Journal Home page: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI :10.21474/IJAR01/18592
DOI URL : http://dx.doi.org/10.21474/IJAR01/18592

RESEARCH ARTICLE

WATER QUALITY INDEX OF LAKE NOKOUÉ PREDICTION USING RANDOM FOREST AND ARTIFICIAL NEURAL NETWORK

N. Dabire[1,2], E.C. Ezin[3] and A.M. Firmin[1,4]

1. Institut National de l'Eau (INE), Centre d'Excellence d'Afrique Pour l'Eau et l'Assainissement (C2EA), Université d'Abomey Calavi (UAC).
2. Ecole Doctorale des Sciences de l'Ingénieur (ED-SDI), Université d'Abomey Calavi.
3. Institut de Formation et de Recherche en Informatique (IFRI), Université d'Abomey Calavi.
4. Laboratoire d'Hydrologie Appliquée (LHA), Université d'Abomey Calavi.

Manuscript Info

Manuscript History
Received: 26 February 2024
Final Accepted: 30 March 2024
Published: April 2024

Key words: -
Benin, Lake Nokoué, Machine Learning Algorithms, Surface Water Quality, Surface Water Quality Index

Abstract

Poor water quality is a serious problem in the world which threatens human health, ecosystems, plant and animal life. Prediction of surface water quality is a main concern in water resource and environmental systems.The lake Nokoué in Benin, the country's main water body, is experiencing without equivox pollution. Usually, the water quality index is evaluated manually by complex mathematical formulas with major risks of errors. This study discusses the development and validation of a Random Forest and Artificial Neural Network (ANN) model in estimating water quality index (WQI) in the lake Nokoué.The two models have been developed and tested using data from 20 monitoring stations over a period of 12 months. The modeling data was divided into two sets. For the first set, RF and ANN were trained, tested and validated using 12 physical-parameters as input parameters. A detailed comparison of the overall performance showed that prediction of the random forest (RF) model was better than artificial neural networks with coefficient of correlation (R^2)=0.98, root mean squared error (RMSE)=0.12, explained variance score (EVS)=0.98 and mean absolute error (MAE)= 0.14 at training phase while and at the validation phase their values are0.80, 0.19, 0.23, 0.74 respectively which demonstrates that RF is capable of estimating WQI with acceptable accuracy. This method simplifies the calculation of the WQI and reduce substantial efforts and time by optimizing the computations. This will help in taking appropriate preventive measures to control the water quality of lake Nokoué through associated chemical treatments.

Copy Right, IJAR, 2024, All rights reserved.

Introduction:-

Water is an essential natural resource whose physical and chemical quality is the foundation of the ecosystem(Dovonouet al. 2011).According toLalèyè et al.(2004)a good chemical and ecological status of surface water body is a major concern for a society that has to meet increasingly important water needs. Thus, water resources are a major concern in West Africa countries because they are absolutely essential for the development of human, economic and social activities.Lake Nokoué in Benin plays a very important role in socio-economic

Corresponding Author:- N. Dabire
Address:- Institut National de l'Eau (INE), Centre d'Excellence d'Afrique pour l'Eau et l'Assainissement (C2EA), Université d'Abomey Calavi (UAC).

development at the local, regional and national levels (Avahouin et al. 2018). However, it is subject to numerous environmental pressures related to hydrocarbon traffic, artisanal fishing, household wastewater discharges from the communes bordering lake Nokoué.In addition to salt water inputs from the Atlantic Ocean, and inputs of pesticide and fertilizer residues by leaching from soils subjected to intense and diversified agriculture (Zandagba et al. 2016). Lake Nokoué, like all bodies of water, hosts the largest lake villages with a galloping demographic growth. No doubt, this leads a strong enthronization and consequently chemical pollution of the water of lake Nokoué, an increasing eutrophication compromising biodiversity and promoting the proliferation of invasive plants such as the water hyacinth (Mama 2010 ; Dovonou et al. 2011; Zandagba et al. 2016). Thus, to ensure its good management within the framework of sustainable development, it is judicious to make a permanent follow-up of the medium and long term evolution of the qualitative physicochemical state of the water (Sèdami & Bokossa 2016). Ensuring freshwater quality appropriate to human and ecological needs is therefore an important aspect of integrated environmental management and sustainable development. In terms of environmental and ecological problems, the number of water quality parameters is quite extensive. Hence, a robust mathematical technique is required to combine the physicochemical characterization of water into a single variable which describes the water quality.A water quality index (WQI) is a single number which uses a set of physicochemical water parameters to express the quality of water at a certain place and time(Vasistha & Ganguly 2020). This method has been initially proposed by (Horton, 1965)and (Brown et al. 1972). To compute this index, (Horton1965)proposed the first formula that takes into account all parameters needed to determine surface water quality and reflects the composite influence of different parameters important for water quality assessment and management(Liou et al. 2004; Tyagi et al. 2013). This index was used for the first time to highlight the physical-chemical changes that may occur during the year (House 1990; House & Ellis 1987).Based on this parameter, the water quality was classified into five different classes according to the water's suitability for various uses such as water supplies, irrigation, and fish culture. The conventional method suggested by Horton requires lengthy transformations to estimate subindices. In addition, the subindices required the inclusion of different equations, which need lengthy effort and time to calculate the final WQI.Therefore, estimation of such a WQI is cumbersome and can lead to occasional mistakes. The management of water pollution in lake Nokoué poses a significant challenge because there is no reliable predictive model for the water quality index of the lake. This is due to the fact that physically-based conceptual models struggle to learn from physical-chemical parameter data to accurately estimate the water quality index of lake Nokoué. However, the random forest (RF) and artificial neural networks (ANN) can be suggested as alternatives for estimation of WQI, as both employ the raw data instead of subindices. The performance of machine learning models to enhance water quality and reduce a wide range of wastewater was reported in several studies (Iorliam et al. 202; Kalhori & Zeng 2013; Wang 2017; Zheng 2018).Since the number of variables which affect water quality is too high, recently machine learning techniques such as RF, geneticprogramming (GP), and ANN have been successfullyemployed to solve the problems related to engineering in hydrology(Sampurno et al. 2022).The RF is proposed as a leading technique which can be used for regression and classification purposes (Maier & Dandy 2000 ; Zhu et al. 2022).The RF has high ability for generalization and is less prone to overfitting. Furthermore, it simultaneously minimizes the estimation of error and model dimensions. Huangapplied the RF and ANNs for prediction of water level in rivers. Amir et al. (2018) recommended the RF as the appropriate tool to forecast lake water levels and obtained quite acceptable results. The ANNs have been recommended as an effective tool for the prediction of water pollution and water quality (Abobakr et al. 2019 ; Liou et al., 2004).The ANNs are a useful technique that was used to speed up the calculation of water quality index in rivers (Dahal et al. 2021; Tyagi et al. 2013).The main objective of this study is to develop a decision support tool based on the prediction of the water quality index of lake Nokoué for pollution management.In this research, both RF and ANN were used as robust techniques for rapid and direct prediction of the WQI in the lake Nokoué which can be used as another alternative for some long-lasting conventional methods. Twenty points in the wetland were monitored twice a month over a period of 12 months and an extensive dataset was collected for 12physico-parameters. Finally, the RF result was compared with neural networks models.

Materials and Methods:-

Study Area, Sample Collectionand Analyses

Located in the South-East of the Republic of Benin (6º 25'N, 2º 26'E) Figure 1, lake Nokoué has a surface area varying from 150-170 km²between the low water period and the high-water period (Gnohossou 2006; Mama 2010; Dehotin et al. 2007). The lake Nokoué measures 20 km in the East-West and 11 km in the North-South direction. In the east, lake Nokoué is linked to Porto-Novo Lagoon and forms a freshwater lake with a surface area of about 180 km^2. lake Nokoué is connected with the Atlantic Ocean by a channel named Cotonou channel, which has a total length of 4.5 km. The hydrological regime of lake Nokoué is characterized by a low flood from May to June, which is the main rainy season in southern Benin and a major flood from September to November because of water supply

from the Ouémé river. The depth of the lake is between 0.3 and 3.4m (Adandedji et al. 2022). The average depth of the lake is 1.3m according to a bathymetric study reported in the article by Mama et al. (2011).

Sampling locations were selected according to three criteria: i. good geographical distribution of stations through the complex lagoon; ii. shrimp and oysters fishing locations; iii. lake village and residential house's locations of the complex lagoon.A total of twenty stations numbered from(St1 to St20) surrounding the complex lagoon were chosen.The sampling, preservation, and analysis procedures were in accordance with the national guidelines for surface water monitoring from the Applied Hydrology Laboratory (LHA). Twenty water samples are collected per month over a three-year period from 2016 to 2018.The data collection was carried out twice a month over a period of 12 months.Initially, these physical-chemical parameters were selected based on the standards set by the World Health Organization (OMS) for monitoring the quality of surface water. Totally, 12 physical-parameters were collected in the lake Nokoué , including the physical-chemical parameters still called sub-indicators such as oxygenation state (quantity of oxygen present in the water in dissolved form and available for aquatic life and the oxidation of organic matter O2_dissolved), temperature (T), electrical conductivity (EC), hydrogen potential (PH), organic pollution, turbidity, nutrient load (pollutants responsible for eutrophication phenomena), chemical oxygen demand (COD), suspended solids (SS), total nitrogen, salinity, nitrite and nitrate. The in-situ parameters such as temperature, hydrogen potential, electrical conductivity, salinity, total dissolved solids and dissolved oxygen were measured using a multi-parameter instrument (AQUAREAD AP-700) and suspended solids with a DR-890 colorimeter. Laboratory analyses were performed with a DR-2800 spectrometer to determine nitrite, nitrate and ortho-phosphate. The geology of lake Nokoué is characterized by a combination of sand, muddy sand, and layers of mud.

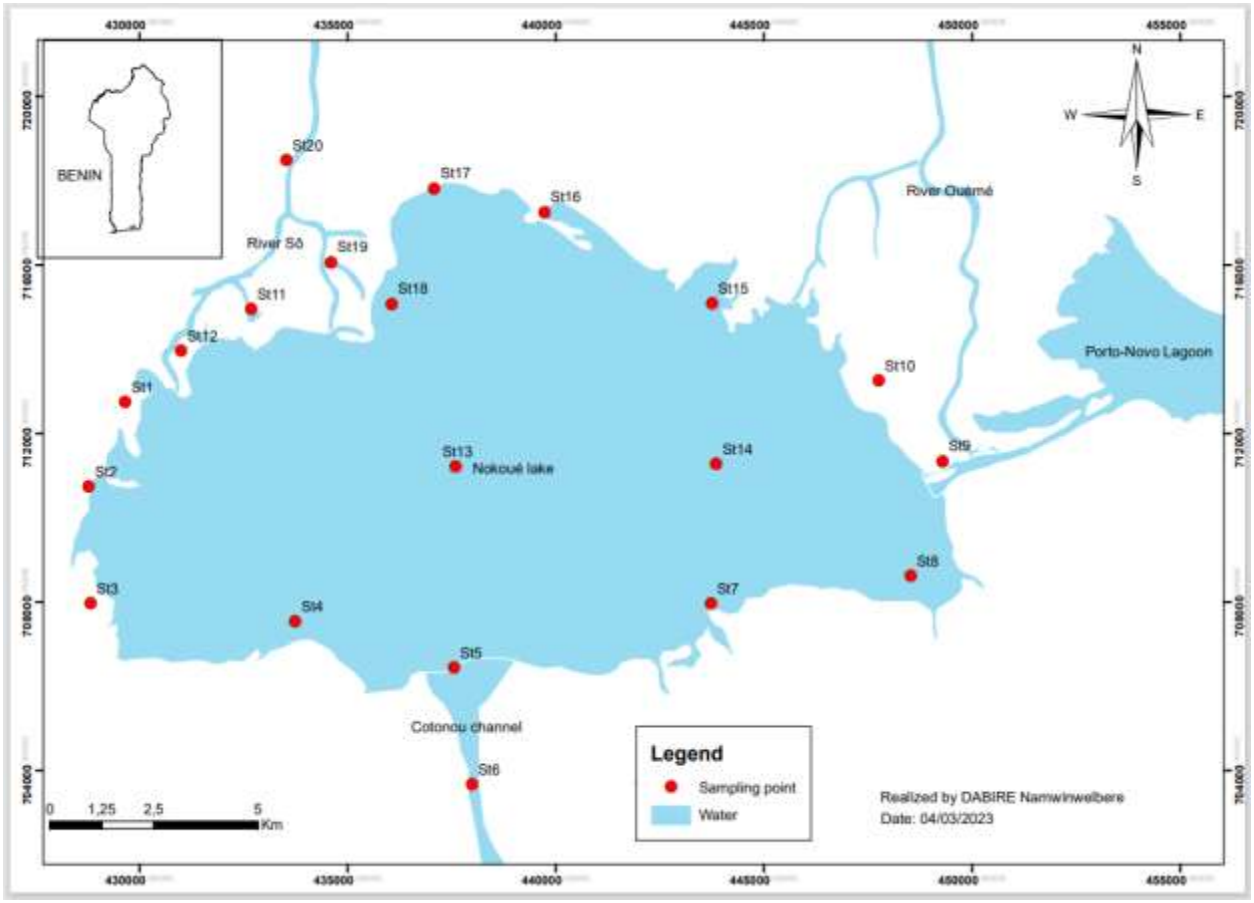


Figure 1:- Map of water sample collection points.

Water Quality Index computing

The tedious mathematical formulas proposed by the authors Horton (1965), Brown etal. (1970), Brown etal. (1972), Chatterji and Raziuddin (2002), Yidana and Yidana (2010) are usually used to evaluate the surface water quality

index. These formulas take into account all the parameters needed to determine surface water quality and reflects the composite influence of different parameters important for the assessment and management water quality.In this study this index is described by the key physicochemical parameters (called sub-indicators) describing potentialhydrogen(pH), dissolved oxygen (DO), salinity, Electrical Conductivity (EC), temperature (T°C), Suspended Solids (MES), turbidity, Total Dissolved Solids (TDS), Chemical Oxygen Demand (COD), Ortho phosphorus, Nitrites, and Nitrates (pollutants responsible for eutrophication).In this approach, a numerical value called relative weight (W_i), specific to each physicochemical parameter, is calculated according to the following formula, showing in equation (1):

$$W_i = \frac{K}{S_i} \quad (1)$$

Where K is a constant of proportionality and S_i in the column two in the table 1 is a maximum value threshold of the standard for surface water of each parameter in milligram per liter(mg/l) except for pH, T°C and electrical conductivity. Those thresholds are set by standard OMS (Balan et al., 2012; Firminetal., 2018). The unit of measurement of turbidity is the NephelometricUnit of Turbidity (NTU).

K can also be calculated using the following equation (2):

$$K = \frac{1}{\sum_{i=1}^{n}\left(\frac{1}{S_i}\right)} (2)$$

Where n is the number of parameters.

Then, a quality rating scale (Q_i) is calculated for each parameter by dividing each parameter concentration by the standard maximum value for that parameter and multiplying the whole by 100 as in the following formula giving in equation (3):

$$Q_i = \left(\frac{C_i}{S_i}\right) \times 100 \ (3)$$

Where Q_i is quality assessment scale for each parameter and $C_i$is the concentration of each parameter in mg/l.

Finally, the overall water quality index is calculated by the following in equation (4):

$$WQI = \frac{\sum_{i=1}^{n} W_i \times Q_i}{\sum_{i=1}^{n} W_i} (4)$$

where Q_i represents the quality rating scale for each parameter, W_i is a numerical value called relative weight and $w_i \times Q_i$is the sub-index value specific to each physical-chemical parameter retained. Five quality classes can be identified according to the values of the water quality index table 2.

Table 1:- Sub-index calculation for each value WQI.

	S_i	$1/S_i$	C_i	K	W_i	Q_i	$W_i \times Q_i$	WQI
T (°C)	25	0.04	29.8		0.0324	119.2	3.85917	
Turbidity (NTU)	5	0.2	5		0.1619	100	16.1878	
Conductivity (µS/cm)	1000	0.001	7.5		0.0008	0.75	0.00061	
SS (mg/l)	50	0.02	1		0.0162	2	0.03238	
TDS (mg/l)	500	0.002	96		0.0016	19.2	0.03108	
Salinity(mg/l)	150	0.00667	0		0.0054	0	0	
O_2 Dissolved(mg/l)	100	0.01	0.62		0.0081	0.62	0.00502	
PH	9	0.11111	6.94		0.0899	77.1111	6.93476	
Ortho-P(mg/l)	2	0.5	0.218		0.4047	10.9	4.41117	
Nitrite (mg/l)	3.2	0.3125	0.013		0.2529	0.40625	0.10275	
Nitrate(mg/l)	45	0.02222	2.9		0.018	6.44444	0.11591	
DCO (mg/l)	100	0.01	57.12		0.0081	57.12	0.46232	
Total		1.2355		0.8094	1		32.1429	32.143

Table 2:- Standard WQI classification and status (Brown et al. 1972; Aher et al. 2016).

Classes	Water status	Possible use
0-25	Excellent quality	Potable water, irrigation and industry
26 - 50	Good quality	Drinking water, Irrigation and industry
51 - 75	Poor quality	Irrigation and industry
76 - 100	Very poor quality	Irrigation
> 100	Unsuitable	Appropriate treatment required before use

Data preprocessing

There are two techniques for preprocessing data before training machine learning models: management of outliers and scaling of variables. The management of outlier'svalues answers the following equation (5):

$$\text{Vab} = \begin{cases} < Q_1 - 1{,}5 \times IQ \\ > Q_3 + 1{,}5 \times IQ \end{cases} \quad (5)$$

where Q_1 is the first quartile, Q_3 is the third quartile and (IQ) is the inter-quartile range. If a value in the series does not fall within these ranges, it is considered an outlier.

Feature scaling is therefore a very important step to take care before training a machine learning model (Dahal et al. 2021). The data for machine learning in this study were preprocessed to reduce the impact on the accuracy of the water quality index prediction caused by the large order of magnitude difference between the water quality index data and the physical-chemical parameters (influencing factors). There are many methods of scaling features, the most common and popular techniques used in the machine learning community are normalization also namelystandardization. The standardization equation is formulated in three steps: (1) calculate the mean and standard deviation (std), (2) subtract the mean (X_{mean}) from each explanatory variable to be standardized (X_i), (3) the previous result is divided by the standard deviation. Analyzes and modeling were done with python language.
Its mathematical expression is the following:

$$X_{\text{normalized}} = \frac{X_i - X_{\text{mean}}}{\text{std}} \quad (6)$$

The data base used has a dimension of 48 rows and 12 columns. All variables (physical-chemical parameters) were considered to build the two models.

Scenario construction

On the basis of the correlation matrix, twelve scenarios recorded in the table 3were constructed based on decreasing coefficients to better understand the sensitivity of each physico-chemical parameter on the water quality index of lake Nokoué.

Table 3:- The scenarios built based on decreasing correlation coefficients.

Scenarios	Physicochemical parameters											
S1	Tur											
S2	Tur	MES										
S3	Tur	MES	Ortoph									
S4	Tur	MES	Ortoph	NO_2								
S5	Tur	MES	Ortoph	NO_2	NO_3							
S6	Tur	MES	Ortoph	NO_2	NO_3	PH						
S7	Tur	MES	Ortoph	NO_2	NO_3	PH	OD					
S8	Tur	MES	Ortoph	NO_2	NO_3	PH	OD	DCO				
S9	Tur	MES	Ortoph	NO_2	NO_3	PH	OD	DCO	TSD			
S10	Tur	MES	Ortoph	NO_2	NO_3	PH	OD	DCO	TSD	Temp		
S11	Tur	MES	Ortoph	NO_2	NO_3	PH	OD	DCO	TSD	Temp	Sal	
S12	Tur	MES	Ortoph	NO_2	NO_3	PH	OD	DCO	TSD	Temp	Sal	CE

Structure of random forest and artificial neural network

Machine learning is a set of statistical methods for analyzing trends, finding relationships, and developing models to make predictions about a data set. In this study, we use the random forest and neural networks technique to predict water quality index of lake Nokoué automatically.

Structure of random forest

Decision tree forests, also known as Random Forest, are a set learning technique that uses decision trees to build decision support models (Dalai et al. 2021). They use split data from historical data by randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode for all predictions in each decision tree. This method reduces the risk of error in an individual tree by relying on a majority prevalence model (i.e., where the majority prevails). For example, if we create a random forest with four (04) decision trees, the third decision tree below will predict zero (0), but if we rely on the mode of the four decision trees, the predicted value will be one (01) as shown inFigure 6.

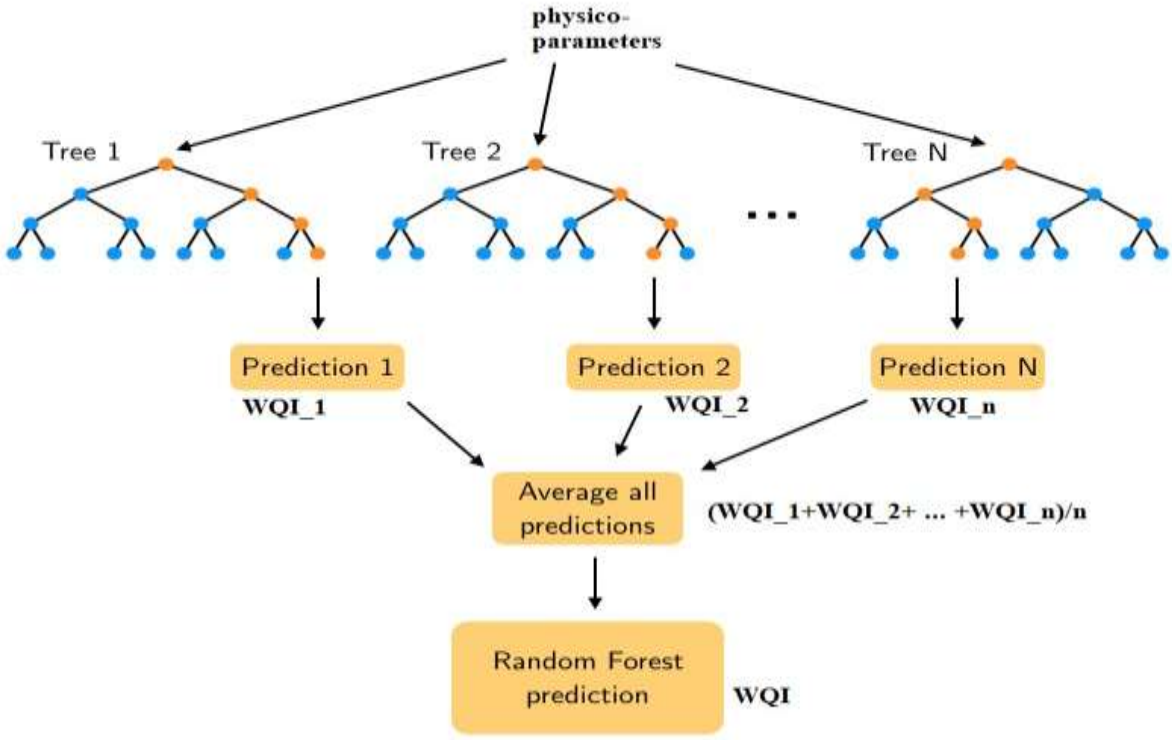


Figure 6:- Visual representation of a random forest structure.

Structure of artificial neural network

They were first described in 1943 by the neurophysiologist Warren McCulloch and the mathematicianWalter Pitts and they are often much better than other machine learning methods on large and complex problems.Chu et al.(2020)andTaher et al.(2022)explains the composition and various steps in creating a neural network model.This means that the structure consists neural layers that work together in parallel (Zahiri et al. 2015;Jimeno-Sáez et al. 2018; Ali & Shahbaz 2020; Hosseini et al. 2022).In this studyas shown in figure 7 one models, Multi-Layer Perceptron (MLP) was presented and a brief description of this is given here.

A Multi-Layer Perceptron networkinvented in 1957 by Frank Rosenblatt at the Cornell Aeronautical Laboratory, also known as a multi-layer neural network,is composed of multiple layers of neurons, namely the input layer, one or more hidden layers, and the output layer.The neurons within these layers are connected in specific ways.In this study the input layer composed of twelve neuronsthat represent thephysical-chemical parameters of the water of lake Nokoué, two intermediate layers of which the first layer is composed of 64 neurons and the second is composed of 32 neurons.The randomized algorithm was used to determine the appropriate number of neurons for the formation of themodel. The output layer is composed of a single node that represents the target (WQI). Each neuron in the input layer is connected to all the neurons in the next layer, the first hidden layer. These intermediate layers between the input layer and the output layer perform complex calculations by combining information from neurons in the previous layer. Each neuron in a hidden layer is connected to all neurons in the previous layer and the next layer. The output layer neuron generates the outputs from the network. The neuron in the output layer is connected to all neurons in the last hidden layer. Each connection between two neurons is associated with a weight that adjusts the impact of input on output. During the training phase, these weights are adjusted to minimize the error between the predicted and actual outputs. In this study we have applied a Rectified Linear Unit activation function using (ReLu)on eachintermediate neuron. With default values, this returns the standard ReLU activation:max (x, 0), the element-wise maximum of zero and the input tensor. Where x is theinput tensor or variable. The mathematical expression of the neural computing is given in the equation (7):

$I_j = f(\sum_i w_{ij} \times \alpha_i + \theta_i)$(7)

Where$\alpha_i$are inputs (explanatory variables), $w_{ij}$are the weights, $\theta_j$are bias, f is an activation function that governs each cell and acts on its inputs(Bhutani 2014).During the learning phase, after having calculated the errors of the

neural network, it is necessary to correct them in order to improve its performance. To minimize these errors - and thus the objective function - the stochastic gradient descent (SGD) algorithm is used.

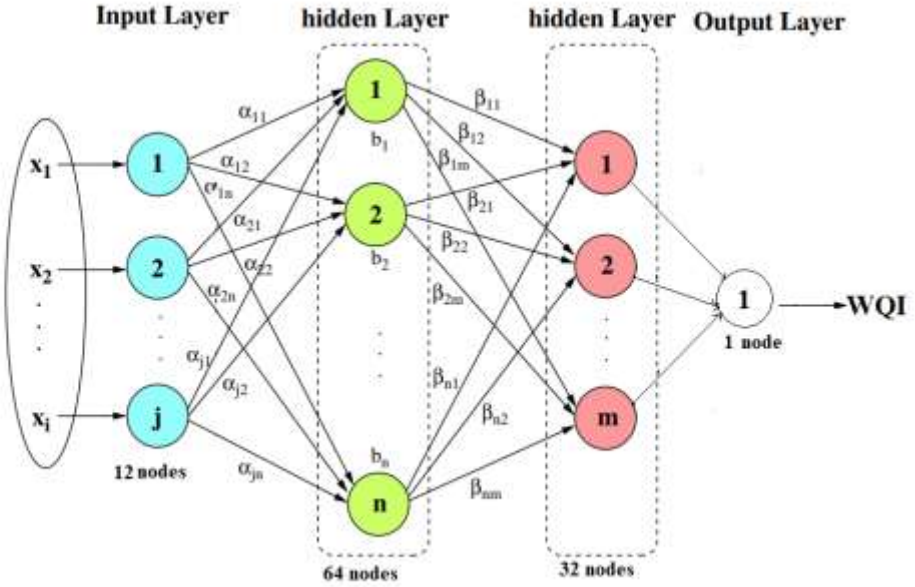


Figure 7:- Architecture of the artificial neural network for the prediction of the water quality index of lake Nokoué.

Performance evaluation metrics

This study considered four commonly used statistical evaluators or metrics, namely, Coefficient of Determination (R^2) and Root Mean Square Error (RMSE) also shown respectively in equation (8) and equation (9) (Dahal et al. 2021).Correlation Determination (R^2) explains the strength and direction of relationships. Its values range from -1 to 1. The value -1 indicates a perfectly inverse relation, while 1 represents a perfectly proportional relation; nearly zero or zero shows poor or no relation at all. The R^2 is the square of the correlation determination that highlights the collinearity or strength of the relationship in terms of a positive number. It is a common metric for measuring the goodness of fit of the models. The least value 0, represents no relation, while the maximum value 1, represents a perfect fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Q_{obs}^{i} - Q_{sim}^{i}\right)^2}{\sum_{i=1}^{n}\left(Q_{obs}^{i} - \overline{Q}_{obs}\right)^2} \quad (8)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{\left(Q_{obs}^{i} - Q_{sim}^{i}\right)^2}{n}} \quad (9)$$

where n is the length of the observation or simulation; Q_{obs}^{i} and Q_{sim}^{i} are the observations and simulations of the quality index at time step i;$\overline{Q}_{obs}$ and $\overline{Q}_{sim}$ are the averages of the observations and simulation of the quality index respectively.

Results and Discussion:-

The results obtained in this study are analyzed, interpreted and presented in tables and graphs, in order to give a general idea of the statistical analyses and the performance of the models on the prediction of the water quality index of lake Nokoué.

Descriptive statistics of the data

The descriptive statistics of the data recorded in the following table4 show that the variables are in disproportionate magnitudes. It is observed that the highest values are seen in salinity. This can be justified by the fact that the Atlantic Ocean feeds into lake Nokoué during the low water period. There are also high values in total dissolved solids (TDS). It is also noticeable that nitrites (NO_2) and nitrates (NO_3) have the lowest values.

Table 4:- Descriptive statistics of physical-chemical parameters.

		mean	std	min	25%	50%	75%	max
Temperature	°C	29.13	2.34	23.0	27.35	29.7	30.38	34.1
Turbidity	NTU	21.08	24.75	1.0	9.69	12.32	22.75	144.0
Conductivity	µS/cm	12.54	12.48	0.07	2.06	10.85	17.97	64.2
MES	mg/l	13.15	14.74	1.0	6.0	8.32	14.0	77.0
TDS	mg/l	247.33	433.23	0.94	7.36	32.90	243.25	1805.0
Salinity	mg/l	4330.0	5292.72	0.0	75.0	1000.0	7600.0	19700.0
O_2dissolved	mg/l	4.76	3.68	0.1	0.96	4.9	7.06	12.2
Ph	---	7.31	0.73	5.45	6.81	7.38	7.78	8.78
OrthoPhosporus	mg/l	0.17	0.22	0.04	0.10	0.12	0.17	1.52
Nitrites	mg/l	0.01	0.01	0.001	0.01	0.01	0.02	0.04
Nitrates	mg/l	2.42	1.81	0.01	0.64	2.7	3.3	5.9
DCO	mg/l	86.16	57.98	3.67	48.2	63.75	115.93	281.97
WQI	-----	99.76	77.36	23.67	58.43	81.49	119.41	485.59

In Figure 4 below, which represents the boxplots of the water quality index of lake Nokoué, a disproportionate monthly variation in the water quality index is observed. So, the water quality of lake Nokoué can be categorized into "Good water" during the rainy season to "Poor water" during the dry season. Several other studies based on WQI for untreated natural water conducted in India have also reported poor quality of water in summer and monsoon compared to the winter season (Puri et al., 2015) have also observed this situation in Ambazaria lake. This seasonal variation in the water quality index of lake Nokoué can be attributed to the seasonal effects on waterquality.

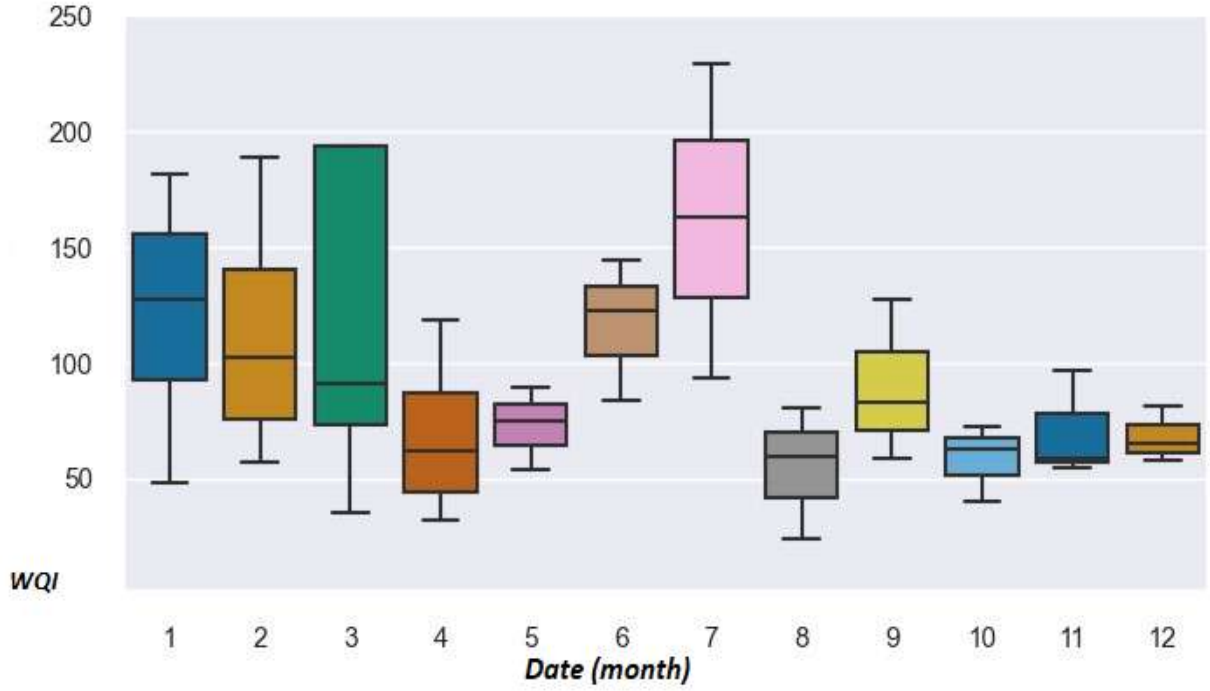


Figure 4:- Monthly change in the lake Nokoué boxplot of the water quality index.

Seven variables show a positive correlation with the water quality index of lake Nokoué as shown in the correlation matrix, which the turbidity, suspended solids and Orto phosphorus are have the greatest influence on the variation of the water quality index of lake Nokoué with correlation coefficients of 0.97, 0.91 and 0.37. The rest of the variables are negatively correlated with the water quality index of lake Nokoué. In the Figure 5indicate the relative influence of the entered variables. All variables were considered for training and validation. The data base is partitioned into two sets: training data set (x_train; y_train) and test data set (x_test; y_test). The size of the training data set is 80 percent and that of the testing data set is 20 percent.

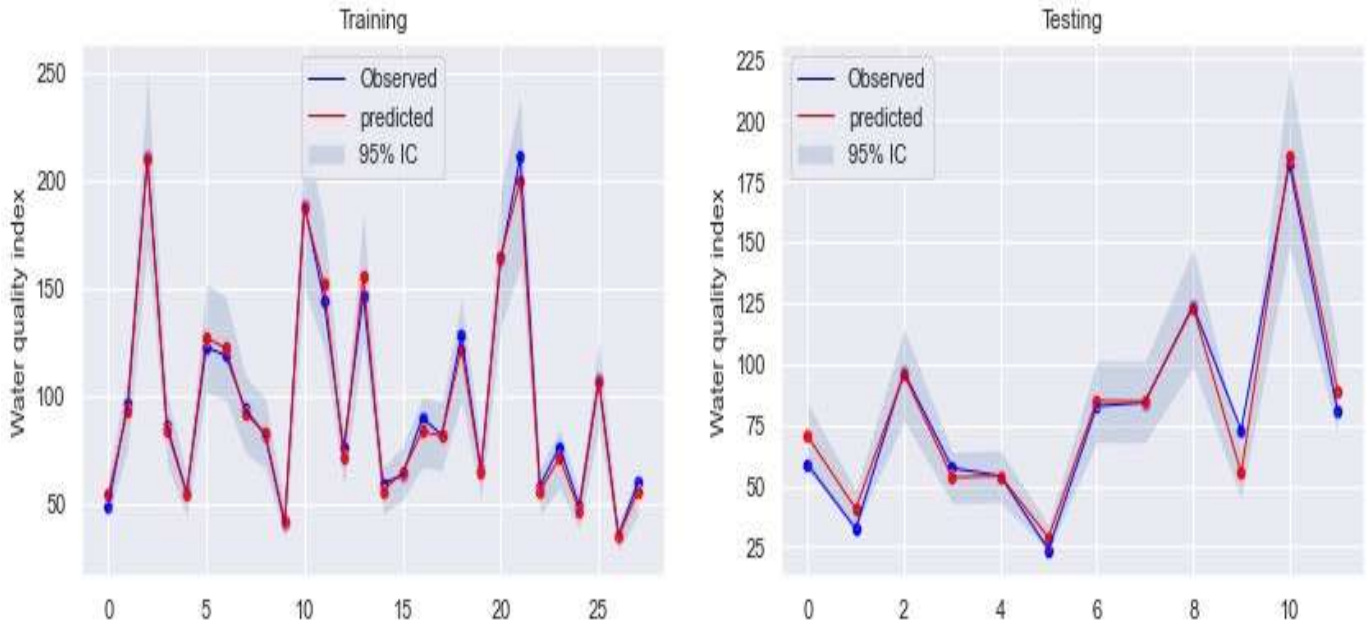
Figure 5:- Correlation matrix of variables.

Results of the random forest model

The observed and predicted from the training and validation phase of the best scenario for the RF model is graphically shown in Figure8. The graphical of RF model results also indicate that the three predictors (hyperparameters)perform a lot better than other combination predictor. To justify the best predictor combination (hyperparameters selected) for RF modelthe grid search method indicates: 100 for the number of trees used, 4 for the maximum depth of a tree (the maximum number of entities), and 1 for the minimum number of samples required to be at a leaf node. The summary of the performance on the twelve scenarios of the random forest modelthe training and testing phase is shown in the Table 5. The twelfth scenario involving all physical-chemical parameters is the most effective scenario with R^2=0.98, RMSE=0.12which means that all predictors' accounts for 98% of the variation in WQIat the training phase. This demonstrates that, in practical chemistry, the correlation matrix does not adequately interpret the influence of physical-chemical parameters on the water quality index of lake Nokoué. All figures clearly show a better correlation between the observed values and the values predicted by the random forest model during training and testing phase (Figure 8-10).

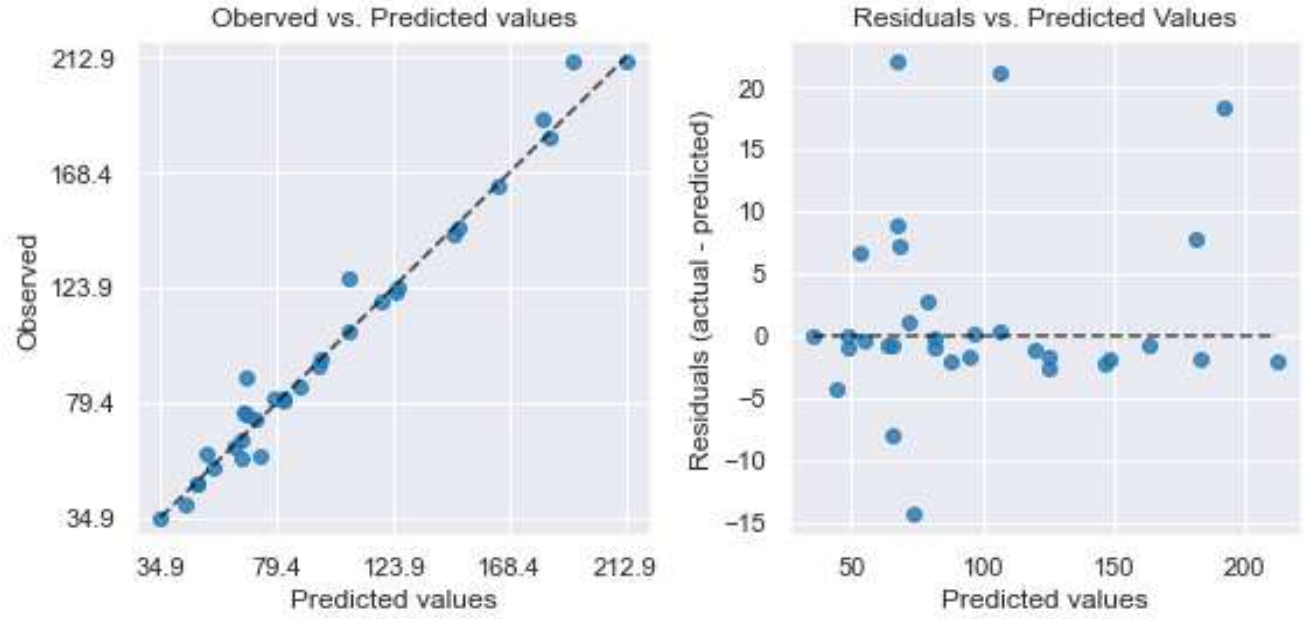
Table 5:- The performances of the twelve scenarios of the random forest model.

RF model	Metrics	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**
Train	RMSE	2.38	1.42	1.28	1.46	1.14	1.24	1.01	1.13	0.66	0.8	0.8	0.12
	R^2	0.81	0.86	0.88	0.86	0.92	**0.90**	**0.93**	**0.92**	**0.97**	**0.96**	**0.96**	**0.98**
Test	RMSE	2.5	1.5	1.29	1.52	1.26	1.29	1.02	1.3	0.20	0.21	0.21	0.19
	R^2	0.55	0.58	0.60	0.58	0.65	0.63	0.66	0.65	0.77	0.72	**0.72**	**0.80**



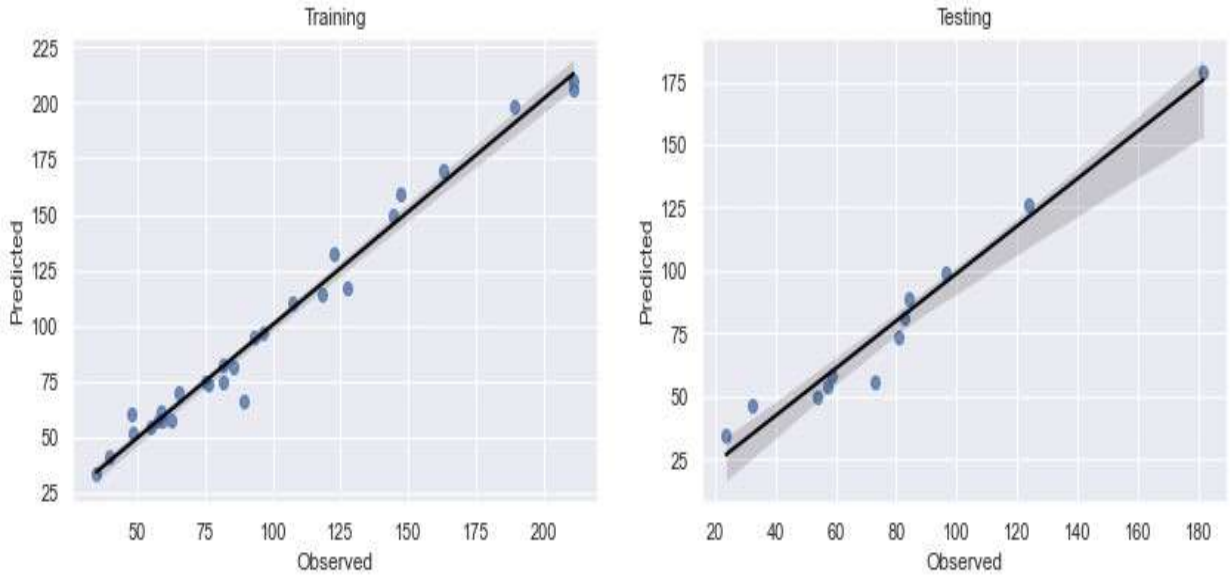
(a) (b)

Figure 8:- (a)Predicted and observed water quality index of the random forest model training phase (b) Predicted and observed water quality index of the random forest model testing.



(a) (b)

Figure 9:- (a)Observed vs predicted values graphic of randomforest model (b) residuals values of random forest model.



(a) (b)

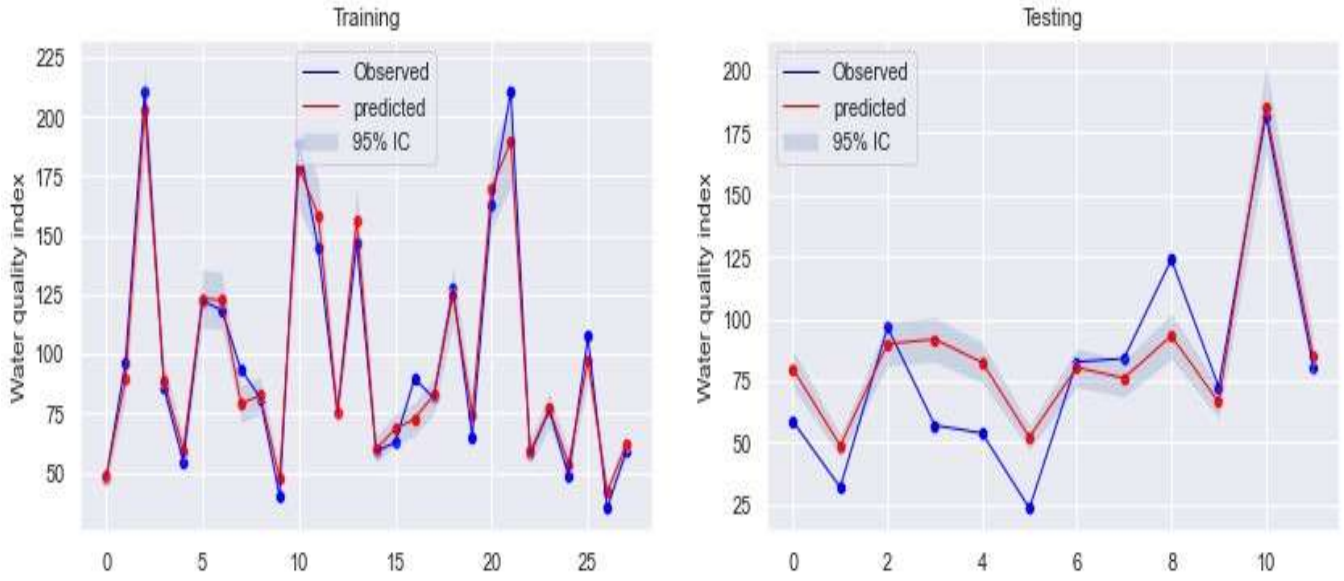
Figure 10:- (a)training phase graphic of random forest(b) testing phase graphic of random forest.

Results of Neural network model:-

Table 6 displays the performance of the artificial neural network model across all constructed scenarios. A notable performance is observed in scenario Twelve during the training phase, achieving 95% for R^2 and 0.72 for RMSE. However, during the testing phase, the performance drops to 57% for R^2 and 0.75.Across all scenarios, the performance of the neural network model is mediocre during the testing phase and relatively good during the training phase.The weakest performances are observed in scenarios 8 and 10 (S8 and S10), respectively.This situation indicates that the neural network model fails to capture the non-linear relationships between physico-chemical parameters and the water quality index of lake Nokoué.The comparison between the results of the water quality index prediction and the real data observed show that the predictionof the neural network model is more or less correlated.This means that the artificial neural network is less efficient to the prediction of the water quality index.The predicted values of the artificial neural network model in the Figure 11show overfitting and underfitting in some places. This is due to the fact that these artificial neural network models are generally applied to a large volume of data in training (Weng 2022).Figure 11 and figure 14shown the change in training and validation error with the number of iterations in the artificial neural network model.An early stopping process that optimizes the model by monitoring the performance of the model on a set of test data and stopping the learning procedure once the training and validation error values reach a constant value on the training and test data beyond a certain number of iterations. We recorded better predictive results with 100 iterations and estimated loss function value (training RMSE) of4.29 at the training phase and 5.59 at the validation RMSE. We used stochastic gradient descent (SGD) and adaptive moment estimation (Adam) as the optimization algorithm to update the network weights.

Table 6:- The performances of the twelve scenarios of the artificial neural network model.

ANN model	Metrics	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**
Train	RMSE	1.7	1.68	1.52	3.17	2.68	1.42	2.68	5.61	3.05	3.54	1.03	0.72
	R^2	0.74	0.79	0.81	0.52	0.68	0.82	0.69	0.38	0.55	0.49	**0.90**	**0.95**
Test	RMSE	1.85	1.79	1.5	3.64	2.88	1.66	2.87	6.91	3.16	4.40	1.12	0.75
	R^2	0.33	0.37	0.42	0.24	0.29	0.43	0.29	0.16	0.26	0.24	**0.50**	**0.57**



(a) (b)

Figure 11:- (a) Predicted and observed water quality index of the artificial neural network model training phase (b) Predicted and observed water quality index of the artificial neural network testing phase.

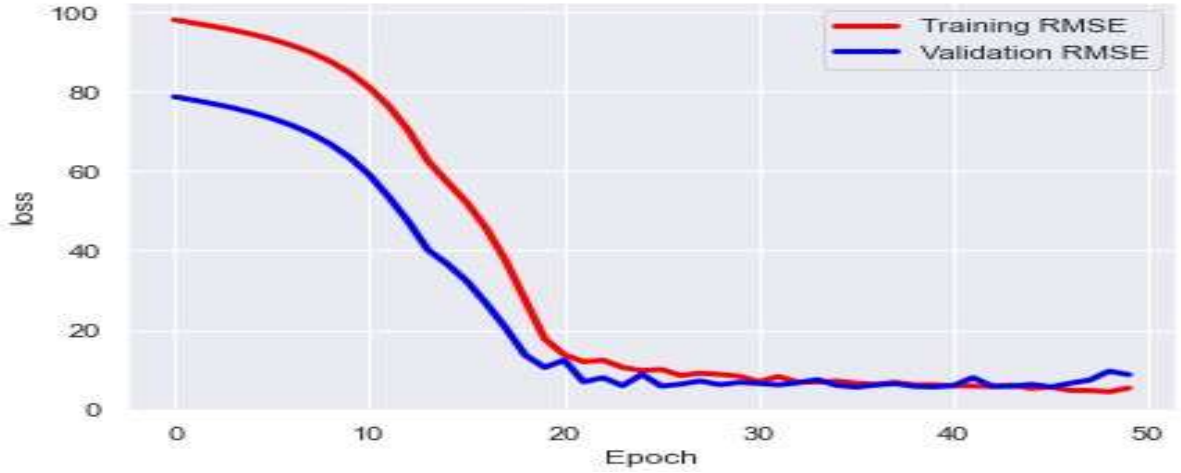
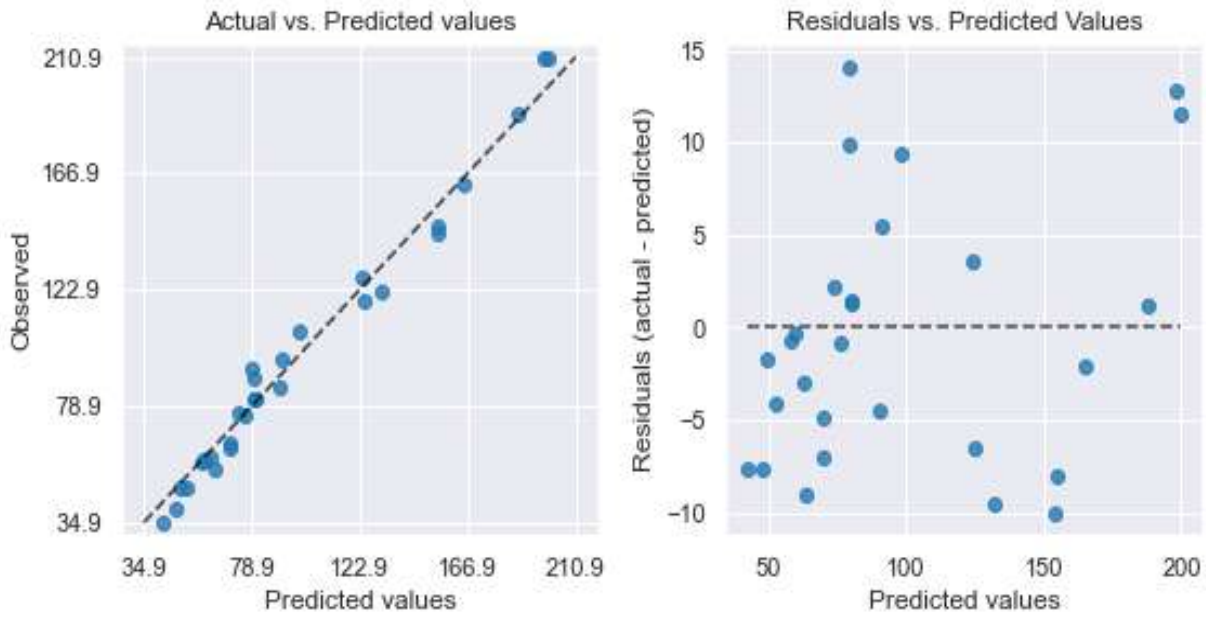
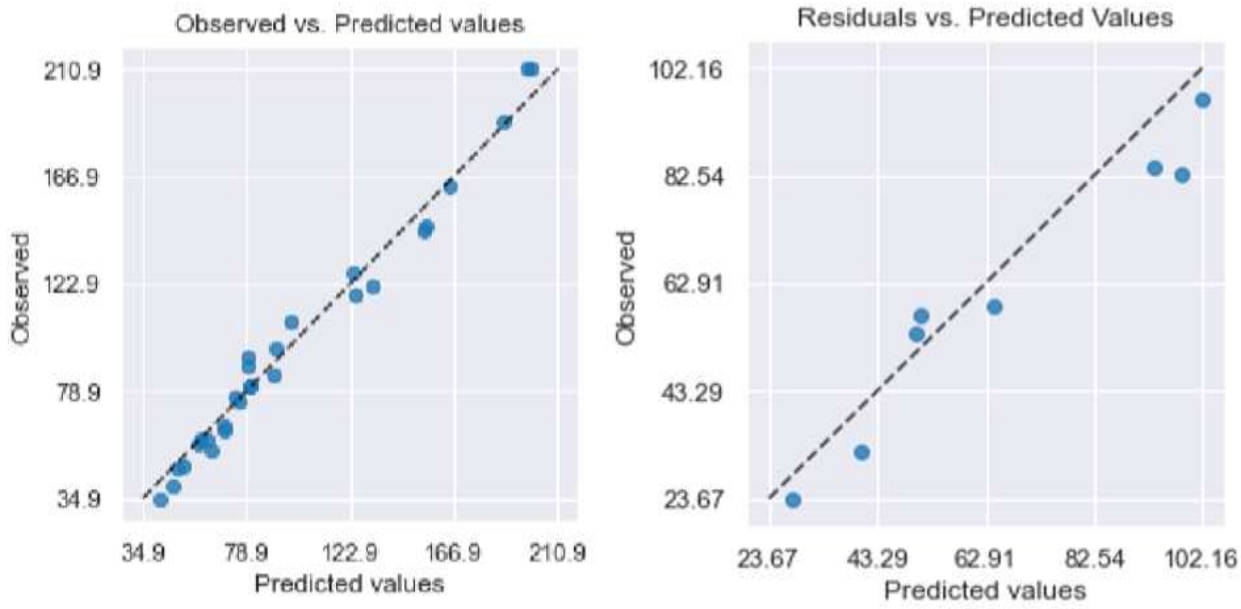


Figure 12:- Comparison of the loss and val_loss of training and validation of the artificial neural network.



(a) (b)

Figure 13:- (a)Observed vs predicted values graphic of artificial neural network model (b) residuals values of artificial neural network model.



(a) (b)

Figure 14:- (a)Training phase graphic of artificial neural network (b) testing phase graphic of artificial neural network.

Conclusions:-

This work has shown that various statistical analyses can be used to analyze data on physical-chemical water parameters to determine the water quality index. This study clearly shows that it is possible to predict the water quality index of Lake Nokoué with a high accuracy using a large amount of information on physical-chemical parameters from the machine learning models. Undoubtedly, it can afford early warnings when the water quality changes as well as it can reduce the adverse consequences resulting from the poor water quality. Herein, the RF and

ANNs approach were introduced to estimate the water quality index of Lake Nokoué using 12 parameters. The presented RF accurately estimated the water quality index with relatively minor prediction errors, proving a quite efficient and robust performance. Prediction precision with RF was equal to 0.98 at the training phase and equal to 0.80 at the testing phase.Even though the outcomes seem to be reasonable, the application of water quality parameters is quite sensitive to the error level. For better decision making based on the model results we suggest collecting samples in order to have more data to improve the quality of the model results. We also recommend experimenting with other than machine learning models in order to compare results with our models.

Acknowledgements:-

This work is supported by the African center of excellence for water and sanitation project (C2EA) funded in part by the World Bank and the French Development Agency.

We would like to particularly thank the Beninese Government for its impetus in building the capacities of young people in the water sector through the National Water Institute (INE).

Conflict of interest

The authors declare that they have no conflict of interest in the publication of this article.

Data availability statement

Data cannot be made publicly available; readers should contact the corresponding author for details.

References:-

1. Abobakr Saeed Abobakr Yahya, Ali Najah Ahmed, Faridah Binti Othman, Rusul Khaleel Ibrahim, Haitham Abdulmohsin Afan, Amr El-Shafie, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, & and Ahmed Elshafie. 2019 Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. Water 2019, 11(1231), 16.
2. Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, & Abbas Parsaie. 2018 Water quality prediction using machine learning methods. Water Quality Research Journal 53(1), 11.
3. Balan, I. N., Shivakumar, M. and Kumar, P. D. M. 2012 An assessment of ground water quality using water quality index in Chennai, Tamil Nadu, India, Chronicles Young Scient., 3(2). 146-150.
4. Bhutani, G. 2014 Application of machine-learning based prediction techniques in wireless networks.Int'l J. of Communications, Network and System Sciences.
5. Brown, R. M., McClelland, N. I., Deininger, N. I., & O'Connor, M. F. 1972 A water quality index-crashing the psychological barrier. In Indicators of environmental quality. 173-182.
6. Calèche NehemieNounagnon AVAHOUIN, Henri Sourou TOTIN VODOUNON, Ernest, & AMOUSSOU. 2018. Variabilité climatique et production halieutique du lac Nokoué dans les Aguégués au Bénin.8(2), 51-61.
7. Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. 2021 Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11(2), Art. 2.
8. Dalai, C., Azizian, J. M., Trieu, H., Rajan, A., Chen, F. C., Dong, T., Beaven, S. W., & Tabibian, J. H. 2021 Machine learning models compared to existing criteria for noninvasive prediction of endoscopic retrograde cholangiopancreatography-confirmed choledocholithiasis. Liver Research, 5(4), 224-231.
9. Daouda MAMA. 2010 Methodologie et resultats du diagnostic de l'eutrophisation du lac Nokoue (BENIN) [PhD Thesis]. UNIVERSITE DE LIMOGES.
10. Firmin M. A., Josué E. B. Z, Bruno E. L., Amédée C., Oswald D., Daouda M.2017.Application Use of Water Quality Index (WQI) and Multivariate Analysis for Lake Nokoué Water Quality Assessment. American Journal of Environmental Science and Engineering. Vol. 1, No. 4, 2017, pp. 117-127.
11. Flavien DOVONOU, Martin AINA, Moussa BOUKARI, & Abdoukarim ALASSANE. 2011 Pollution physico-chimique et bactériologique d'un écosystème aquatique et ses risques écotoxicologiques : Cas du lac Nokoue au Sud Benin.5(5), 1590-1602.
12. HaïboChu, WenyanWu, Q.J. Wang, RoryNathan, &JiahuaWei. 2020 Un cadre de modélisation d'émulation basé sur ANN pour la modélisation des inondations : Application, défis et orientations futures. Volume 124, 10(45-87), 13.
13. Horton, R. K. 1965. An index number system for rating water quality. 3(37), 300-306.
14. House, M. A. 1990. Water quality indices as indicators of ecosystem change. Environmental Modeling & Assessment, 255-263.

15. House, M. A., & Ellis, J. B. 1987 The development of water quality indices for operational management. Water Science and Technology, 145-154.
16. Huang, H., Lin, Z., Liu, S., & Zhang, Z. 2022 A neural network approach for short-term water demand forecasting based on a sparse autoencoder. Journal of Hydroinformatics, jh2022089.
17. Iorliam, I. B., Ikyo, B. A., Iorliam, A., Okube, E. O., Kwaghtyo, K. D., & Shehu, Y. I. 2021Application of Machine Learning Techniques for Okra Shelf-Life Prediction. Journal of Data Analysis and Information Processing, 9(3), Art. 3.
18. Josué Zandagba, Firmin M. Adandedji, Daouda MAMA, Amédée Chabi, & Abel Afouda. 2016 Assessment of the Physico-Chemical Pollution of a Water Body in a Perspective of Integrated Water Resource Management: Case Study of Lake Nokoué. Journal of Environmental Protection, 2016, 7(656-669), 14.
19. Kalhori, S. R. N., & Zeng, X.-J. 2013 Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. Journal of Intelligent Learning Systems and Applications, 5(3), Art. 3.
20. Lalèyè, Chikou, Teugels, &Dewalle, V. 2004 Etude de la diversité ichtyologique du bassin du fleuve Ouémé au Bénin (Afrique de l'Ouest). Cybium, 28(4), 329-339.
21. Liou, S. M., Lo, S. L, & Wang, S. H. 2004 A generalized water quality index for Taiwan. 35-52.
22. Maier, H. R., & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling & Software, 15(1), 101-124.
23. P. Vasistha & R. Ganguly. 2020 Water quality assessment of natural lakes and its importance: An overview. Proceedings 32(544–552), 9.
24. P. J. Puri, M. K. N. Yenkie, D. B. Rana and S. U. Meshram. 2015Application of water quality index (WQI) for the assessment of surface water quality (Ambazari Lake). European Journal of Experimental Biology, 5(2):37-52.
25. Rosenblatt, Frank 1958A Probabilistic Model For Information Storage And Organization in the Brain" Psychological Review. 65 (6), 386–408
26. Sampurno, J., Ardianto, R., &Hanert, E. 2022 Integrated machine learning and GIS-based bathtub models to assess the future flood risk in the Kapuas River Delta, Indonesia. Journal of Hydroinformatics, jh2022106.
27. Sèdami Pivot Amour SACHI & Innocent BOKOSSA YAOU. 2016 Evaluation de la connaissance et de la mise en œuvre des bonnes pratiques d'hygiène par les populations riveraines du lac Nokoué (Sud-Bénin). Int. J. Biol. Chem. Sci.10(4),1823-1831.
28. Taher Rajaee, SalarKhani, &MasoudRavansalar 2022 Modèles uniques et hybrides basés sur l'intelligence artificielle pour la prédiction de la qualité de l'eau dans les rivières : Un examen. 200(103-978), 4.
29. Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. 2013 Water quality assessment in terms of water quality index.American Journal of Water Resources, 34-38.
30. Wang, N. 2017 Bankruptcy Prediction Using Machine Learning. Journal of Mathematical Finance, 7(4), Art. 4.
31. Weng, C.Y. 2022 Land-Use Classification via Transfer Learning with a Deep Convolutional Neural Network. Journal of Intelligent Learning Systems and Applications, 14(2), Art. 2.
32. Zheng, H. 2018 Analysis of Global Warming Using Machine Learning. Computational Water, Energy, and Environmental Engineering, 7(3), Art. 3.
33. Zhu, X., Guo, H., Huang, J. J., Tian, S., Xu, W., & Mai, Y. 2022 An ensemble machine learning model for water quality estimation in coastal area based on remote sensing imagery. Journal of Environmental Management.3(23),116-187.