**163_{ANN};** 146_{DT}; 158_{SVM}	17_{ANN}; 11_{DT}; 12_{SVM}	0_{ANN}; 0_{DT}; 0_{SVM}
	III	0_{ANN}; 0_{DT}; 0_{SVM}	5_{ANN}; 20_{DT}; 9_{SVM}	301_{ANN}; **307_{DT};** 305_{SVM}	10_{ANN}; 11_{DT}; 9_{SVM}
	IV	0_{ANN}; 0_{DT}; 0_{SVM}	0_{ANN}; 0_{DT}; 0_{SVM}	1_{ANN}; 1_{DT}; 2_{SVM}	22_{ANN}; 21_{DT}; **23_{SVM}**

Based on the confusion matrix results shown in Table 9, the rate of TP for models ANN, DT, and SVM is compared and analyzed in the graph, as shown in Figure 4. The model that reaches the highest TP for 'Class I' is the SVM model, with a TP of 22 correctly classified versus the DT model of 16, followed by the ANN model of 14. The next model to reach the highest TP for 'Class II' is the ANN model with 163 correctly classified versus the SVM model of 158, followed by the DT model of 146. On the other hand, the highest TP for 'Class III' is the DT model, with the TP of 307 correctly classified versus the SVM model of 305, followed by the ANN model of 301. While the model to reach the highest TP for 'Class IV' is the SVM, with a TP of 23, versus the ANN model of 22, followed by the DT model of 21.

Figure 4. Comparison of TP between NN, DT, and SVM models.

Table 10 compares the three models' TP, FP (Type I error), and FN (Type II error). The value of FN (Type II error) for each true class can be calculated according to Equations (3) and (5), respectively, by adding all the classification errors in the respective class column. Table 10

shows that for the ANN, the value of FN (Type II error) and FP (Type I error) is 48. While for the DT, the value of FN (Type II error) and FP (Type I error) is 58. Whereas for the SVM, the value of FN (Type II error) and FP (Type I error) is 40. In contrast, each class's true negative (TN) can be calculated by following Equation (4). As for the comparison of the three models, the model to reach the highest TP is the SVM, with a total number of TP of 508 correctly classified as compared to the ANN of 500 and followed by model DT of 490. The results show that the SVM classifier performs better in classifying data into different classes.

Table 10. Comparison of TP, FP (Type I error), and FN (Type II error) between ANN, DT, and SVM models.

	ANN			DT			SVM		
CLASS	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
I	14	1	14	16	3	12	22	2	6
II	163	31	6	146	23	23	158	18	11
III	301	15	18	307	31	12	305	18	14
IV	22	1	10	21	1	11	23	2	9
Total	500	48	48	490	58	58	**508**	40	40

In order to select the best model, the confusion matrices were used to measure accuracy, precision, recall, and F1-Score by using macro average and micro average approaches. ROC curves are also a good method for evaluating models, but the classification model developed in this study is based on an imbalanced dataset; consequently, the ROC curve is not a good visual representation of imbalanced data because decision thresholds are not explicitly depicted in the ROC curve, and the distinction between the models is difficult to define [32]. Consequently, we present the total performance of classification models using macro-averaged scores to avoid bias for major categories in the imbalanced data associated with micro-averaged scores. This is because we are particularly concerned with the performance of minor categories. Macro averaging assumes all classes are equal and important, while micro averaging favors a larger class. Because of the imbalanced data, macro averaging is used to evaluate the performance of the classifier model for each class, while micro averaging is used to evaluate the performance of the classifier model as a whole. The macro averaging and micro averaging of accuracy, precision, recall, and F1-Score can be calculated according to the following Equations ((6)–(13)):

$$\text{Micro Average Accuracy } = \frac{\sum_{k=1}^{n} TP_k}{N} \tag{6}$$

$$\text{Micro Average Precision} = \frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n} (TP_k + FP_k)} \tag{7}$$

$$\text{Micro Average Recall} = \frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n} (TP_k + FN_k)} \tag{8}$$

$$\text{Micro Average F1} - \text{Score} = 2 * \frac{\textit{Micro Average Recall} * \textit{Micro Average Precision}}{\textit{Micro Average Recall} + \textit{Micro Average Precision}} \tag{9}$$

$$\text{Macro Average Accuracy} = \frac{\sum_{k=1}^{n} \frac{TP_k + TN_k}{TP_k + FN_k + FP_k + TN_k}}{n} \tag{10}$$

$$\text{Macro Average Precision} = \frac{\sum_{k=1}^{n} \frac{TP_k}{TP_k + FP_k}}{n} \tag{11}$$

$$\text{Macro Average Recall} = \frac{\sum_{k=1}^{n} \frac{TP_k}{TP_k + FN_k}}{n} \tag{12}$$

$$\text{Macro Average F1} - \text{Score} = 2 * \frac{\sum_{k=1}^{n} \frac{Recall_k * Precision_k}{Recall_k + Precision_k}}{n} \quad (13)$$

Table 11 shows the performance of ANN, DT, and SVM based on the accuracy (Acc), precision (Pr), recall (Rc), and FI-Score (F1) using macro and micro averaging. The model that provides predictions with higher accuracy, precision, and recall can perform the classification task.

Table 11. Performance of ANN, DT, and SVM models based on the accuracy (Acc), precision (Pr), recall (Rc) values, and F1-Score (F1).

CLASS	ANN				DT				SVM			
	Acc (%)	Pr (%)	Rc (%)	F1 (%)	Acc (%)	Pr (%)	Rc (%)	F1 (%)	Acc (%)	Pr (%)	Rc (%)	F1 (%)
I	97.26	93.33	50	65.12	97.26	84.21	57.14	68.08	98.54	91.67	78.57	84.62
II	93.25	84.02	96.45	89.81	91.61	86.39	86.39	86.39	94.71	89.77	93.49	91.59
III	93.98	95.25	94.36	94.80	92.15	90.83	96.24	93.46	94.16	94.43	95.61	95.02
IV	97.99	95.65	68.75	80.00	97.81	95.45	65.62	77.77	97.99	92.00	71.88	80.70
Macro averaging	95.62	92.06	77.39	82.43	94.71	89.22	76.35	81.43	**96.35**	**91.97**	**84.89**	**87.98**
Micro averaging	91.24	91.24	91.24	91.24	89.42	89.42	89.42	89.42	92.7	92.7	92.7	92.70

Figure 5 shows a comparison graph of model performance based on the percentage of prediction accuracy, precision, recall, and F1-Score. The micro averaging accuracy shows that SVM produces the highest micro accuracy of 92.7% compared to the ANN at 91.24% and the DT at 89.42%. The same value for micro accuracy, micro precision, micro recall, and micro F1-Score of each model, where *Acc* = *Pr* = *Rc* = *F1*, is shown in Figure 5.

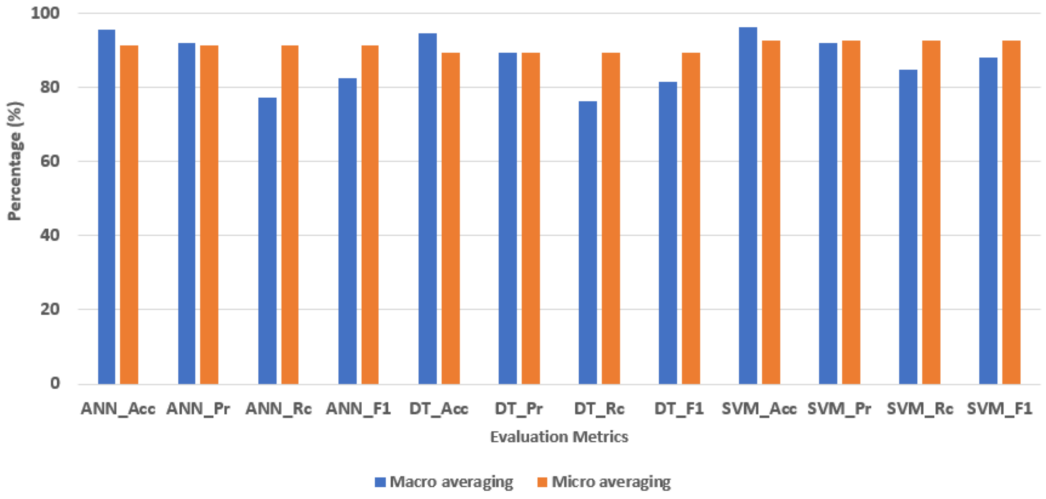

Figure 5. Macro average percentage vs. micro average percentage.

Results of *Pr* micro and *Rc* micro are the same with precision micro when each data point is only given to one class. This can be explained where the count value *Pr* micro and *Rc* micro by Equations (7) and (8), and the rate of TP, FP (Type I error), and FN (Type II error) for all label classes must be known in advance (refer to Table 10). Therefore, when a data point from one class label is predicted, if the result is a false positive FP (Type I error), then there will be a false negative FN (Type II error) and vice versa. For example, if the label 'CLASS I' is predicted and the true class label is 'CLASS II', then the prediction result or 'CLASS I' is false positive FP (Type I error) and false negative FN (Type II error) or 'CLASS II'. If the prediction is correct, the 'CLASS I' is predicted, and the class label is 'CLASS I', the results of the prediction are true positive TP.

Referring to Figure 5, SVM again shows the highest macro accuracy of 96.35% compared to the ANN and DT models. Additionally, SVM shows the highest macro precision of 91.97%, macro recall of 84.89%, and micro F1-Score at 87.98% (refer to Figure 5) compared to the ANN and DT models. The macro accuracy of ANN was at 95.62%, supported by the

macro precision at 92.06%, macro recall at 77.39%, and macro F1-Score at 82.43%. Next, the macro accuracy of DT results at 94.71% is supported by the macro precision at 89.22 %, macro recall at 76.35%, and macro F1-Score at 81.42%.

Statistical Significance Tests

Comparing the machine learning algorithms and selecting the best model is a common process in the machine learning task. In this study, the performance of algorithms is compared using statistical tests, namely the paired corrected *t*-test on 548 instances of water quality data. The ANN, DT, and SVM classifiers were evaluated against the water quality data sets with a significant level of 0.05 (95%). In the *t*-test, ANN is used as a baseline model. For the comparison with DT and SVM, 10-fold cross-validation is used.

From Table 12, the algorithm ANN, as the baseline for comparison, is marked as (1) and has an accuracy of 95.62%. These results are compared with the DT algorithm marked as (2) and the SVM algorithm marked as (3). The symbol '*' next to the DT result indicates that the result is different from the ANN result, but the score is lower. In contrast, the symbol 'v' next to SVM showed that the SVM results are larger than ANN, and the difference is significant, with an accuracy of 96.35%. This shows that the SVM is the best multiclass classifier and is statistically significant at a 0.05 confidence level.

Table 12. The paired corrected *t*-test results. * indicates that the result is different from the ANN.

Tester	**Paired Corrected *t*-Test**		
Analyzing	Percent_correct		
Dataset	1		
Result sets	3		
Confidence	0.05 (two tailed)		
Date	15/01/2020		
Dataset	**(1) ANN**	**(2) DT**	**(3) SVM**
Water quality	95.62 (v / / *)	94.71 * (0/0/1)	96.35 v (1/0/0)

4. Discussions

This study examines the use of machine learning algorithms for multiclass classification models in assessing water quality in the Langat River Basin. A set of water quality data for a period of five years (2012–2016) were used in this study. Three series of experiments were conducted using a neural network algorithm, a decision tree algorithm, and a support vector machine algorithm with six water quality parameters as model input, as well as four WQI class labels as the target output.

In this study, the outcomes revealed that three machine learning models perform well in predicting the WQI; however, SVM is the most accurate model developed using a small dimensional space of 560 rows and 10 columns of dataset. According to previous studies [5,33,34], the SVM is an efficient method and has outperformed artificial neural networks in many studies related to the classification of water quality data. In prior studies, the SVM algorithm effectively predicts water quality using small datasets, whereas the ANN approach is superior for large, high-dimensional datasets. Moreover, it is necessary to specify a kernel function of good quality to achieve an SVM with high classification accuracy. The experimental results show that for the stated SVM parameter setting, the linear kernel is found to be the best choice for the classification process. In this study, the model used the kernel function to map input data into the high-dimensional feature space and find the optimum hyperplanes to separate the two data classes. The experimental results also show that the SVM model can achieve the highest performance using the 1 against 1 multiclass approach.

Furthermore, the ANN model achieved better accuracy in classifying given input data based on the water quality parameters of the Sungai Langat Basin compared to DT. The experimental results show that the ANN model can improve the accuracy of

water quality class classification and model the complex non-linear relationship between the input and output data compared to other conventional techniques. Several studies have also supported this; ANN shows high-performance accuracy compared with other conventional methods in modeling and predicting the water quality index in a tropical environment [11,19–22,33–35]. In addition, the DT model is good for identifying the WQI class label because it is simple to comprehend and implement. It is capable of facilitating, analyzing, and classifying water quality data. However, further investigations are required to improve the accuracy of the DT model.

The classification model developed in this study involves an unbalanced dataset because the number of data points in each class is different. In other words, one class label has a very large number of observations whereas the other has a very small number of observations. According to Haixiang et al. [31], an imbalanced dataset is referred to as the dataset in which one or more classes have a larger number of samples than the other. The class with the highest number of samples is called the majority class, while the lowest number of samples is called the minority class. Nevertheless, in this study, each class is important because it refers to the type of water class and the suitability of the type of water use as outlined by the National Water Quality Standards. Therefore, to address the imbalanced dataset problem, a confusion matrix can be used to show a more detailed breakdown of the true and false classifications for each class. The confusion matrix used includes a column matrix representing true class labels, while the row matrix represents predicted classes.

On the other hand, more detailed fractional information for correct classification and incorrect classification for each class would be lost if model performance evaluations were only measured using the overall accuracy of all classes. The overall accuracy of all classes will give a less accurate picture because larger classes will dominate the results. Therefore, the model performance evaluation for each class was measured using the average of each class's accuracy because there were different numbers of data for each class in the dataset used in this study. The average of each accuracy class is also known as accuracy using macro averaging. The evaluation based on macro averaging is important when the study dataset has unbalanced classes. Since we are highly concerned with the performance of every category, particularly the minor ones, we present the overall performance using macro-averaged scores to prevent bias for major categories in the imbalanced data associated with micro-averaged scores [36]. Other metrics, such as accuracy, recall, and F1-score, are frequently used in the research community to evaluate models trained on imbalanced data [37]. Therefore, the confusion matrix, together with accuracy, precision, recall, and F1-Score (based on macro and micro averaging), is used as the performance measure. In general, for a given class label, different combinations of precision and recall represent the following meanings; (1) high precision and high recall indicate that the class label is handled perfectly by the classifier model; (2) high precision and low recall indicate the class label cannot identify the class label well but is very reliable when it occurs; (3) low precision and high recall indicate the class label is well identified, but the classifier model also identifies other classes in it; and (4) low precision and low recall indicate the class label is not well handled by the classifier model.

Thus, the experimental results and model evaluation based on the multiclass confusion matrix, including accuracy, precision, recall, and F1-Score using macro averaging, show the SVM model is the best multiclass classifier compared to ANN and DT. The *t*-test results have also shown that the SVM model is the best multiclass classifier, and the test results are statistically significant at the 0.05 confidence level. The findings strengthen the argument that machine learning models, particularly SVM, may be used to forecast WQI with a high accuracy level, hence enhancing water quality management. Overall, the experimental and evaluation results of the models presented could change the way WQI classes are classified, and water quality monitored in the future, thus enabling better water resources management by reducing costs and time involved in monitoring and evaluation processes.

Although this study was successful in accomplishing its objectives, there is still an opportunity for continuous improvement to be carried out. This is because, when developing classification and prediction models, the primary focus is on how to generate better models and results. The use of larger datasets to predict water quality classes is one of the recommended enhancements. Using larger datasets involves training on historical data and developing classifier models to predict new data. Furthermore, using larger datasets allows for the discovery of more hidden patterns. The models generated can be improved by conducting more detailed investigations into the associations between water quality measures. At the same time, the experimental design for model configuration can be modified automatically to give additional functions such as loss function optimization, utilizing search methods such as random search that explore over hyperparametric space to find a desirable configuration value.

Next, the use of feature selection algorithms can be introduced in future studies to test the prediction and accuracy of classification models based on various scenarios consisting of different water quality parameters. In addition, supervised machine learning algorithms can be utilized for time series prediction problems on the raw water quality dataset of a time series dataset. Several supervised machine learning algorithms have recently been developed for the R and Python programming environments. This opportunity can be taken to explore algorithms developed to solve time series prediction problems.

5. Conclusions

This paper presents the experimental results of the model development phase that has been implemented. The experimental results were analyzed comparatively based on the multiclass confusion matrix, followed by accuracy, precision, and recall using macro and micro averaging to evaluate the performance of the ANN, DT, and SVM models. Overall, the experimental results show that the ANN, DT, and SVM performance is good while having advantages and disadvantages, respectively.

Next, the performance of each classification model was measured and compared using evaluation metrics that include confusion matrix, accuracy, precision, and retrieval to determine the best multiclass classification model. Comparative analysis based on accuracy, precision, and retrieval using macro and micro averaging showed that all three models had achieved more than 85% performance. The best model that achieved the highest classification performance was SVM, with an accuracy rate of 96.35%, supported by a precision value of 91.97% and a recovery of 84.89% based on macro averaging. The SVM model is also a multiclass model that is good at classifying data from different classes based on a confusion matrix. As a result, the SVM model was selected as the best classifier model in this study, and the *t*-test showed that the test results were statistically significant at the 0.05 confidence level.

This study succeeded in identifying the best techniques and algorithms among the three models developed to classify the suitability of water use types according to the standards as outlined by SKAN. Earlier, monitoring data recorded over five years from January 2012 to December 2016 from seven rivers managed by DOE were processed in this study using data preprocessing techniques. The classification model developed has successfully identified water quality from various classes based on clean and quality Langat River Basin data.

Moreover, the multiclass classification model developed using the decision tree algorithm in this study has achieved higher accuracy with six optional parameters determined by the expert panel as compared to the decision tree model developed in the previous study by Ho et al. [3]. As a result, as compared to utilizing the traditional IKA-JAS formula, the research strategy based on data mining and machine learning approaches reduced the cost and complexity of calculating sub-indices of six water quality parameters and classifying water quality.

One of the most important tasks in developing a machine learning model is evaluating its performance to assure the classifier model's success and the study's effectiveness. This

study has mediated the use of metric evaluation based on micro and macro averaging to address the classification problem of various classes and unbalanced datasets. The effectiveness of the study has indirectly contributed to the improvement of water quality management by providing data mining approaches and machine learning techniques that offer a variety of classification and forecasting methods to meet the specific needs of policymakers, environmental experts, and the general public.

This study achieved its goals, yet there is room for improvement. When developing classification and prediction models, the focus is on improving results. For predicting water quality classes, larger datasets are recommended. Training on greater historical data and developing classifier models to predict new data are required when using larger datasets. Larger datasets reveal more hidden patterns. More extensive research of water quality measures can improve the models. The experimental design for model configuration can be modified automatically to give additional functions such as loss function optimization using search methods such as random search to explore hyperparametric space. Future studies can use feature selection algorithms to test the accuracy of classification models based on diverse water quality parameter scenarios.

Furthermore, supervised machine learning algorithms can be used to solve time series prediction issues on a raw water quality dataset. Several supervised machine learning techniques for the R and Python programming environments have recently been created. This is an excellent opportunity to investigate techniques designed to handle time series prediction challenges.

Author Contributions: Conceptualization, I.I.S.S. and Z.O.; data curation, I.I.S.S.; formal analysis, I.I.S.S.; funding acquisition, Z.O. and N.S.S.; investigation, I.I.S.S. and Z.O.; methodology, I.I.S.S. and Z.O.; project administration, I.I.S.S., Z.O. and N.S.S.; resources, Z.O. and N.S.S.; software, I.I.S.S.; validation, Z.O. and N.S.S.; supervision, Z.O. and N.S.S.; visualization, I.I.S.S.; writing—original draft, I.I.S.S.; writing—review and editing, Z.O. and N.S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Universiti Kebangsaan Malaysia: GGP-2020-032 and GUP-2019-060.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Li, X.; Cheng, Z.; Yu, Q.; Bai, Y.; Li, C. Water-quality prediction using multimodal support vector regression: Case study of Jialing River, China. *J. Environ. Eng.* **2017**, *143*, 04017070. [CrossRef]
2. Seyed Asadollah, S.B.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
3. Ho, J.Y.; Afan, H.A.; El-Shafie, A.H.; Koting, S.; Mohd, N.S.; Jaafar, W.Z.; Sai, H.L.; Abdul Malek, M.; Ahmed, A.N.; Wan Mohtar, W.H.M.; et al. Towards a time and cost effective approach to water quality index class prediction. *J. Hydrol.* **2019**, *575*, 148–165. [CrossRef]
4. Abba, S.I.; Hadi, S.J.; Sammen, S.S.; Salih, S.Q.; Abdulkadir, R.A.; Pham, Q.B.; Yaseen, Z.M. Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *J. Hydrol.* **2020**, *587*, 124974. [CrossRef]
5. 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]
6. Danades, A.; Pratama, D.; Anggraini, D.; Anggriani, D. Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In Proceedings of the 2016 6th International Conference on System Engineering and Technology (ICSET) 2016, Bandung, Indonesia, 3–4 October 2016; pp. 137–141.
7. Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods. *Water Qual. Res. J.* **2018**, *53*, 3–12. [CrossRef]

8. Muhammad, S.Y.; Makhtar, M.; Rozaimee, A.; Aziz, A.A.; Jamal, A.A. Classification model for water quality using machine learning techniques. *Int. J. Softw. Eng. Its Appl.* **2015**, *9*, 45–52. [CrossRef]
9. Prakash, R.; Tharun, V.P.; Devi, S.R. A comparative study of various classification techniques to determine water quality. In Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) 2018, Coimbatore, India, 20–21 April 2018; pp. 1501–1506. [CrossRef]
10. Wang, X.; Zhang, F.; Ding, J. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Sci. Rep.* **2017**, *7*, 12858. [CrossRef]
11. Sani, N.S.; Mohamed Nafuri, A.F.; Othman, Z.A.; Ahmad Nazri, M.Z.; Mohamad, K.N. Drop-out prediction in higher education among B40 students. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 550–559. [CrossRef]
12. Abdulkareem, A.B.; Sani, N.S.; Sahran, S.; Alyessari, Z.A.A.; Adam, A.; Abd Rahman, A.H.; Abdulkarem, A.B. Predicting COVID-19 based on environmental factors with machine learning. *Intell. Autom. Soft Comput.* **2021**, *28*, 305–320. [CrossRef]
13. Kulisz, M.; Kujawska, J.; Przysucha, B.; Cel, W. Forecasting water quality index in groundwater using artificial neural network. *Energies* **2021**, *14*, 5875. [CrossRef]
14. Wang, S.; Peng, H.; Liang, S. Prediction of estuarine water quality using interpretable machine learning approach. *J. Hydrol.* **2022**, *605*, 127320. [CrossRef]
15. Malek, N.H.A.; Wan Yaacob, W.F.; Md Nasir, S.A.; Shaadan, N. Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water* **2022**, *14*, 1067. [CrossRef]
16. Khoi, D.N.; Quan, N.T.; Linh, D.Q.; Nhi, P.T.T.; Thuy, N.T.D. Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. *Water* **2022**, *14*, 1552. [CrossRef]
17. Ma, C.; Zhao, J.; Ai, B.; Sun, S.; Yang, Z. Machine Learning Based Long-Term Water Quality in the Turbid Pearl River Estuary, China. *J. Geophys. Res. Ocean.* **2022**, *127*, e2021JC018017. [CrossRef]
18. Alqahtani, A.; Shah, M.I.; Aldrees, A.; Javed, M.F. Comparative Assessment of Individual and Ensemble Machine Learning Models for Efficient Analysis of River Water Quality. *Sustainability* **2022**, *14*, 1183. [CrossRef]
19. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using supervised machine learning. *Water* **2019**, *11*, 2210. [CrossRef]
20. Al-Adhaileh, M.H.; Alsaade, F.W. Modelling and prediction of water quality by using artificial intelligence. *Sustainability* **2021**, *13*, 4259. [CrossRef]
21. Mansor, N.; Sani, N.S.; Aliff, M. Machine learning for predicting employee attrition. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 435–445. [CrossRef]
22. Hameed, M.; Sharqi, S.S.; Yaseen, Z.M.; Afan, H.A.; Hussain, A.; Elshafie, A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. *Neural Comput. Appl.* **2017**, *28*, 893–905. [CrossRef]
23. Ahmed, M.F.; Lim, C.K.; Mokhtar, M.B.; Khirotdin, R.P.K. Predicting Arsenic (As) Exposure on Human Health for Better Management of Drinking Water Sources. *Int. J. Environ. Res. Public Health.* **2021**, *18*, 7997. [CrossRef]
24. Nasif, A.; Othman, Z.A.; Sani, N.S. The Deep Learning Solutions on Lossless Compression Methods for Alleviating Data Load on IoT Nodes in Smart Cities. *Sensors* **2021**, *21*, 4223. [CrossRef]
25. Rahman, M.A.; Sani, N.S.; Hamdan, R.; Othman, Z.A.; Abu Bakar, A. A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. *PLoS ONE* **2021**, *16*, e0255312. [CrossRef]
26. Holliday, J.D.; Sani, N.; Willett, P. Ligand-based virtual screening using a genetic algorithm with data fusion. *Match Commun. Math. Comput. Chem.* **2018**, *80*, 623–638.
27. Bakar, A.A.; Hamdan, R.; Sani, N.S. Ensemble Learning for Multidimensional Poverty Classification. *Sains Malays.* **2020**, *49*, 447–459. [CrossRef]
28. Muhammad, A.; Abdullah, S.; Sani, N.S. Optimization of Sentiment Analysis Using Teaching-Learning Based Algorithm. *Comput. Mater. Contin.* **2021**, *69*, 1783–1799. [CrossRef]
29. Othman, Z.A.; Bakar, A.A.; Sani, N.S.; Sallim, J. Household overspending model amongst B40, M40 and T20 using classification algorithm. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 392–399. [CrossRef]
30. Zhang, J.; Williams, S.O.; Wang, H. Intelligent computing system based on pattern recognition and data mining algorithms. *Sustain. Comput. Inform. Syst.* **2018**, *20*, 192–202. [CrossRef]
31. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
32. Vuttipittayamongkol, P.; Elyan, E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Inf. Sci.* **2020**, *509*, 47–70. [CrossRef]
33. Sakizadeh, M. Assessment the performance of classification methods in water quality studies, A case study in Karaj River. *Environ. Monit. Assess.* **2015**, *187*, 573. [CrossRef] [PubMed]
34. Rajaee, T.; Khani, S.; Ravansalar, M. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103978. [CrossRef]
35. Ahmad, Z.; Rahim, N.A.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River Basin Malaysia through a combination of multiple neural networks. *Int. J. River Basin Manag.* **2017**, *15*, 79–87. [CrossRef]

36. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [CrossRef]
37. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]