Computational and Numerical Modeling for Classification of Water Quality of Lake

Jonalyn G. Ebron, Rommel Ivan D. De Leon, Arviejhay D. Alejandro, and Basaron A. Amoranto

***Abstract*—In this study, the Multivariate Linear Regression (MLR), Artificial Neural Network (ANN), k-Nearest Neighbour (kNN), and Support Vector Machine (SVM) models had been developed to simulate and to predict the water quality of Laguna Lake. The input variables for the MLR model had been determined through linear regression. The ANN, kNN, and SVM had been modelled per water quality parameter with cross validation and evaluated through its accuracy. The performance of the MLR models had been evaluated with the statistical metrics R-squared, Mean Absolute Error, and Root Mean Square Error. A web-based water quality monitoring had been developed to incorporate in their monitoring.**

The results had indicated that the performance of SVM is superior in the prediction of classes in most water quality parameters. The study results had shown that the poor correlation between the water quality parameters indicated that the data cannot be modelled. The results had shown that the correlation had not reached the threshold to be significant of 60% for R-squared. As per the classification models, the results of the comparison had shown that SVM had been the best model in the majority of parameters.

***Index Terms*—Artificial neural network, classification, multivariate linear regression, machine learning, predictive model.**

I. Introduction

Water pollution has become the most concerned problem in many countries around the world. The assessment of long-term water quality changes is also a challenging problem. Water is a prime natural resource and precious national asset forms the chief constituent of the ecosystem [1]. It is one of the essential things in the world which is not only used by us humans but also all of the other living things around us. However, because of the continuous anthropogenic activities in bodies of water around the world, bodies of water are continually deteriorating and assisting in destroying the environment because of other chemicals that it contains. Thus, the classification of water quality is one of the most important ways to manage and monitor the quality of water resources.

Laguna de Baý or A lake in the Philippines is the largest lake and most import lake in the Philippines [2], [3]. The status of Laguna de Baý is deteriorating at alarming levels and dying. At present Laguna, de Baý currently classified as Class C. Laguna de Baý is suitable for fisheries and water tapping in parts of West Bay [3]. The agency responsible for the welfare and maintenance of lake used a spreadsheet in data representation, computation, and data repository.

At present, the water level is monitor on a monthly basis. As a result, volumes of data were gathered from nine major stations and recorded in different excel worksheet. Consequently, it takes much time in data processing and preparing the analysis. As a result, they cannot provide immediate information and preventive measures before making the status of the lake worse. Therefore, there is no way to forecast or anticipate dangerous siltation levels in the water.

After considering all categories of the interviews and researches and studies, the researcher realized that there is a need of a centralized system that they can use to share data for data analysis, processing, and visualization. Given the above scenario, a data-driven intelligence is a key to the next-generation organization, according to Sephora Asia's head of business intelligence. Nowadays, machine learning (ML), including ANN of different architectures and SVM, provides essential tools for intelligent geo and environmental data analysis, processing and visualization.

As described in different related studies different methods on monitoring and classifying the water quality of lakes, rivers or reservoir presented and most water-quality associated studies have computed the water quality index as a part of predicting or monitoring [4]-[6].

In this paper study aims to predict the classification of the water quality using MLR, ANN and SVM. For each model is made up of a number of the predictor, which are variables that are likely influence the future results. The model classifies the water quality of the lake based on the primary water quality parameters. Different model was formulated using the ANN, SVM, and KNN be to employ, mapped out by an application of artificial intelligence in the system software. As the historical data becomes trained and the model is validated.

A graphical user interface created for monitoring results in a statistical control process. These objectives serve as reference points for classifying the lake's status and its response over time to natural and anthropogenic influences. These objectives may in the future also enable the government to evaluate the impacts of development activities on the lake's water quality that serve as essential criteria for environmental planning and management.

One of the significant problems which are faced nowadays is how to handle big data and understand it and to create a model the data if there are too many or too few of them. The first objective is to design a predictive model to classify the water quality of the ten (10) significant stations. The model

Manuscript received September 15, 2019; revised July 23, 2020. This work was supported in part by Malayan Colleges Laguna, A Mapua School, Philippines.

The authors are with the College of Computer and Information Science at Malayan College Laguna, A Mapua School, Pulo-Diezmo Road, Cabuyao, Laguna, Philippines (e-mail: jgebron@mcl.edu.ph, riddeleon@live.mcl.edu.ph, arvhiejhay123@gmail.com, amorantobasaron101@gmail.com).

doi: 10.18178/ijesd.2020.11.9.1285

will help in data analysis, of the months and years data providing the ease of access and query optimization in preparing preventive measures and actions before the status of the lake worsens.

The second objective which is to design a map for visualization of data makes it easier for them to analyze the data collected because the raw data that raised and transformed into visual information that readily understood by the users. It can help to reduce the bulk of the data into a summarize value to express the data in a simplified and logical form through visualization, which simplify their extensive collection of data [5], [6].

While the third objective which is to develop a GUI that provide statistical process control of the reports of classification of water quality a big help for the users to have power to the data and also on what visualization of data they want to see.

II. Methodology

A. Methods Data Analysis Plan

The ten water quality parameters were considered in the study. BOD is the amount of oxygen microorganisms requires for stabilizing biologically decomposable organic matter under aerobic conditions in water. BOD is tested to identify the pollution load of water, degree of pollution, and the efficiency of wastewater treatments. BOD is monitored through five-day sampling of DO under 20 °C and is measured in mg/L. High BOD levels is associated with organic pollutions seen in sewages and excessive growth of algae and aquatic plants. BOD does not directly harm aquatic life because it is not a pollutant and only effects when levels reach a low point that are cause insufficient supply to aquatic life. Azide Modification (Dilution Technique) is the method analyzing. DO in the water is important for the survival of most aquatic organisms. Lack of it thereof, could be damaging to the respiration of aquatic organisms. Two main sources of DO are diffusion of oxygen from the air and photosynthetic activity. Azide Modification (Winkler Method) is used as the method for analyzing.

Color, measured in total color unit (TCU), is the amount of dissolved organic compound typically is ten times the organic carbon. It greatly affects absorption of color as opposed to suspended solids. It can be measured by comparator and colorimetric methods. Comparator methods rely on visual comparison of a water sample with a standard color solution or a set of colored filter disks. The most common comparator method involves matching a water sample with one of a series of dilutions of a standard color solution of platinum and cobalt chloride salts of molar ratio 2:1 where the platinum concentration in mg/L.

Chloride, in form of Cl- ion, one of the major inorganic anions, or negative ions, in freshwater and it's originated from of salts, like sodium chloride or calcium chloride. Possible sources of manmade salts that may contribute to increased chloride concentrations are chlorinated drinking water and sodium-chloride water softeners. Argentometric Titration is the method of analysing.

Nitrate is the most oxidized form of nitrogen and result of aerobic decomposition of organic nitrogenous matter and commonly occurs in small amounts in surfaces. Nitrates are essential plant nutrients, but an excessive amount can cause significant water quality problems and it also causes hypoxia, or low level of DO. Sources include agricultural run-offs, domestic, and industrial discharges and Sodium Salicylate Method is the method for analysing the nitrate. Phosphate is an essential nutrient for aquatic life occurs naturally in small amounts and high levels can lead to algal bloom and excessive nutrients in water. Phosphorus occurs in the forms of organic and organic. Common sources include agricultural runoff, animal waste and sewages. Ascorbic Acid Method is the method used to analyse.

Temperature measures the warmness and coldness of water body; it is important to assess because aquatic lifeforms are sensitive to changes in water temperature and changes naturally based on seasons. Mercury-filled temperature is used to measure the values. pH, measures the hydrogen ion concentration, indicates water whether acidic or basic. pH scales from 0 to 14 with pH seven as neural point, below seven as acidic and above seven is basic. The temperature is typically monitored for assessing in aquatic ecosystems, health, recreational waters, irrigation sources, and discharges, drinking water, industrial discharges, and flood water. Glass electrode method is used as the method for analysing.

Fecal Coliform presence in water bodies indicates that the water has been contaminated by feces from man or other animals. At the time of its occurrence, the water may have been contaminated by pathogens or disease producing bacteria or viruses and the method for monitoring Fecal Coliform is the Multiple Tube Fermentation Technique.

Total Suspended Solids are particles of silt, clay, sand, and any organic material including plankton that move with water. Measured in mg/L. Suspended solids reduced visibility and absorbs more light, which increases water temperature and reduce photosynthesis. Water with high suspended solids is unsatisfactory for bathing, industrial and other purposes. High levels could be damaging to aquatic life. Gravimetric method is the method used for analysing.

B. Water Quality Guidelines

This research used the 2009-2016 water quality historical data with nine (9) major monitoring stations. Data from the stations of the lake were preprocessed as the input for the modeling system. Table I shows the water quality guidelines prepared by the government agency for classifying freshwater.

TABLE I: Water Quality Guidelines for Primary Parameters

Water Quality Parameters	Unit	Water Body Classification				
		AA	A	B	C	D
BOD	mg/L	1	3	5	7	15
Chloride	mg/L	250	250	250	350	400
Color	TCU	5	50	50	75	150
DO	mg/L	5	5	5	5	2
Fecal Coliform	MPN /100mL	<1.1	<1.1	100	200	400
Nitrate	mg/L	7	7	7	7	15
Ph (Range)		6.5 -8.5	6.5-8.5	6.5-8.5	6.5-9.0	6.5-9.0
Phosphate	mg/L	<0.003	0.	0.5	0.5	5
Temperature	C^0	26-30	26-30	26-30	25-31	26-32
TSS	mg/L	25	50	65	80	100

Notes: MPN/100Ml – Most probable Number per 100 millilitre

C. Data Requirement

The dataset of water quality parameters based on the Water Quality Guidelines and General Effluent Standards by government agency was used in this study. The period of the dataset was covered from ranges between 2009 to 2016 and from the nine major stations. The monitored water quality values are quantitative data which is an advantage in the study for the development of the model. Each water quality parameter is represented by a unit of measurement. Since that every water quality parameter differs in unit, every water quality parameter is treated individually. The following water quality parameter and its units are seen in Table II below.

TABLE II: WATER QUALITY PARAMETERS UNITS

Water Quality Parameters	Unit
Biochemical Oxygen Demand(BOD)	mg/L
Chloride	mg/L
Dissolved Oxygen (DO)	mg/L
Fecal Coliform	MPN/100 mL
Nitrate	mg/L
pH	
Phosphate	mg/L
Temperature	C^0
Total Suspension Solids (TSS)	mg/L
Transparency	cm

1) Data collection

We used the 2009-2016 water quality historical data with nine (9) major monitoring stations. Nine years' data from the stations of the lake were preprocessed as the input for the modeling system.

2) Data preparation

During pre-processing in data preparation we followed three approaches. First, consolidating all data by year to a single sheet or flat file. After consolidating, the attributes parameter, year, and station were transformed in a numerical value to standardized all values of parameters.

In treatment of the missing values, the average value of the water quality parameter per station would be used for replacing the missing values of a water quality parameter in its respective station. For example, the average of BOD of Station I is 10, then all the missing values of BOD of Station I is replaced by 10. The researchers faced challenges in data preparation during the treatment of missing values because the data are in different numerical format and in different data format. The treatment of missing values was the most time-consuming procedure and requires meticulous attention because of the amount of data the researchers are working on. Another challenge we encountered is the difficulty of creating metadata in a csv format for dimensional model. To pre-process the data, we used the Jupyter Notebook running on Python 3.6.

In normalization process, the values per parameters and stations are normalized and transformed into an equivalent numerical value. The water quality parameters are replaced of the numerical values. The stations names are originally represented by roman numerals but are replaced by the corresponding numerical value. For example, a data set for pH, pH is replaced by the number 1 and the Station I to IX value are replaced by the equivalent numerical value of 1 to 10.

In handling of outliers, the researchers run a code in python. We plotted a histogram of the residuals, and then examine the normality of the residuals. As a result, even though is slightly skewed, but it not hugely deviated from being normal distribution, that means that the assumption is satisfied.

A dimensional modelling for database design was adapted for the database architecture in this study for faster retrieval of data and query optimization. The star schema is composed of four dimensions which are the *parameter, year, month* and *station*. The parameter dimension which contains a primary key id and non-primary key attributes, parameterName and the unit that is used for that parameter which is connected to the table rules which is comprised of the range for each class depending on the parameter. The fact table is then comprised of an id and the foreign keys *parameterID_id, yearID_id, monthID_id, stationID_id* and also the value for that parameter for a certain year, month and station.

3) Data understanding

The researchers used descriptive statistics to test the different model and data visualization in providing initial insight with the data. We also used the basic assumptions of the statistics to study trends in Python 3.0, with the library seaborn, contains tools that allowed the researchers to visualize the data. Most of the library needed for data analysis was provided by the library pandas used by the researchers for descriptive statistics. We calculated the mean, standard deviation, and variance in understanding the data.

D. Modelling

Modelling was performed by the researchers with the help of library sklearn from python 3.6, and used a module called Linear Regression. For evaluation of the models, the following statistical metrics were used which are R-squared (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). R^2 tests whether the model was able to explain the variability of the observations. RMSE is the square root difference between the predicted values and the observations values. MAE is the difference between two continuous values.

In the mentioned statistical metrics, R^2 score is used to see if the models can be used to predict [7]. R^2 score should exceed 60% in pure sciences studies for a model to be considered significant. RMSE is a tool for the researchers to tests the spread of the difference of the predicted values to the observed values. MAE is also used similarly to RMSE.

$$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2)][n\sum y^2 - (\sum y)^2]}} \quad (2)$$

n = number of pairs of variables x = independent t variable value y = dependent variable value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^{n}(P_i - O_i)}{n} \quad (4)$$

where P = Predicted, O = Observed and n = number of observations

E. Multivariate Linear Regression Model

Multivariate Linear Regression (MLR) is a method in regression analysis to attempt to model between two or more explanatory variables by fitting a linear equation to the observable data [8], [9]. In the research, there are two different multivariate linear regressions to perform. Each water quality parameter is modelled individually. In the first multivariate linear regression to be modelled, the variables include the dependent variable water quality parameter, with the predictive variables time and station. This MLR models the relation between time and water quality parameters values. As mentioned that each water quality is exclusively modelled, so in total ten models are prepared in this MLR.

In another MLR, the predictive variables, once again, includes ten of the water quality parameters, with the dependent variables, DO, Phosphate, and Chloride. A similar study of [31] modelled a predictive model of the same variables, these water quality parameters are often used as indicators of the eutrophication of a body of water. This model predicts the water quality parameter values with only the water quality parameter DO, Phosphate or Chloride.

The selection of an appropriate set of input variables for the MLR model is important for predicting the water quality variables in the lake [6]. To identify the input variables for predicting the dependent variables, the correlation of coefficient, R-score, of the paired variables must be at least 10%. All the water quality variables that has 10% R-score with a dependent variable will be used to predict the dependent variable in Fig. 1.

$$DO = 2.8083 \times pH + 0.0887 \times Chla - 0.074 \times WT$$
$$-0.0045 \times TP - 0.4928 \times NO_3 - 0.0055 \times BOD - 13.851$$

Fig. 1. Equation for predicting the dependent variable.

F. Artificial Neural Network Model

In this model, inputs are parameters, month, station and year from different stations. The hidden layer is unknown and will be constructed while training the ANN. Multilayer Perceptron Classifier (MPC) is used to perform neural networks. Backpropagation is the learning method used to compute for the errors in output based on the inputs assigned weights. The model process input data thru multiple parallel processing parameters, which do not store any data.

Supervised Learning is applied in the study to identify the class of the lake. In the study, historical data of water quality parameters treated as the dataset for the models. The dataset was split randomly by 70% training set and 30% testing set similarly to the procedure done from a similar study [10]. With the training set, cross validation was executed with machine learning algorithms. The model with the best accuracy was trained again, then the trained model was used to classify new data.

The dataset was split into two the training dataset that contains the 70% and the testing dataset that contains the 30 using the k-Fold Cross-Validation [2]. ANN, SVM and KNN are trained and tested using the k groups that are split in the training dataset and evaluate the score of each unique testing k group that it represents per iteration of k value. [4.] The procedure that came from of using k-Fold Cross-Validation for the model of the technique are as follows:

1) Shuffle the training dataset randomly
2) Split the training dataset into k groups based on k value
3) For each unique group (All groups must be able to experience being a testing dataset per iteration)
a) Choose a group that will represent a testing dataset
b) After choosing, the remaining groups will be training datasets
c) Train the model of the technique using the training datasets and test the model of the technique using the testing dataset
d) Compute for the evaluation score based on the result of the testing
4) Summarized and Combine all the evaluation scores of all groups to be able to get the average evaluation score of the technique

$$Meanaccuracy = \frac{\sum_{i=1}^{n} Accuracy_i}{N} \quad (5)$$

For configuring the k value, [4] stated that the *k* value must be chosen carefully for the dataset or it may result of a wrong idea of the skill of the model of the technique such as a score with a high variance that can change a lot according to the data used to fit the model of the technique or a high bias such as overestimate of the skill of the model of the technique if the k value was poorly chosen. In choosing ten as a k value means that the result of the skill of the model is estimate with a low bias or a modest variance, it is common in the field of the applied machine learning and it is recommended.

G. Classification Models Assessment

When a model is trained, it is evaluated using Accuracy using the equation in Fig. 2. The trained model with the best accuracy is usually favoured, since it was able to adapt well to new data. Based on the freshwater classification, it is crucially important to evaluate the model, in this paper, to test how accurately the three model predicts water quality. The researchers do so by trying out the model predictors for water quality for where they already know the true results. Researchers are very careful not to do testing on data that has already been used for training otherwise, it ends up over-estimating the prediction quality. In this paper, the training data was divided into 10 number of iterations using cross validation, which was measured based on performance in each iteration and then assessed the performance of the mean performance of the iterations.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Fig. 2. Equation for accuracy valuation of the model.

The classification of the water quality parameters and its water quality index was implemented in the web-based application using Python and Django which is used for

creating web-based application and the monitoring of these parameters. Determining the value of the model is dependent on the parameters used. Linear models were produced to see the correlation between time and station to the dependent variable water quality parameter value. The performance of the model with the total of 864 observations, with 604 observations in the training set and 0 with the testing set are evaluated and compared as seen in Fig. 3. The predicted value of the model is compared with the observed values using statistical methods to test the applicability of the model for the study. These statistical methods include the R-squared, RMSE, and MAE.

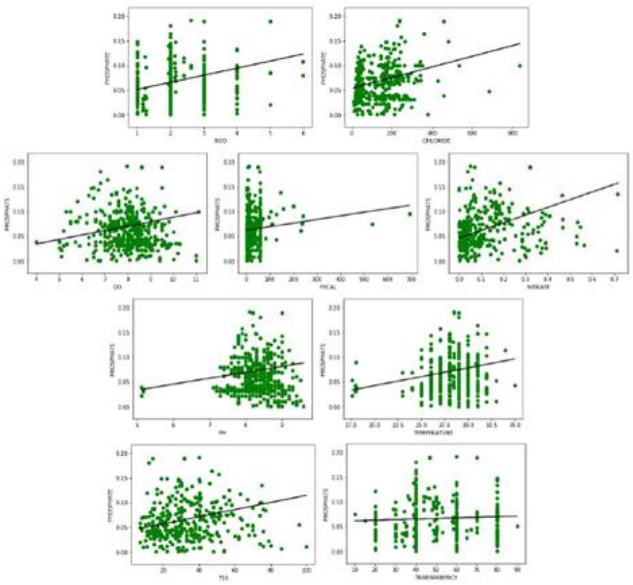

Fig. 3. Graph of the relationship of the class to water quality parameters.

The results from the training phase indicates that the R2 is observed to be least in TSS, which got a very low value of 0.0001%, for all the periods and at all the monitoring stations. A good R2 is noted for pH and Temperature which got 24.26% and 24.05% respectively, at all monitoring stations and for all s4gvtudy periods. Another water quality parameter to be noted is Transparency which achieves a R2 of 12.67%. With results from the testing phase, the least R2 is indicated in TSS which scored -0.22%. The highest observed R2 resulted in temperature which achieves a score of 17.59%.

III. Results and Discussion

After data cleaning, K-means clustering algorithm was applied to the data set using python programing language. The researchers create numerical models for prediction and develop the web based system. The researcher spends more time and effort on data cleaning process which is the most critical process for the success of any machine learning function in identifying the input variables. Linear Regression was used for getting the correlation between two parameters for data cleaning using MLR, ANN and SVM in creating a predictive model.

A summary of the observed result is seen in Table III. BOD got a mean of 2.19, a standard deviation of 1.03, and variance of 1.06. Based on the collected data of BOD, the standard deviation of the dataset is decently spread. Chloride scored 131.22 in mean and 104.33 in standard deviation. The deviation of this parameter is highly spread-out.

TABLE III: Statistical Summary of Parameters

Parameter (Unit)	Mean	SD	Variance	Min	Max
BOD(mg/L)	2.19 mg/L	1.03	1.06	1.0 mg/L	6.0 mg/L
Chloride(mg/L)	131.22 mg/L	104.33	1088.77	8.0 mg/L	642.0 mg/L
DO (mg/L)	7.87 mg/L	0.94	0.88	5.3 mg/L	11.6 mg/L
Fecal Coliform(mL)	59.69 mg/L	59.71	3565.46	2.0 mg/L	241.0 mL
Nitrate(mg/L)	0.150 mg/L	0.109	0.012	0.001 mg/L	0.715 mg/L
pH(range)	8.04	0.73	0.53	6.8	10.0
Phosphate(mg/L)	0.073 mg/L	0.036	0.001	0.001 mg/L	0.2 mg/L
Temperature(C^0)	27.75 C^0	2.41	5.81	22.0 C^0	36.0 C^0
TSS (mg/L)	35.64 mg/L	15.89	252.49	7.0 mg/L	117.0 mg/L
Transparency(cm)	51.13 cm	17.06	289.20	10.0 cm	100 cm

With DO, it obtains a mean of 7.87 a deviation of .94. An example of a low spread of values is seen in this parameter as what its standard deviation exhibits.

Fecal Coliform have a mean of 59.69 and a standard deviation of 59.71. Its standard deviation is high relative to the dataset's mean. Nitrate garners a mean of 0.15 and a standard deviation of 0.11. Its standard deviation is high in relates to its mean.

pH collected a mean of 8.04, a standard deviation of 0.73. The standard deviation of this water quality parameter is lowly spread-out.

Phosphate mean is .07 and got a deviation of 0.04. The deviation of the data for phosphate is decently spread. For the water quality parameter

Temperature, a mean of 27.75 and a standard deviation of 2.41 are calculated. The deviation of the data for this parameter is lowly spread.

For TSS a mean of 35.62 and a standard deviation of 15.89. The standard deviation of the parameter is evenly spread. Transparency collected a 51.13 in mean and 17.06 in standard deviation. Another to take note of is the water quality parameter pH which scored 12.29% as presented in Table III.

The performance evaluation using Linear Regression were presented in table below.

TABLE IV: Performance Evaluation Using Linear Regression over Time

Parameter (Unit)	Training			Training		
	R^2	RMSE	MAE	R^2	RMSE	MAE
BOD(mg/L)	0.20	1.05	0.70	0.65	0.93	0.62
Chloride(mg/L)	7.03	95.13	63.63	9.76	86.17	65.20
DO (mg/L)	11.66	1.28	1.06	7.67	1.16	0.96
Fecal Coliform (mL)	7.04	104.04	58.86	7.42	103.10	57.26
Nitrate(mg/L)	0.40	0.10	0.08	0.97	0.09	0.07
pH(range)	24.26	1.11	0.91	12.29	1.08	0.86
Phosphate(mg/L)	0.53	0.04	0.03	3.73	0.03	0.03
Temperature(C^0)	24.05	3.81	3.24	17.59	3.57	2.99
TSS (mg/L)	0.0001	15.41	12.22	0.22	15.76	11.90
Transparency(cm)	12.67	13.85	11.04	3.55	14.11	11.07

Criteria of R^2 > 60% to be significant

The R2, RMSE, and MAE for BOD in the training phase are 0.20%, 1.0505 mg/L and 0.6967 mg/L respectively.

During the testing phase BOD achieves a result of -0.65%, 0.9338 mg/L, and 0.6232 mg/L. The RMSE and MAE from the training phase is slightly higher in comparison with the RMSE and MAE of testing phase.

With Chloride the R2, RMSE, and MAE resulted in 7.03%, 95.1313 mg/L, and 63.6349 mg/L. The testing phase scored 9.76%, 86.1719 mg/L, and 65.2029 mg/L. A small difference of R2 is evident with this water quality parameter. In the water quality parameter

DO training phase, the R2 is 11.66%, RMSE shows 1.2851 mg/L, and 1.0555 mg/L MAE. The testing phase resulted in 7.67% R2, 1.1628 mg/L RMSE, and 0.9642 mg/L MAE. A small difference among the statistical metrics between the training and testing is again evident in this water quality parameter. As indicated in the results from the water quality parameter

Fecal coliform, the training phase achieved a R2 7.04%, RMSE of 104.0455 mL, and MAE of 58.8645 mL. The testing phase resulted in 7.42%, 103.0997 mL, and 57.2664 mL.

Nitrate training phase resulted in R2 of 0.4%, RMSE of 0.1012 mg/L, and MAE of 0.0773 mg/L. The testing phase results shows that the R2 is 0.97%, RMSE is 0.0954 mg/L, and MAE 0.0723 mg/L.

pH results the highest R2 score in the training phase among the ten water quality parameters. In its training phase, a R2 24.26% is scored, a RMSE of 1.1103, and 0.9148 MAE. The testing phase scored R2 of 12.29%, RMSE of 1.0811, and MAE of 0.8646.

Phosphate training phase scored a low R2 score of 0.53%, which explains its high RMSE of 0.0369 mg/L and MAE 0.0306 mg/L. Similar goes with it testing phase which resulted in -3.73% R2, 0.0392 mg/L RMSE, and 0.0332 mg/L MAE.

Temperature training phase achieve the second highest R2, which is not too far off compare with the highest. A R2 of 24.05%, RMSE of 3.8074 C°, and MAE of 3.2403 C° resulted in the training phase. The testing phase results include a R2 of 17.59%, RMSE of 3.5708 C°, and MAE of 2.989 C°. TSS training phase receives a very low R2 of 0.0001%, its RMSE 15.4075 mg/L, and MAE of 12.2201 mg/L. Testing phase achieve a R2 -0.22%, RMSE of 15.7603 mg/L, and MAE of 11.9074 mg/L.

Lastly, transparency training phase results with the R2 of 12.67%, RMSE of 13.8552 cm, and MAE of 11.0435 cm. Its testing phase achieved a score of 3.55% R2, 14.1114 cm RMSE, and 11.0762 cm MAE. The results from the training phase and testing phase shows that the R2 is higher during the training phase. Although the difference is not that huge in water quality parameters with low R2 in the training phase. As evident in the training phase and testing phase of the water quality parameters which are pH and Temperature, a drop-off from its R2 score can be observed in its testing phase. This indicates that the model over-fits with the training set, wherein unseen observations are not explained well by the model.

It is also evident, with the evaluation of the performance of the model, that the R2 score in all water quality parameters were not able to achieve a significant enough score of 60% or 0.60. R2 ranges from 0.0001% (TSS) to 24.26% (pH) in training phase.

Therefore, none of this water quality parameters can be used as a predictive model of its respective water quality parameters. Using time as the predictor variable resulted in a low correlation with the water quality parameters, the researchers turned in to the methodology of another study by [6] the MLR models were used to predict the values of a parameter with independent variables to year or with DO, Chloride, and Phosphate. The multivariate regression models in Equation 1, were obtained from the input water quality variables

The result of the evaluation of performance of cross validation of the three-classification algorithm. The results show that independence when we have longitudinal dataset. Longitudinal dataset is one where we collect observations from the same entity over time. SVM got the most count of the best model. Some of the water quality parameters are not legible to model SVM due the restriction that the dataset of the specified water quality parameter only contains one class. Individually ANN was not able to be the best model. Hidden layers are factors and the nature of the defined classes could have an effect on the layers of ANN. kNN, with a count of 2 best models, produces near perfect results too considering that these water quality parameters contain classes. In the evaluation of the correlation of time and water quality parameters, the results showed that most of the water quality parameters present a weak correlation to time.

Based on the models' R-squared score with the testing phase, the highest, temperature, attain a score of 17.59%, and succeeding is pH with 12.29% score. These small R-squared scores are too low to produce a linear model that can predict with low error. An acceptable R-squared score in science research must be above 60%. The variability of the data, as seen in the plots of correlation of water quality parameters and time, indicates that model strive to generalize the data. Therefore, with the dataset present, predicting with water quality parameter based on time would not be definite. The results indicate that monitored data could possibly be unreliable or the inconsistency nature of the water quality parameters.

IV. Conclusions

ANN models, kNN, and SVM were used to predict the classes of the lake. The water quality variables predicted using the kNN model did yield a good result with a low mean accuracy result which shows that a low percentage of the predicted value did not match the actual values from the test dataset. In general, the SVM model better predicts in more water quality variables than the kNN and ANN models. Although SVM were not able to be modelled in some water quality parameters due to singular available class in the dataset, it still got the highest number of best mean accuracy. The ANN model, including the kNN and SVM model, can preserve the nonlinear characteristics between the input and output variables, which are superior to conventional statistical approaches. Overall with the ten water quality parameters, SVM is the best model to perform at the most.

In the real world, temporal and spatial distributions in observational data do not exhibit simple regularities; therefore, they are difficult to accurately predict. It is necessary to use nonlinear models, such as the ANN model,

which are suitable for complex nonlinear systems. The proposed approach using the ANN model has yielded valuable information that can be used by decision-makers for aiding reservoir water quality management. In the present study, we focus on the prediction of water quality instead of forecast. improve the prediction of classification of water quality in the lake.

AUTHOR CONTRIBUTIONS

J. G. Ebron was the over-all in-charge of the methods, design and functionalities needed for the training the data.

I.R.D.D was in-charge of the coding and data analysis in Python.

A. D. Alejandro was in-charge in the review of related literature, data collection and testing.

B. A. Amoranto was in-charge of the development of the web-based system.

ACKNOWLEDGMENT

The authors would like to thank Malayan Colleges Laguna, A Mapua School for funding the presentation and publication of this research. Likewise, to the College of Computer and Information Science for all the support. Also to their family and friends on unparalleled support and motivation.

REFERENCES

[1] S. Ewaid and S. Abed, "Water quality index for Al-Gharraf River, souther Iraq," *Egyptian Journal of Aquatic Research*, vol. 6, 2017.
[2] A. S. Borja and D. N. Nepomuceno, *Laguna de Bay. Pasig: A Lake in the Philippines Development Authority*, 2006.
[3] S. H. Ewaid, S. A. Abed, and S. A. Kadhum, "Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis," *Environmental Technology & Innovation*, pp. 391-398, 2018.
[4] J. Brownlee. (May 2018). *k-Fold Cross-Validation. Retrieved from Machine Learning Mastery*. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/
[5] S. H. Ewaid, S. A. Abed, and S. A. Kadhum, "Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis," *Environmental Technology & Innovation*, pp. 391-398, 2018.
[6] B. Wei and C. Wen, *Water Quality Modeling in Reservoirs Using Multivariate Linear Regression and Two Neural Network Models.* Hindawi Publishing, 2015, pp. 1-12.
[7] J. F. Hair, M. Sarstedt, L. Hopkins, and V. Kuppelwieser, *Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results*, 2013.
[8] N. Cameron, "Sephora Asia details its journey to data-driven decision making," *CMO Australia*, 2018, p. 1.
[9] W. Thoe and J. H. Lee, "Daily forecasting of Hong Kong beach water quality by multiple linear regression models," *Journal of Environmental Engineering.*
[10] M. Khadr and M. Elshemy, "Data-driven modeling for water quality prediction case study: The drains system associated with Manzala Lake," *Egypt. Ain Shams Engineering Journal*, pp. 549-557, 2016.

Jonalyn G. Ebron is a professor in the College of Computer and Information Science at the Malayan Colleges Laguna – A Mapua School. She is currently the program chair for BS Computer Science at Malayan Colleges Laguna – A Mapua School. She already completed her academic requirements in Ph.D in computer science and currently doing her dissertation in the field of machine learning. She also holds a MS degree in computer science and a BS in computer science. She got her Licensure Examination for Teacher (LET) and license to teach in the Philippines. She's a certified IBM instructor for business analytics/business intelligence. She got her diploma as foreign technical graduate in database management and network administration by the Authorization Committee of Nagano Prefecture at Information Technology Research Center (ITRC), Japan.

Rommel Ivan D. De Leon is a fourth-year bachelor of science in computer science student from Malayan Colleges Laguna, A Mapua School. In the four years of his college, he has gain experiences with using different programming language. He's intermediate in C++, C#, and Java are among those programming languages and he's also skillful in handling database in MySQL environment. He has beginner level skills with python and mobile development, and he is willing to learn and takes failures for learning.

Arviejhay D. Alejandro is a fourth year student taking bachelor of science in computer science from Malayan Colleges Laguna, A Mapua School. In his experiences in the past four years, he has gain experiences on programming different applications using four different programming languages which are C++, C#, Java and python. Among these four programming languages that he learned, he has the most experience in programming using C#. He has a novice skill level with assembly which is a low-level programming language. He had gain work experience on his On-Job Training at Gleent Incorporated on full-stack web development.

Basaron A. Amoranto is a fourth-year student from Malayan Colleges Laguna, A Mapua School. currently taking bachelor of science in computer science. He had learned and improved his experience on software and web development over these past four years in his college life using various of programming languages and web frameworks.